The Author(s) *BMC Bioinformatics* 2016, **17**(Suppl 11):321
DOI 10.1186/s12859-016-1171-x

BMC Bioinformatics

RESEARCH Open Access

MiRNATIP: a SOM-based miRNA-target interactions predictor

Antonino Fiannaca*, Massimo La Rosa, Laura La Paglia, Riccardo Rizzo and Alfonso Urso

From 11th International Meeting on Computational Intelligence Methods for Bioinformatics and Biostatistics (CIBB 2014) Cambridge, UK. 26-28 June 2014

Abstract

Background: MicroRNAs (miRNAs) are small non-coding RNA sequences with regulatory functions to post-transcriptional level for several biological processes, such as cell disease progression and metastasis. MiRNAs interact with target messenger RNA (mRNA) genes by base pairing. Experimental identification of miRNA target is one of the major challenges in cancer biology because miRNAs can act as tumour suppressors or oncogenes by targeting different type of targets. The use of machine learning methods for the prediction of the target genes is considered a valid support to investigate miRNA functions and to guide related wet-lab experiments. In this paper we propose the miRNA Target Interaction Predictor (miRNATIP) algorithm, a Self-Organizing Map (SOM) based method for the miRNA target prediction. SOM is trained with the seed region of the miRNA sequences and then the mRNA sequences are projected into the SOM lattice in order to find putative interactions with miRNAs. These interactions will be filtered considering the remaining part of the miRNA sequences and estimating the free-energy necessary for duplex stability.

Results: We tested the proposed method by predicting the miRNA target interactions of both the Homo sapiens and the Caenorhbditis elegans species; then, taking into account validated target (positive) and non-target (negative) interactions, we compared our results with other target predictors, namely miRanda, PITA, PicTar, mirSOM, TargetScan and DIANA-microT, in terms of the most used statistical measures. We demonstrate that our method produces the greatest number of predictions with respect to the other ones, exhibiting good results for both species, reaching the for example the highest percentage of sensitivity of 31 and 30.5 %, respectively for Homo sapiens and for *C. elegans*. All the predicted interaction are freely available at the following url: http://tblab.pa.icar.cnr.it/public/miRNATIP/.

Conclusions: Results state miRNATIP outperforms or is comparable to the other six state-of-the-art methods, in terms of validated target and non-target interactions, respectively.

Keywords: miRNA, SOM, mRNA, Target prediction, miRNA-mRNA interactions

Background

MicroRNAs (miRNAs) are small non-coding single stranded RNA molecule, 22–25 nucleotides (nt) long, found in many organisms (plants, animals, and some viruses) [1]. MiRNAs are important players in gene regulation. The most important step in their regulatory function is the targeting of RNA messengers (mRNAs). MiRNAs, in fact, are responsible for degradation or repression of mRNAs at post-transcriptional level, when their sequences bind with partially complementary sites. This way, they play a crucial role in the cell differentiation and proliferation, apoptosis, and many other physiological and pathological processes [1].

Expression patterns of miRNAs are highly related to specific external stimuli, developmental stage or tissue. For example, in cancer disease the expression levels of miRNAs are known to change considerably [2].

Many recent works proved a different behaviour of cellular actors mediated by a differential expression of miRNAs that are cell condition or tissue/specific [3]. In the pathology of cancer this is relevant as they can act as tumour suppressors or oncogenes by targeting different

*Correspondence: fiannaca@pa.icar.cnr.it
National Research Council of Italy, ICAR-CNR, via Ugo La Malfa 153, 90146 Palermo, Italy

type of targets, leading respectively to decrease or accelerate the tumorous processes. Thus, analysing the miRNA-mRNA interaction would mean to better understand the molecular mechanism of the pathological condition compared to the normal cell behaviour, through the main actors that are proteins, and moreover to hypothesize new therapeutic strategies of intervention to stop the malignant processes [4, 5].

MiRNAs were first identified in 1993 [6] via classical genetic techniques in Caenorhabditis Elegans (Nematoda; Rhabditidae).

It is just over the last decade that thousands of miRNAs have been discovered in all kinds of taxa and their regulations in cancer have been analysed [7–9]. Unfortunately, most of these studies were focused only on a specific subset of miRNAs, or a limited group of patients. The role of miRNAs was also demonstrated in the early stages of the disease progression and metastasis. In fact, several experimental evidences showed miRNAs are involved in the regulation of those biological processes, leading to the acquisition of metastatic potential, including adhesion, invasion, migration, epithelial-mesenchymal transition and angiogenesis [10, 11].

MiRNAs interact with their mRNA targets especially by base pairing in the 3'-untranslated regions (3'UTR) of mRNA sequences. In living species, near perfect base pairing is required between the so called miRNA seed, i.e. the first 8 nt in the 5' miRNA sequences, and a target site in the 3'UTR mRNA sequences. In plants, the whole miRNA sequences usually have near-perfect pairing with their mRNA targets, which induces gene repression through cleavage of the target transcripts. In contrast, with few exceptions, in animals, the base pairing between the whole miRNA sequences and their mRNA targets is imperfect. However, some authors have identified three main rules for miRNA-target base pairing by experimental and in silico analysis [12]:

1. Perfect and contiguous base pairing of miRNA seeds, made of nucleotides 2 to 8 in 5' miRNA, which nucleates the miRNA-mRNA association. In general, conditions as mismatches and bulges in the seed region should be avoided because it greatly affect on repression.
2. There must be enough complementarity to the miRNA 3' half in order to stabilize the interaction. In this region bulges and mismatches are generally allowed.
3. The central region of the miRNA-mRNA duplex should have bulges or mismatches, in order to preclude the endonucleolytic cleavage of mRNA.

Because experimental identification of miRNA targets is a difficult work, the aid of computational tools for target predictions is a valuable instrument to investigate miRNA functions and to guide related wet-lab experiments. Usually two research problems involving miRNA are addressed with computational methods, i.e. miRNA genes detection and miRNA targets prediction. The former consists in the identification of those regions in the genome that produces the miRNAs; the latter searches for the mRNAs that could interact with the miRNAs. Machine learning methods have improved the performance of both miRNA gene detection and target prediction [13–15]. These approaches typically make use of sequence data (e.g. of short 6–8 nt miRNA binding motifs), secondary structure (e.g. stem-loops using thermodynamic modelling) and evolutionary conservation to identify putative candidates, using algorithms such as Hidden Markov Models (HMM) [15], Random Forest classifiers [14] or Support Vector Machines (SVM) [13].

In this work, we present miRNATIP (miRNA Target Interaction Predictor), a method for miRNA target predictions based on Self-Organizing Maps (SOM). SOM networks [16] are artificial neural networks widely used to categorize large high-dimensional datasets by mapping the data into a smaller dimensional space, typically into a two-dimensional lattice of interconnected neurons. Each neuron of a SOM represents a reference model, corresponding to a local domain in the input space [17]. By using a competitive learning, rather than an error-correction learning like the back-propagation with gradient descent adopted by other artificial neural network algorithms, the SOM algorithm tries to reproduce the self-organizing mechanism that creates the somatosensory in some areas of the brain. In this sense the SOM algorithm can be defined as an artificial neural network, like many other algorithms inspired by neurons in the brain. The SOM is more than a clustering algorithm because it gives a visualisation of the distribution of the patterns in the input space. When the input patterns are projected on the map the clusters can be visualised, and the map can be divided into areas where the input patterns share some feature values. In our work, the SOM algorithm is able to cluster together the miRNA seeds and, consequently, to project on the trained lattice the 3'UTR mRNA sequences in order to find a preliminary list of putative targets. This list will be filtered out considering the remaining parts of both miRNA and mRNA sequences and finally it will be shortened using a threshold over the free-energy, whose values provides hints about the thermodynamic stability of the miRNA-mRNA duplex [18]. In bioinformatics, SOM has been previously applied to issues like clustering of protein sequences [19] and molecular compounds [20], gene finding [21], and identification of transcription factor binding sites [22].

The paper has the following structure: the next Section describes some related works about miRNA-target

predictors; "Methods" section reports in details our proposed algorithm and the datasets used in our experiments; the basic SOM algorithm and some details of the other algorithms used for comparison; "Results and discussion" section reports both the methodology to tune the the parameters of miRNATIP algorithm and the the experimental prediction results compared with other six state-of-the-art miRNA-target prediction tools. Finally, some conclusion as well as our future work are reported in "Conclusion and future work" section.

Related works

All known miRNA-targets are mainly based on experimentally validated miRNA-mRNA interactions [23, 24]. However, they represent only a very small part of all existing interactions. For this reason, in recent years several miRNA-target predictors have been developed. The available algorithms were recently reviewed in [18, 25, 26] focusing on their bioinformatics, mathematical and statistical features. In the following it will be discussed some of these in more detail.

The miRanda algorithm [27] searches for target sites on the 3'UTR regions of mRNAs. It considers both the binding energy for the duplex stability and the conservation of the target site among different species. Those miRNAs having multiple binding sites within 3'UTR are highly scored.

PicTar [28] identifies a list of putative targets searching for almost fully complementarity sites between miRNAs and 3'UTR mRNAs. The free energy between the binding sites is then computed and finally the results are ranked by means of a score obtained using an HMM, and miRNAs having multiple binding sites are highly scored. In order to refine the identified targets, PicTar looks for the target site conservation among eight vertebrate species.

TargetScan [29] algorithm is based on the identification of full complementary zones between the miRNA seed (nucleotide 2 to 7) and 3'UTR mRNA. Starting from those sites, TargetScan searches for larger interactions, ranking the results in three groups according to the length of the matches. In particular, the presence of an adenine in the first position of the target site is highly scored because of its evolutionary conservation.

DIANA-microT [30] scans putative target sites by means of a 38 nt-long sliding window moved over the 3'UTR region of mRNA. At each shift, the minimum free energy between the miRNA-mRNA binding sites is computed and then it is compared with the energy related to the supposed full (100 %) miRNA-mRNA complementarity. miRNA seed matches of 7, 8 and 9 nt are allowed, as well as 6 nt-long matches if there are further complementary sites in the remaining region of miRNA.

PITA [31] algorithm consists of two steps. In the first one, it looks for putative target sites considering near perfect complementarity between miRNA seed and 3'UTR mRNA. In the second step, PITA takes into account the actual accessibility of the target site, related to the transcript secondary structure, by combining the free energy of the miRNA-mRNA bound and the energy needed to unfold the mRNA and make it accessible.

RNAhybrid [32] searches for miRNA target sites considering the hybridization sites having the most advantageous energy content. Hybridization is a technique that measures the degree of genetic similarity between groups of DNA/RNA sequences [33] and it is usually used to determine the genetic distance between two organisms. RNAhybrid looks for targets in the 3'UTR mRNA.

MirSOM [34] is, at the best of our knowledge, the only other miRNA target prediction tool implementing a SOM. It takes *C. elegans* 3'UTR sequences and cluster them using a SOM in order to identify potential miRNA target sites. The SOM is built upon is a 32×32 grid and it is trained considering all the overlapping 22 nt-long fragments extracted from the 3'UTR mRNA sequences. At the end of the learning phase, MirSOM produces clusters of putative target sites. Then miRNA sequences are assigned to a mRNA cluster if they have a perfect match between their 7 nt-long seed and the last 8 nucleotides of the sequences belonging to that cluster. Because the SOM clusters together not only identical but also similar sequences, it is possible to identify miRNA-mRNA interaction having near perfect seed matching. At this point, each cluster contains a list of putative miRNA targets. Those list are filtered lefting out those miRNA-mRNA couples whose free energy is below a certain threshold.

MirSOM performed well against most other tools with high sensitivity and vastly improved specificity. Unfortunately, it currently supports only *C. elegans* data. The mirSOM interface allows the user to enter an miRNA and the predicted mRNAs are returned as output. mirSOM can be accessed from https://bioinformatics.uef.fi/mirsom/.

Methods

In this Section, it is described the representation adopted for the miRNA and mRNA sequences; then it is presented the four-steps algorithm for the identification of miRNA targets and finally all the datasets used for our experiments are introduced.

Genomic sequence representation

One of the major challenges in bioinformatics is finding the best representation of the DNA/RNA sequences. In our approach, similarly to [34], we represented RNA (miRNA and mRNA) sequences by means of a numerical encoding derived from the position weight matrices (PWMs) [35]. A PWM is a commonly used representation model in biological sequence analysis, obtained

by computing the frequency of each specific base (A, C, G and T or U) at each nucleotide position in the sequence. The PWM model has been successfully applied to many problems in DNA and protein sequence analysis, for example in the identification of functional sequence elements [36].

In particular, within our method, each RNA sequence is represented with a PMW of $4 \times k$ elements, where 4 are the nucleotide symbols and k is the length of the sequence. Each column j has a fixed value according to the corresponding nucleotide in the j-th position, with $1 \leq j \leq k$. Numeral encoding for each nucleotide was the following: A = $[1000]^t$, C = $[0100]^t$, G = $[0010]^t$, T/U = $[0001]^t$

To measure the dissimilarity between two PWMs we considered, among the others [37], the normalised Euclidean Distance defined as:

$$D(a,b) = \frac{1}{\sqrt{2k}} \cdot \sum_{j=1}^{k} \sqrt{\sum_{b \in \{A,C,G,T\}} \left(P^1_{j,b} - P^2_{j,b}\right)^2} \qquad (1)$$

where P^1 and P^2 are two PWMs, k is the length of the sequences and $P_{j,b}$ is the values in column j with base b. This distance ranges from 0 (perfect identity) to 1 (complete dissimilarity).

MiRNATIP pipeline

The miRNATIP algorithm is composed of four main steps. Figure 1 shows the whole pipeline used in this work and it will be explained in detail in the following subsections. Steps 1 to 3 have been implemented using the Java programming language, so that miRNATIP is platform-independent.

SOM training

In the first step, a set of miRNAs seeds, fixed at a length of 8 nt, is used for the training of a SOM. More details on the SOM training algorithm can be found in [16]. We considered only the 8-mer miRNA seeds, because it has been demonstrated that the seed is mainly responsible of the miRNA target binding (cfr. Background). Each neuron is represented by a PWM of size 4×8 that are first initialised using random values. In this work, we used the batch method for the training of the SOM [38]. Furthermore, the neurons are arranged in a rectangular lattice, in which each neuron is connected to its four neighbours, except for those at the edge of the grid. To locate the best matching unit (bmu), it is calculated the distance between the input vector and the weights of each neuron according to Eq. 1. The result of this step is a set of clusters composed of the 8-mer seeds belonging to each miRNA.

SOM projection

The second step consists in the projection of a mRNA sequence over the trained SOM. For this reason, we extracted all the 8-length mRNA fragments through a 8-mer sliding window with step = 1. This way, we obtained a set of 4×8 PWMs that can be projected over the trained SOM. The result of this step is, for each neural unit (cluster), a list of couples (miRNA_seed, mRNA fragment). Each cluster can be considered as a preliminary set of predicted miRNA-mRNA interactions.

Tail filtering

In this step, we filtered those preliminary interactions considering the remaining part of the miRNA sequences, called miRNA_tail. For each couple (miRNA_seed, mRNA fragment), we considered respectively the miRNA_tail

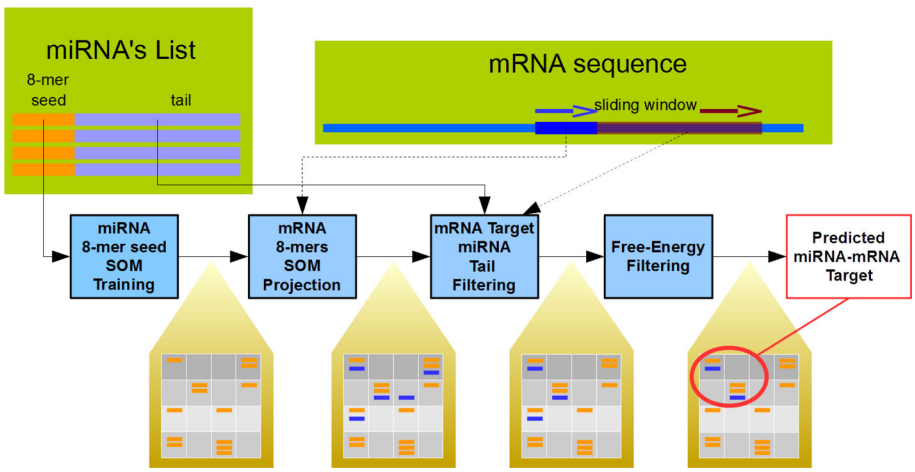

Fig. 1 The proposed miRNA target prediction method. It is composed of four steps, each one represented by a light-blue coloured box. We described SOM training in section "SOM training", SOM projection in section "SOM projection", miRNA tail filtering in section "Tail filtering" and the free-energy filtering in section "Free-energy filtering"

and the mRNA sequence of the same length of miRNA_tail, next to the projected mRNA fragment. Then we computed a dissimilarity measure based on normalised euclidean distance (Eq. 1) between the PWM representation of those two sequences and, according to the the rule no. 2 of the "Background" Section, we retained only the couples of miRNA-mRNA interactions whose distance is below a certain threshold. As reported in the third rule of the "Background" section, in order to take into account also the presence of possible bulge loops between the 8-mer seed and the tail of the miRNA, we considered an offset of few nucleotides (2–3), causing a shift of the mRNA fragment corresponding to the miRNA_tail.

Free-energy filtering

In the last step, we applied a further filtering process to the couples list, based on the minimum free-energy required to form the miRNA-mRNA duplex. For this purpose we used the IntaRNA tool [39, 40]. IntaRNA is able to calculate a free-energy value (given in kcal/mol) from a couple of genomic sequences, considering two different contributions: (1) the free-energy required to unfold the interaction sites both in miRNA and mRNA and (2) the hybridization free-energy between interacting nucleotides of genome sequences. The sum of these two contributions represents the final free-energy score. In our method, we introduced a threshold on this free-energy score: in this way, the putative interactions that obtain a free-energy score over the threshold are removed from the final miRNA-mRNA interaction list.

Datasets

In our study, we focused on two species: *C. elegans* (cel) and human (Homo sapiens - hsa). Human species of course has been chosen for the importance that miRNA target interactions have with regards to regulatory functions involving many diseases, such as cancer. Moreover, we considered *C. elegans* because Nematodes have been studied in a wide range of fields, and they are organisms that allow to help to understand the molecular biology of humans and animals. They are easy to study thanks to their intrinsic features and handiness in cultivation and manipulation. miRNA mature sequences, both for cel and hsa, have been downloaded from miRBase [41] (release 21, update in June 2014), the most comprehensive online database of published validated miRNA sequences and annotation. We obtained 434 and 2588 miRNA sequences for cel and hsa, respectively. As for cel, the validated 3'UTR mRNA sequences are available on WormBase [42] (release WBcel 235, update in April 2013), an online genome database of the nematode model organism *C. elegans*, and they have been downloaded through the BioMart [43] online service. As for the hsa, the validated 3'UTR mRNA sequences were downloaded from Ensembl repository (release 80) [44]. We obtained a total of 30939 and 154666 3'UTR mRNA sequences for cel and hsa, respectively.

Experimentally validated miRNA-mRNA interactions, representing positive examples, were downloaded from mirTarBase [45], a repository of manually verified miRNA target interactions for the most studied species, including cel and hsa. We collected 3209 and 39111 positive validated interactions for cel and hsa, respectively. Finally we considered a set of validated non-target interactions, representing negative examples: for cel we had 16 non-target interactions; for hsa we collected 123 negative validated interactions. Thirteen out of 16 negative interactions for cel have been provided by [34], the remaining 3 and all the negative interactions for hsa have been found in Tarbase (release 7.0) [46], that is a publicly available database containing both miRNA-mRNA target and non-target interactions.

Results and discussion

In this Section we describe how we selected the best parameter configuration for miRNATIP algorithm and then we analyse the prediction results against other six state-of-the-art miRNA-target interaction prediction tools.

MiRNATIP configuration

During the SOM training step (see Fig. 1), in order to obtain the best parameters for the network learning, we performed several tests at varying of network size and learning rate α [16]. The quality of the trained map was measured by means of two evaluation criteria: resolution and topology preservation. These two measures are calculated respectively with the average quantization error (QE) and the topographic error (TE), as defined in [47]. We chose the configuration of SOM parameters that minimize both the QE and TE, according to the Eq. 2, where c is the configuration we adopted, i is a triple of SOM parameters (size, α_{max} and α_{min}) , QE'_i and TE'_i are respectively the normalized value of QE and TE for the triple i.

$$c = \underset{i}{\operatorname{argmin}} \left(mean \left(QE'_i, TE'_i \right) \right) \tag{2}$$

Figure 2 reports a box-plot that shows a trend of performed test at varying of i for the homo sapiens. At the end of the training phase, we obtained a configuration c with the following values: map size= 65×65, initial α= 0.85, final α= 0.1. The same configuration process has been computed for the *C. elegans* species. After this phase, the projection of mRNA fragments over the SOM lattice was performed.

As regards the third step of the proposed method, i.e. the tail filtering, we used a threshold of 0.7 over the euclidean distance (Eq. 1), that allow to preserve at least

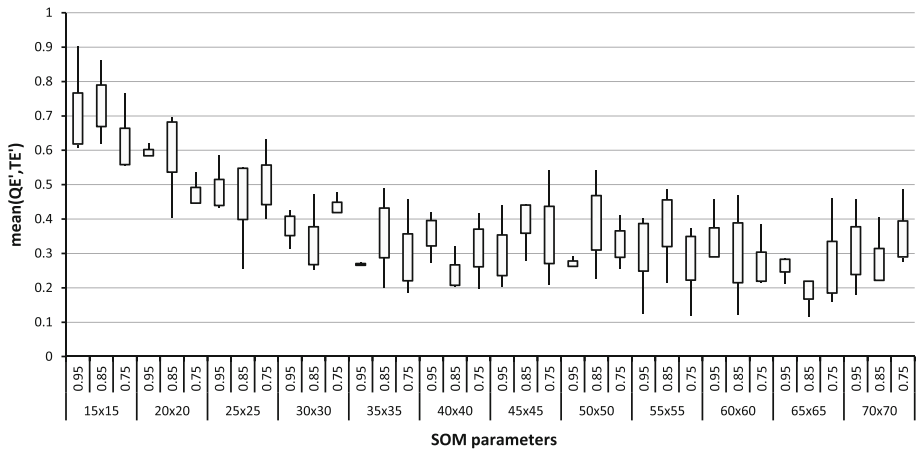

Fig. 2 Measure of SOM training for hsa species at varying of SOM parameters. The box-plot reports the distribution of the mean between QE' and TE' (as defined in section "MiRNATIP configuration"), at varying of SOM parameters, i.e. map size (from 30 × 30 to 70 × 70) and α_{max} (from 0.75 to 0.95). Values of α_{min} (from 0.001 to 0.1) are omitted from the graph for the clarity of image. According to Eq. 2, the best configuration for hsa species is map size = 65 × 65, $\alpha_{max} = 0.85$ and $\alpha_{min} = 0.1$

the 30 % of miRNA-mRNA match in the tail region. In addition, to simulate the presence of a bulge between the seed and the tail of the miRNA binding site, we set the offset = 3, i.e. we supposed a bulge could contain at most three nucleotides. Finally, for establishing the minimum free-energy score, we performed different measurements of the obtained predictions and estimated the optimal threshold for cel and hsa, respectively equal to –6 and –7 kcal/mol.

All the configuration parameters used in this work are reported in Table 1.

Prediction results

MiRNATIP has been run using the datasets presented in "Datasets" section. Prediction results have been compared with those provided by other miRNA target predictors: PITA [31], MiRanda [27], MirSOM [34], PicTar [28], DIANA-microT [30], TargetScan [29]. These predictors have been considered for comparison because they allow to directly download the whole set of miRNA target predictions. MirSOM only provided prediction for cel species; as for TargetScan, the predictions were extracted by means of the miRDIP portal [48]. In order to obtain the most reliable and comparable results as possible, we filtered out the predictions of the other algorithms according to the following criteria. PITA and MiRanda predictions have been chosen considering the same free-energy thresholds we adopted for miRNATIP algorithm (–6.0 kcal/mol for cel and –7.0 kcal/mol for hsa). DIANA-microT predictions have been selected according to the default scores suggested by the authors (0.6 for cel and 0.7 for hsa). Finally for TargetScan we considered the conserved predicted targets, representing the most reliable interactions. In order to evaluate the correctness of a predicted miRNA-mRNA interaction, for each predictor we considered only the set of interactions that involve at least one miRNA/mRNA belonging to the datasets of experimentally validated miRNA-mRNA interactions. Prediction scores have been computed considering the following statistical measures [49]:

$$Accuracy\ (ACC) = \frac{TP + TN}{P + N} \tag{3}$$

Table 1 Parameters used for cel and hsa miRNA-target predictions

	MiRNATIP parameters					
Species	SOM training			Tail filtering		Free-energy filtering score threshold
	Map size	α_{max}	α_{min}	offset	distance threshold	
C. elegans	30×30	0.95	0.1	3	0.7	–6 kcal/mol
Homo sapiens	65×65	0.85	0.1	3	0.7	–7 kcal/mol

The first column reports the species, the next three columns contain parameters for SOM training (section "SOM training"). Forth and fifth columns report parameters for miRNA tail filtering process (section "Tail filtering"). Finally, the last column shows the free-energy threshold score (section "Free-energy filtering")

Table 2 Comparison among the proposed method and the other prediction algorithms for the *C. elegans* species, in terms of true positive and true negative interactions

Validation of miRNA target prediction algorithms for *C. elegans*				
Algorithm	Last update (year)	Predicted interactions	3209 positive validated interactions	16 negative validated interactions
			True positive	True negative
PITA	2008	4874	979	14
MiRanda	2010	3307	829	12
MirSOM	2011	1734	588	15
DIANA-microT	2012	1232	172	16
MiRNATIP	2015	6533	994	15

$$Precision\ or\ positive\ predictive\ value\ (PPV) = \frac{TP}{TP + FP} \quad (4)$$

$$Sensitivity\ or\ true\ positive\ rate\ (TPR) = \frac{TP}{TP + FN} \quad (5)$$

$$Specificity\ or\ true\ negative\ rate\ (TNR) = \frac{TN}{FP + TN} \quad (6)$$

$$Miss\ rate\ or\ false\ negative\ rate\ (FNR) = \frac{FN}{FN + TP} \quad (7)$$

$$Fall - out\ or\ false\ positive\ rate\ (FPR) = \frac{FP}{FP + TN} \quad (8)$$

$$F1\ score = \frac{2TP}{2TP + FP + FN} \quad (9)$$

$$Matthews\ correlation\ coefficient\ (MCC) = \\ = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (10)$$

where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positive, FN is the number of false negative.

Prediction results are reported in Tables 2 and 3 for cel; in Tables 4 and 5 for hsa, respectively. As for Tables 2 and 4, the first two columns of the tables contain respectively the algorithms we compared and the year of the last update. Third column contains, for each method, the subset of the miRNA-mRNA predicted interactions that have at least one miRNA or mRNA belonging to the validated target dataset. Finally, the last two columns report the number of TP and TN with regards to the total number of positive validated interactions and negative validated interactions.

Tables 3 and 5, for cel and hsa respectively, show the prediction scores computed using all the statistical measures presented in Eqs. 3 to 10.

Observing these results, it is possible to notice that our miRNATIP algorithm reaches the best scores regarding the sensitivity, with a score of almost 31 % for both cel (30.97 %) and hsa (30.54 %) species. At the same time we produced the largest number of total predicted interactions. The other best predictors are PITA for cel (30 %) and MiRanda for hsa (about 25 %). Best results are also reached in terms of accuracy and F1 score, confirming the fact that our algorithm predicts the most number of correct interactions.

As for the specificity scores, our miRNATIP algorithm is able to reach one of the best results especially for cel species. In fact the proposed predictor obtained a score of 93.75 %. The best result in terms of specificity for cel

Table 3 Performances of prediction algorithms related to validated interactions in Table 2

Algorithm	ACC %	PPV %	TPR %	TNR %	FNR %	FPR %	F1 %	MCC
PITA	30.8	99.8	30.5	87.2	69.5	12.5	46.7	0.02750
MiRanda	26.7	99.5	25.8	75.0	74.1	25.0	41.0	0.00133
MirSOM	18.7	99.8	18.3	93.7	81.6	6.2	30.9	0.02195
DIANA-microT	5.8	100.0	5.3	100.0	94.6	0.0	10.2	0.01676
miRNATIP	31.2	99.9	30.9	93.7	69.0	6.2	47.3	0.03761

Statistical measures reported in this table are accuracy (*ACC*), precision (*PPV*), sensitivity (*TPR*), specificity (*TNR*), miss-rate (*FPR*), F1-measure (*F1*) and Matthews correlation (*MCC*), respectively

Table 4 Comparison among the proposed method and the other prediction algorithms for the Homo sapiens species, in terms of true positive and true negative interactions

Validation of miRNA target prediction algorithms for Homo sapiens				
Algorithm	Last update (year)	Predicted interactions	3209 positive validated interactions	16 negative validated interactions
			True positive	True negative
PITA	2008	43823	1971	109
MiRanda	2010	420800	9962	73
TargetScan	2012	105407	4367	96
Pictar	2012	40497	2713	100
DIANA-microT	2012	367379	7805	91
MiRNATIP	2015	968798	11945	86

is reached by DIANA-microT tool (100 %), but it produced very low sensitivity score (5.36 %). As regards hsa, mirRNATIP specificity score (69.92 %) is consistent with the scores reached by the other algorithms, whose best specificity score is reached by PITA (88.62 %). Once again, however, PITA reached a very low sensitivity score of about 5 %.

miRNATIP proves to reach the lowest miss-rate (FNR), whereas the fall-out score (FPR) is the second best for cel and and the fifth for hsa. Finally, miRNATIP is the only algorithm producing a positive MCC, confirming the its goodness of the overall predictive power.

It is important to notice that although we predicted the largest number of interactions with respect to the other methods for both species, we obtained a fair specificity score, with regards to the other predictors, and the best sensitivity score. That that means our method could predict more potentially true miRNA-mRNA interactions than the other algorithms.

All the predicted interaction are freely available at the following url: http://tblab.pa.icar.cnr.it/public/miRNATIP/.

Conclusion and future work

The interaction between miRNA and mRNA is of fundamental importance in the post-transcriptional regulatory mechanism. In this paper we presented miRNATIP, a SOM-based predictor for the identification of miRNA-target interactions. MiRNATIP simulates the main features of the miRNA-mRNA interaction, including near perfect seed pairing, the presence of bulges, free energy constraints for stability of the duplex. In particular a SOM is trained considering only the miRNA seeds (first 8 nucleotides), that are represented by means of a numerical encoding derived from as PWM, and then 8-mer mRNA fragments are projected over the trained lattice in order to identify a preliminary list of putative interactions. Then that list is filtered out taking into account the distance between the remaining parts of the miRNA and mRNA sequences and the free energy values. The obtained predictions, for cel and hsa species, have been validated in terms of sensitivity and specificity scores against six other state-of-the-art predictors (miRanda, PITA, DIANA-microT, mirSOM, PicTar, TargetScan) with regards to a manually curated dataset of both validated miRNA-mRNA interactions and validated non-target interactions. Results demonstrated that our methods reached the best sensitivity score for both species and a specificity score consistent with the other predictors, even if we produced the largest number of putative interactions. As future work, we are going to test our method with other species, and at the same time we

Table 5 Performances of prediction algorithms related to validated interactions in Table 4

Algorithm	ACC %	PPV %	TPR %	TNR %	FNR %	FPR %	F1 %	MCC
PITA	5.3	99.3	5.0	88.6	94.9	11.4	9.6	–0.01617
MiRanda	25.5	99.5	25.4	59.3	74.5	40.6	40.5	–0.01946
TargetScan	11.4	99.3	11.2	78.0	88.8	21.9	20.0	–0.01912
Pictar	7.1	99.1	6.9	81.3	93.1	18.7	12.9	–0.02581
DIANA-microT	20.1	99.5	19.9	74.0	80.0	26.0	33.2	–0.00847
miRNATIP	30.6	99.7	30.5	69.9	69.4	30.1	46.7	0.00055

Statistical measures reported in this table are accuracy (*ACC*), precision (*PPV*), sensitivity (*TPR*), specificity (*TNR*), miss-rate (*FPR*), F1-measure (*F1*) and Matthews correlation (*MCC*), respectively

will provide a web service that will allow to download already computed predictions or to test our algorithm with customized sets of miRNA and/or mRNA sequences.

Declaration

The publication costs for this article were funded by the CNR Interomics Flagship Project "- Development of an integrated platform for the application of "omic" sciences to biomarker definition and theranostic, predictive and diagnostic profiles". This article has been published as part of BMC Bioinformatics Volume 17 Supplement 11, 2016. Selected articles from the 11th International Meeting on Computational Intelligence Methods for Bioinformatics and Biostatistics (CIBB 2014). The full contents of the supplement are available online https://bmcbioinformatics.biomedcentral.com/articles/supplements/volume-17-supplement-11.

Availability of data and materials

All the predicted interaction are freely available at the following url: http://tblab.pa.icar.cnr.it/public/miRNATIP/.

Authors' contributions

AF: project conception, implementation, experimental tests, discussions, assessment, writing. MLR: project conception, experimental tests, writing, assessment, discussions. LLP: project conception, writing, assessment, discussions. RR: project conception, discussions, assessment, writing. AU: project conception, discussions, assessment, writing, funding. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

Published: 22 September 2016

References

1. Ambros V. The functions of animal microRNAs. Nature. 2004;431(7006):350–5. doi:10.1038/nature02871.
2. Kloosterman WP, Plasterk RHA. The Diverse Functions of MicroRNAs in Animal Development and Disease. 2006. doi:10.1016/j.devcel.2006.09.009.
3. Farazi TA, Horlings HM, Ten Hoeve JJ, Mihailovic A, Halfwerk H, Morozov P, et al. MicroRNA sequence and expression analysis in breast tumors by deep sequencing. Cancer Res. 2011;71(13):4443–53. doi:10.1158/0008-5472.CAN-11-0608.
4. Wotschofsky Z, Gummlich L, Liep J, Stephan C, Kilic E, Jung K, Billaud J-N, Meyer H-A. Integrated microRNA and mRNA Signature Associated with the Transition from the Locally Confined to the Metastasized Clear Cell Renal Cell Carcinoma Exemplified by miR-146-5p. PLoS ONE. 2016;11(2):0148746. doi:10.1371/journal.pone.0148746.
5. Lin X, Yang B, Liu W, Tan X, Wu F, Hu P, et al. Interplay between PCBP2 and miRNA modulates ARHGDIA expression and function in glioma migration and invasion. Oncotarget. 2016. doi:10.18632/oncotarget.6869.
6. Lee RC, Feinbaum RL, Ambros V. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell. 1993;75(5):843–54. doi:10.1016/0092-8674(93)90529-Y.
7. Amirkhah R, Schmitz U, Linnebacher M, Wolkenhauer O, Farazmand A. MicroRNA-mRNA interactions in colorectal cancer and their role in tumor progression. Gene Chromosome Cancer. 2015;54(3):129–41. doi:10.1002/gcc.22231.
8. O'Day E, Lal A. MicroRNAs and their target gene networks in breast cancer. Breast Cancer Res BCR. 2010;12(2):201. doi:10.1186/bcr2484.
9. Fiannaca A, La Rosa M, La Paglia L, Rizzo R, Urso A. Analysis of miRNA expression profiles in breast cancer using biclustering. BMC Bioinforma. 2015;16(Suppl 4):7. doi:10.1186/1471-2105-16-S4-S7.
10. Ma L, Teruya-Feldstein J, Weinberg RA. Tumour invasion and metastasis initiated by microRNA-10b in breast cancer. Nature. 2007;449(7163):682–8. doi:10.1038/nature07316.
11. Li N, Fu H, Tie Y, Hu Z, Kong W, Wu Y, et al. miR-34a inhibits migration and invasion by down-regulation of c-Met expression in human hepatocellular carcinoma cells. Cancer Lett. 2009;275(1):44–53. doi:10.1016/j.canlet.2008.09.035.
12. Filipowicz W, Bhattacharyya SN, Sonenberg N. Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? Nat Rev Genet. 2008;9(2):102–14. doi:10.1038/nrg2290.
13. Kim SK, Nam JW, Rhee JK, Lee WJ, Zhang BT. miTarget: microRNA target gene prediction using a support vector machine. BMC Bioinforma. 2006;7:411. doi:10.1186/1471-2105-7-411.
14. Mendoza MR, da Fonseca GC, Loss-Morais G, Alves R, Margis R, Bazzan ALC. RFMirTarget: Predicting Human MicroRNA Target Genes with a Random Forest Classifier. PLoS ONE. 2013;8(7). doi:10.1371/journal.pone.0070153.
15. Agarwal S, Vaz C, Bhattacharya A, Srinivasan A. Prediction of novel precursor miRNAs using a context-sensitive hidden Markov model (CSHMM). BMC Bioinforma. 2010;11 Suppl 1:29. doi:10.1186/1471-2105-11-S1-S29.
16. Kohonen T. Self-Organizing Maps, vol. 30. Berlin Heidelberg: Springer; 2001, p. 501. doi:10.1007/978-3-642-56927-2.
17. Fiannaca A, Di Fatta G, Rizzo R, Urso A, Gaglio S. Simulated annealing technique for fast learning of som networks. Neural Comput & Applic. 2013;22(5):889–99. doi:10.1007/s00521-011-0780-6.
18. Peterson SM, Thompson JA, Ufkin ML, Sathyanarayana P, Liaw L, Congdon CB. Common features of microRNA target prediction tools. Front Genet. 2014;5. doi:10.3389/fgene.2014.00023.
19. Kohonen T, Somervuo P. How to make large self-organizing maps for nonvectorial data. Neural Netw. 2002;15(8–9):945–52. doi:10.1016/S0893-6080(02)00069-2.
20. Di Fatta G, Fiannaca A, Rizzo R, Urso A, Berthold MR, Gaglio S. Context-Aware Visual Exploration of Molecular Datab. In: Sixth IEEE Int Conf Data Min - Workshops (ICDMW'06); 2006. doi:10.1109/ICDMW.2006.51.
21. Mahony S, McInerney JO, Smith TJ, Golden A. Gene prediction using the Self-Organizing Map: automatic generation of multiple gene models. BMC Bioinforma. 2004;5:23. doi:10.1186/1471-2105-5-23.
22. Mahony S, Hendrix D, Golden A, Smith TJ, Rokhsar DS. Transcription factor binding site identification using the self-organizing map. Bioinformatics. 2005;21(9):1807–14. doi:10.1093/bioinformatics/bti256.
23. Sethupathy P, Corda B, Hatzigeorgiou AG. TarBase: A comprehensive database of experimentally supported animal microRNA targets. RNA (New York). 2006;12(2):192–7. doi:10.1261/rna.2239606.
24. Papadopoulos GL, Reczko M, Simossis VA, Sethupathy P, Hatzigeorgiou AG. The database of experimentally supported targets: A functional update of TarBase. Nucleic Acids Res. 2009;37(SUPPL. 1). doi:10.1093/nar/gkn809.
25. Yue D, Liu H, Huang Y. Survey of Computational Algorithms for MicroRNA Target Prediction. Curr Genomics. 2009;10(7):478–92. doi:10.2174/138920209789208219.
26. Witkos TM, Koscianska E, Krzyzosiak WJ. Practical Aspects of microRNA Target Prediction. Curr Mol Med. 2011;11(2):93–109. doi:10.2174/156652411794859250.
27. John B, Enright AJ, Aravin A, Tuschl T, Sander C, Marks DS. Human microRNA targets. PLoS Biol. 2004;2(11). doi:10.1371/journal.pbio.0020363.
28. Lall S, Grün D, Krek A, Chen K, Wang YL, Dewey CN, et al. A genome-wide map of conserved MicroRNA targets in C. elegans. Curr Biol. 2006;16(5):460–71. doi:10.1016/j.cub.2006.01.050.
29. Lewis BP, Burge CB, Bartel DP. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell. 2005;120(1):15–20. doi:10.1016/j.cell.2004.12.035.
30. Kiriakidou M, Nelson PT, Kouranov A, Fitziev P, Bouyioukos C, Mourelatos Z, et al. A combined computational-experimental approach predicts human microRNA targets. Genes Dev. 2004;18(10):1165–78. doi:10.1101/gad.1184704.
31. Kertesz M, Iovino N, Unnerstall U, Gaul U, Segal E. The role of site accessibility in microRNA target recognition. Nat Genet. 2007;39(10):1278–84. doi:10.1038/ng2135.
32. Krüger J, Rehmsmeier M. RNAhybrid: MicroRNA target prediction easy, fast and flexible. Nucleic Acids Res. 2006;34(WEB. SERV. ISS.). doi:10.1093/nar/gkl243.
33. Lesnik EA, Freier SM. Relative Thermodynamic Stability of DNA, RNA, and DNA:RNA Hybrid Duplexes: Relationship with Base Composition and Structure. Biochemistry. 1995;34(34):10807–15. doi:10.1021/bi00034a013.

34. Heikkinen L, Kolehmainen M, Wong G. Prediction of microRNA targets in Caenorhabditis elegans using a self-organizing map. Bioinformatics. 2011;27(9):1247–54. doi:10.1093/bioinformatics/btr144.
35. Hannenhalli S, Wang LS. Enhanced position weight matrices using mixture models. Bioinformatics. 2005;21(SUPPL. 1). doi:10.1093/bioinformatics/bti1001.
36. Orenstein Y, Linhart C, Shamir R. Assessment of Algorithms for Inferring Positional Weight Matrix Motifs of Transcription Factor Binding Sites Using Protein Binding Microarray Data. PLoS ONE. 2012;7(9). doi:10.1371/journal.pone.0046145.
37. Linhart C, Halperin Y, Shamir R. Transcription factor and microRNA motif discovery: The Amadeus platform and a compendium of metazoan target sets. Genome Res. 2008;18(7):1180–9. doi:10.1101/gr.076117.108.
38. Attik M, Bougrain L, Alexandre F. Self-organizing map initialization In: Duch W, Janusz K, Erkki O, Slawomir Z, editors. Artificial Neural Networks: Biological Inspirations - ICANN 2005. Lecture Notes in Computer Science, vol. 3696 LNCS. Berlin Heidelberg: Springer; 2005. p. 357–62. doi:10.1007/11550822_56.
39. Busch A, Richter AS, Backofen R. IntaRNA: Efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions. Bioinformatics. 2008;24(24):2849–856. doi:10.1093/bioinformatics/btn544.
40. Wright PR, Georg J, Mann M, Sorescu DA, Richter AS, Lott S, et al. CopraRNA and IntaRNA: Predicting small RNA targets, networks and interaction domains. Nucleic Acids Res. 2014;42(W1). doi:10.1093/nar/gku359.
41. Kozomara A, Griffiths-Jones S. MiRBase: Integrating microRNA annotation and deep-sequencing data. Nucleic Acids Res. 2011;39(SUPPL. 1). doi:10.1093/nar/gkq1027.
42. Howe K, Davis P, Paulini M, Tuli MA, Williams G, Yook K, et al. WormBase: Annotating many nematode genomes. 2012. doi:10.4161/worm.19574.
43. Kasprzyk A. BioMart: Driving a paradigm change in biological data management. Database. 2011;2011. doi:10.1093/database/bar049.
44. Cunningham F, Amode MR, Barrell D, Beal K, Billis K, Brent S, et al. Ensembl 2015. Nucleic Acids Res. 2015;43(D1):662–9. doi:10.1093/nar/gku1010.
45. Hsu SD, Tseng YT, Shrestha S, Lin YL, Khaleel A, Chou CH, et al. MiRTarBase update 2014: An information resource for experimentally validated miRNA-target interactions. Nucleic Acids Res. 2014;42(D1). doi:10.1093/nar/gkt1266.
46. Vlachos IS, Paraskevopoulou MD, Karagkouni D, Georgakilas G, Vergoulis T, Kanellos I, et al. DIANA-TarBase v7.0: indexing more than half a million experimentally supported miRNA:mRNA interactions. Nucleic Acids Res. 2014;43(D1):153–9. doi:10.1093/nar/gku1215.
47. Kiviluoto K. Topology preservation in self-organizing maps. In: Proceedings of International Conference on Neural Networks (ICNN'96); 1996. doi:10.1109/ICNN.1996.548907.
48. Shirdel EA, Xie W, Mak TW, Jurisica I. NAViGaTing the Micronome–Using Multiple MicroRNA Prediction Databases to Identify Signalling Pathway-Associated MicroRNAs. PLoS ONE. 2011;6(2):17429. doi:10.1371/journal.pone.0017429.
49. Vihinen M. How to evaluate performance of prediction methods? Measures and their interpretation in variation effect analysis. BMC Genomics. 2012;13(Suppl 4):2. doi:10.1186/1471-2164-13-S4-S2.