MicroRNA discovery in the human parasite *Echinococcus multilocularis* from genome-wide data

L. Kamenetzky[1]*, G. Stegmayer[2] , L. Maldonado[1] , N. Macchiaroli[1] ,C. Yones[2] and D.H. Milone[2]
1 IMPAM-UBA-CONICET, Facultad de Medicina - Buenos Aires (Argentina)
2 sinc(i)-FICH-UNL-CONICET, Ciudad Universitaria - Santa Fe (Argentina)

∗ To whom correspondence should be addressed

Abstract

The cestode parasite *Echinococcus multilocularis* is the aetiological agent of alveolar echinococcosis, responsible for considerable human morbidity and mortality. This disease is a worldwide zoonosis of major public health concern and is considered a neglected disease by the World Health Organization. The complete genome of *E. multilocularis* has been recently sequenced and assembled in a collaborative effort between the Wellcome Trust Sanger Institute and our group, with the main aim of analyzing protein-coding genes. These analyses suggested that approximately 10% of *E. multilocularis* genome is composed of protein-coding regions. This shows there is still a vast proportion of the genome that needs to be explored, including non-coding RNAs such as small RNAs (sRNAs). Within this class of small regulatory RNAs, microRNAs (miRNAs) can be found, which have been identified in many different organisms ranging from viruses to higher eukaryotes. MiRNAs are a key regulation mechanism of gene expression at post-ranscriptional level and play important roles in biological processes such as development, proliferation, cell differentiation and metabolism in animals and plants. In spite of this, identification of miRNAs directly from genome-wide data only is still a very challenging task. There are many miRNAs that remain unidentified due to the lack of either sequence information of particular phylums or appropiate algorithms to identify novel miRNAs. The motivation for this work is the discovery of new miRNAs in *E. multilocularis* based on non-target genomic data only, in order to obtain useful information from the currently available unexplored data. In this work, we present the discovery of new pre-miRNAs in the *E. multilocularis* genome through a novel approach based on machine learning. We have extracted the most commonly used structural features from the folded sequences of the parasite genome: triplets, minimum free energy and sequence length. These features have been used to train a novel deep architecture of self-organizing maps (SOMs). This model can be trained with a high class imbalance and without the artificial definition of a negative class. We discovered 886 pre-miRNA candidates within the *E. multilocularis* genome-wide data. After that, experimental validation by small RNA-seq analysis clearly showed 23 pre-miRNA candidates with a pattern compatible with miRNA biogenesis, indicating them as high confidence miRNAs. We discovered new pre-miRNA candidates in *E. multilocularis* using non-target genomic data only. Predictions were meaningful using only sequence data, with no need of RNA-seq data or target analysis for prediction. Furthermore, the methodology employed can be easily adapted and applied on any draft genomes, which are actually the most interesting ones since most non-model organisms have this kind of status and carry real biological and sanitary relevance.

Availability

Web demo: http://fich.unl.edu.ar/sinc/web-demo/mirna-som/
Source code: http://sourceforge.net/projects/sourcesinc/files/mirnasom/

1.Introduction

1.1 MicroRNAs in Echinococcus spp.

Echinococcus multilocularis is a parasitic flatworm that causes human alveolar echinococcosis worldwide. It is amongst the world's most dangerous zoonoses, developing tumor-like flatworm larvae growing in the body (Torgerson *et al*., 2010). The metacestode of this parasite can grow in an aggressive manner budding exogenously, infiltrating and colonizing surrounding and distant tissues due to the metastatic nature of its germinative cells. The genome of *E. multilocularis* was recently sequenced and assembled in a collaborative effort between the Wellcome Trust Sanger Institute and our group (Tsai *et al.*, 2013). Gene content analysis revealed that approximately 10% of the genome are protein-coding regions (Cucher *et al*., 2015). This shows that there is still a vast proportion of the genome that needs to be explored, including non-coding RNAs such as small RNAs (sRNAs).

Within this class of small regulatory RNAs, microRNAs (miRNAs) have been identified in many different organisms. MiRNAs are endogenous ∼22 nucleotide noncoding RNAs, which act as pos-transcriptional regulators involved in the control of nearly all cellular pathways, from development to diseases in animals and plants (Ameres and Zamore, 2013). MiRNAs act mainly silencing gene expression by binding to complementary sequences in the 3' untranslated regions (UTRs) of their target mRNAs. Animal miRNAs are processed in the nucleus from long primary RNA transcripts (pri-miRNAs) into ∼70 nt long stem loop intermediates, known as miRNA precursors (pre-miRNAs), from which mature miRNAs are processed in the cytoplasm (Bartel, 2004). Pre-miRNAs (also known as hairpins) generated during biogenesis have well-known RNA secondary structures derived from primary structures that have allowed the development of computational algorithms for their identification. In a previous report, we experimentally found that miRNAs are expressed in *Echinococcus granulosus sensu lato* (Cucher *et al*., 2011), a species closely related to *E. multilocularis*, suggesting that these small RNAs could be an essential mechanism of gene regulation in this genus. Profiling of miRNAs can be defined as the assessment of miRNA expression in a given cell type and condition (Pritchard et al., 2012). Several methods are available to do this, and are preferentially used depending on a wide range of factors. The most important considerations tend to be related to the amount of biological material available, the experimental design and the final objectives of the study. As with model organisms, this kind of experiments is time-consuming and depends on the expression level of each biological stage. With the advent of new sequencing technologies, it is faster and easier to obtain genomic sequences from new organisms. However, only a few bioinformatics efforts are available to analyze this type of data, which, on the other hand, provide limited capabilities and low prediction performance for non-model organisms. To the best of our knowledge, no miRNA discovery studies from *E. multilocularis* genome wide data have been carried out to date. Thus, knowledge of the *E. multilocularis* miRNA repertoire needs to be explored.

1.2 Tools for miRNA identification

MiRNAs can be identified either by bioinformatics approaches or by sequencing strategies, both of which need computational tools for the analysis of the sequences obtained. Some of the oldest strategies for miRNAs discovery includes RNA conformation based approaches using Mfold (Zuker *et al*., 2003) and RNAfold (Hofacker *et al*., 2003; Hofacker *et al*., 1994; Jacobson *et al*., 1993) as core algorithms. Other approaches are based on homology methods using known miRNA and pre-miRNA sequences from several well-known model organisms. One potential drawback of these homology-based methods is their inability to identify completely novel miRNA sequences in non-model genomes, precisely due to the conservation criteria between related genomes on which they rely and that might not be true or known for brand-new recently sequenced genomes. More recently, machine-learning techniques for miRNA prediction have been proposed, based on properties and features of well-known miRNAs. Among them, mainly supervised machine-learning techniques have been employed, using sequence composition and structural conformation features to train a learning system capable of identifying miRNA candidates (Saetrom *et al*., 2007; Wen-Ching Chan *et al*., 2012). As opposed to homology based methods, this approach could be useful for species-specific miRNA discovery since it does not depend on evolutionary conservation. As mentioned above, many methods have been developed to predict pre-miRNA loci based on the genome sequence and structural properties of the candidate loci. The miRNA classifier methods use different features to evaluate, for example, the structural stability or sequence properties of the candidates, in order to produce a final prediction (Li *et al*., 2010; Liu *et al*.; 2014, Lopes *et al*.;

2014). However, this is a non-trivial problem when addressing it in a purely computerized way, in particular with classical supervised learning because the artificial definition of a negative class is required (Gomes *et al*., 2013). Although methods that use only positive samples to predict new miRNAs have been described (Yousef *et al*., 2008), it is well known that, when the negative class is complex, these methods fail because they do not model these regions of the feature space appropriately. Actually, they do not model the negative class at all or they model it under very simplified assumptions. Furthermore, when the negative class is not artificially defined and genome-wide data wants to be used, a huge imbalance is often present between the positive class (a few known miRNAs) and the unlabeled data (hundreds of thousands of sequences). Since *E. multilocularis* genome was recently generated, mining this new genomic data will provide a deeper understanding of parasite miRNome. In this work, we identify candidate novel miRNA precursors in *E. multilocularis* through a novel approach based on self-organizing maps (SOM) (Kohonen *et. al.,* 2005; Milone *et. al.,* 2010).

2. Materials and methods

2.1 Biologically relevant data set and hairpin features extraction

The main pipeline used for the analysis of the genome-wide data is presented in Figure 1. The complete *E. multilocularis* genome (Tsai *et al*., 2013) was processed by Einverted software (EMBOSS package) as described by de Souza Gomes *et al*. (2011) with the following parameters: gap penalty 6, minimum score threshold 25, match score 3, mismatch score -3, maximum separation between the start and end of the inverted repeat 95. Then, the inverted repeats were folded into 491532 sequences by RNAfold (Supp. file 1). The obtained sequences were then pre-processed. Sequences with minimum free energy (MFE) threshold of -20 and single-loop folded sequences were selected according to the miRNA biogenesis model (Bartel, 2004). The retained sequences were analyzed using BLAST algorithm (Altschul *et al*., 1990) against an in-house database of CDS, tRNAs, rRNAs and long non coding RNAs flatworm sequences (Cucher *et al*., 2015). After this, 77429 sequences were retained. Then, all *E. multilocularis* hairpin sequences were downloaded from miRBase v21, BLAST searches among the 77429 sequences retained were performed and a total of 18 sequences were labeled as positive class. To represent the sequences, the 34 most commonly used features were extracted. We used the smallest and less costly to compute subset of features that are extensively used nowadays to identify novel pre-miRNAs : 32 triplets (Xue *et al*., 2005), sequence length and MFE (Lopes *et al*., 2014). These features were extracted with the web tool miRNAfe (Yones *et al*., 2015) recently developed by us. Then, the features extracted from 77429 sequences were used to train the SOM classifier, which identified 886 sequences as the best pre-miRNA candidates.

2.2 Classifier

In this work, instead of training a classifier in a classical supervised manner, we identified miRNA precursors with a novel approach based on several nested SOMs. For SOM training, there is no need to define the negative miRNA class. Only some examples of positive class examples (well-known pre-miRNAs) are needed to identify the neurons that have the best miRNA candidates associated to them. In this context, each neuron in the SOM is a cluster of sequences. The SOM classifier is actually composed of several nested SOMs, which are hierarchically related. This deep architecture is shown at the top of Figure 2, where a 10-layered (h=10) example is provided. The training process of the hierarchical maps starts with the root SOM on the first layer (left), with the 77429 sequences as input. This map undergoes standard training. After that, all the sequences grouped together in a neuron (cluster) having also well-known pre-miRNAs (painted in dark blue) are labeled as highly likely pre-miRNA candidates. These sequences are chosen as input to train the map in the following layer (indicated with black lines). This process is repeated several times, further refining the classifier level after level. With this approach, each internal map is trained with only a portion of the input data: the data mapped in the pre-miRNA clusters in the previous layer. At the bottom of Figure 2, the number of candidates is shown for each level of the SOM. It can be clearly seen here that this method significantly reduces the number of possible pre-miRNA candidates, level after level, retaining at last the high-confidence pre-miRNAs. After four consecutive levels without changes in the number of data clustered into pre-miRNA neurons (8042 sequences), no more levels are added. These and the following levels are exactly the same since the map is trained with exactly the same data. Therefore, adding more levels does not cause over-training either. In the last level, each well-known pre-miRNA in the miRNA neurons (in blue) is grouped together with unlabeled sequences. Among them, the best bona fide candidates are selected (886) as those having feature values within ranges automatically defined by rules obtained according to the positive class (well-known miRNAs). This reduction was possible because each feature was evaluated individually with respect to its discriminative power for separating the positive class (well-known miRNAs) from the rest of the sequences. This was done iteratively, until all features were analyzed and all positive sequences were correctly classified. This way, several rules for the feature ranges were extracted, which were applied to the 8042 sequences in order to further reduce its number to 886.

2.3 Mature miRNA sequence extraction

The total number of candidate pre-miRNAs discovered by SOM analysis (886) was mapped to the complete *E. multilocularis* genome and sequences with more than 10 hits were removed (highly repetitive sequences, Figure 1). Then, in order to extract mature miRNA sequences from pre-miRNAs retained in the previous step, 26.9 million clean mapped reads from small RNA-seq data of *E. multilocularis* metacestode stage retrieved from Cucher *et al*. (2015) were BLAST searched against the pre-miRNAs sequences. BLAST algorithm was optimized for small sequences with word size set in 7, the filter for low complexity regions off, and an e-value set in 10. For each pre-miRNA with small RNAseq evidence in the stem region of the candidate pre-miRNA, the consensus mature sequence was extracted from alignments showing 100% of identity and 100% of coverage. This data was used for mature miRNA sequence determination and not for miRNA expression quantification. In order to extract additional mature miRNA sequences, all metazoan mature miRNA sequences from miRBase 21 and *Echinococcus* mature miRNAs reported in the literature that were not integrated in miRBase (Bai *et al*., 2014, Macchiaroli *et al*., 2015) were analyzed by BLAST and SSEARCH algorithms against candidate pre-miRNAs. Finally, for conservation analysis, all *E. multilocularis* mature sequences identified in previous steps were BLAST searched against related flatworm genomes: *Echinococcus granulosus*, *Echinococcus canadensis*, *Hymenolepis microstoma* and *Taenia solium*. The genomes were downloaded from http://parasite.wormbase.org/index.html and processed as previously described for *E. multilocularis* whole genome.

2.4 Further evaluation of the approach in a model organism

In order to further evaluate the proposal, a model organism has been used. *Caenorhabditis elegans* genome was processed in a similar way as previously described for E. multilocularis. *The* 1,739,460 sequences obtained were BLAST matched against miRBase v17 for pre-miRNA identification. A total of 200 well-known miRNAs of *C. elegans* included into miRBase v17 were labeled as positive class. All genome data (including the identified positive class) were used to train SOM until the level where the number of candidates did not change (as described previously for *E. multilocularis*). In order to evaluate the prediction performance of new miRNAs in a model organism, the miRNAs added to miRBase in its most recent version have been used as input test sequences. Therefore, the trained SOM was tested with 48 *C. elegans* pre-miRNA obtained from miRBase v19 to v21 (absent in miRBase v17).

3. Results and Discussion

In this work, we discovered 886 pre-miRNA candidates from *E. multilocularis* genome-wide data (Figure 1). Although such quantity can be hard to validate experimentally, this must be interpreted as an important first step towards the discovery of new miRNAs in low explored genomes, such as the *E. multilocularis* one, where only few pre-miRNA sequences are available. Computationally identified miRNAs suggests that miRNA gene numbers are substantially higher than those currently known, as proposed by Piriyapongsa et al. (2007). Most computational methods nowadays require expensive high-throughput RNA sequencing data as input (Friedlander *et al*., 2012, Hackenberg *et al.*, 2011). However, we use NGS data only for validation after finding the pre-miRNA candidates, as in (Saçar *et al*., 2014). The few methods that have been proposed to identify miRNAs from a complete genome without such data obtain a very high number of initial candidates, hundreds of thousands or tens of thousands of sequences (Mendes *et al*., 2010). After that, a reduced list of the best candidates is obtained by manually applying ad hoc rules (Mendes *et al*., 2012) in order to achieve a number of sequences that can be experimentally validated. However, for miRNA prediction most of the published approaches do not really deal with genome-wide data but with class and no-class data (Xue *et al*., 2005; Hertel *et al.*, 2006; Huang *et al.*, 2007; Jiang *et al*., 2007; Xu *et al*., 2008; Gkirtzou *et al*., 2010; Ding *et al*., 2010; Rahman *et al*., 2012; Gudy *et al*., 2013). In these works, in order to train classifiers, and measure sensitivity and specificity in a cross-validation scheme, a reduced subset of negative examples must be artificially defined. Moreover, these unrealistic tests are performed over the genomes of model organisms, such as mammals or round worms, being only useful to precisely measure the performance in cross-validation experiments, but they cannot be applied in real practical scenarios. In the proposed processing pipeline, only obvious non-miRNA sequences are filtered (according to loops, energy threshold and identity to known RNAs other than miRNAs). The remaining sequences from the original genome are all presented to the SOM for training and classification. The first advantage here is that the SOM does not require the artificial definition of negative class, thus it does not perform unrealistic tests. The second advantage is that it works directly on complete genome-wide data, which is being refined level after level, automatically discarding low-quality candidates. With this methodology, artificial examples to represent the negative class (which is actually unknown) must not be defined. The negative examples can be actually very hard to define, even for a model genome (Wei et al., 2014). Thus, SOM is well suited to the analysis of genome data from novel non model organisms.

In order to classify each miRNA as conserved or novel, we analyzed the identity of all pre-miRNA candidates discovered by SOM with already reported metazoan miRNAs (miRBase v21) and *E. multilocularis* miRNAs (Cucher *et al.*, 2015). This analysis allowed us to identify 13 pre-miRNAs previously described (Supplementary Table S1). Taking into account the 18 miRNAs used as positive class, the total of miRNAs found was 31 out of 37 miRNAs expected to be in *Echinococcus multilocularis* (Cucher *et al.*, 2015). Since four miRNAs were absent in the genome input dataset because their folded structure did not match the filter criteria employed, the sensitivity of SOM reached 94% (31/33). Moreover, 10 new pre-miRNAs were also identified totaling 23 pre-miRNAs. The mature miRNA annotation, their clean mapped read counts and the biological function in other organisms are shown in Table 1. *E. multilocularis* RNA-seq clearly mapped to the hairpin stem region with a pattern compatible with miRNA biogenesis indicating them as high-confidence miRNAs. As an example, a schematic representation of the secondary structure from the conserved *E. multilocularis* premiRNA 36b is shown in Figure 3.

These new pre-miRNAs represent, in the first place, flatworm-specific miRNAs since they were not detected in any other phyla. Also, some of them were recently reported in *E. granulosus* (Bai *et al.*, 2014). It can be noticed here the ability of the SOM to discover of new miRNAs, only with genomic data as input. Furthermore, the secondary structure from all new pre-miRNAs discovered by SOM analysis is shown in Figure 4. Structural features such as MFE and mature miRNA sequences that mapped to them clearly showed that they were bona fide pre-miRNAs. All mature and pre-miRNA sequences and structures are available in Supplementary Table S1 and Figure S1. Additionally, our method discovered miRNAs in *E. multilocularis* that were not identified by a recent bioinformatics approach (Jin *et al*., 2013) such as miR-36, miR-307, miR-1992, mir-3479, highlighting the potential of SOM analysis for miRNA discovery. Interestingly, this miRNAs were considered lost in *Echinococcus* (Fromm *et al*., 2013) but SOM discovered them in coincidence with previously reports (Cucher *et al.*, 2015; Macchiaroli *et al.*, 2015).

We have also searched for these 23 pre-miRNA sequences in closely related flatworm genomes. All of them were found in at least one of the four related flatworm species (Figure 3, Supplementary Table S1). Several of the mature miRNAs found in this work are deeply conserved among bilateria such as emu-miR-281 and emu-miR-31, but others are found only in protostomia such as emu-bantam, emu-miR-36 and emu-miR-1992. So far, there is no information about the biological function of these miRNAs in *Echinococcus*. These results could be interpreted as a good indicator of the biological confidence of the predictions obtained with the pipeline proposed in this work, and indicate that the SOM could discover both conserved and novel miRNAs from *E. multilocularis* genome data. Although losses of conserved miRNAs have been previously proposed in parasite flatworms (Fromm *et al.*, 2013; Macchiaroli *et al.*, 2015), the presence of specific miRNAs is expected since novel miRNAs have been recently reported from small RNAseq data in other helminth parasites (Winter *et al.*,2012; Bai *et al.*, 2014). The new pre-miRNA sequences discovered in our work are good candidates to be flatworm-specific miRNAs since they have no identity with miRNAs from other phyla. These miRNA sequences are the most interesting ones because they could have a crucial role in the establishment and/or progression of human alveolar echinococcosis. As future work, it could be interesting to be able to determine the *E. multilocularis* life cycle stage where the new miRNAs discovered in this work are expressed which could be done following approaches previously published by us (Macchiaroli *et al*. 2015). The knowledge of the complete repertoire of miRNAs, conserved and specific ones, is key to understand the development of the parasite and the progression and control of this neglected disease.

The validation of the proposed methodology in a non-model organism has proved its effectiveness. However, benchmarking it in a well-known reference genome can provide evidence of its utility in a wide number of organisms. Thus, we have performed a benchmarking test of the proposed SOM approach with a well- known reference genome. The SOM was trained with the complete genome data plus a total of 200 *C. elegans well-known* pre-miRNA sequences present in miRBase v17. Then, the trained SOM has been tested with 48 pre-miRNAs more recently added to miRBase v18-21 and absent in v17. In this test, 44 out of 48 pre-miRNA have been identified as positive class, resulting in a SOM sensitivity of 92%. Results are available at http://fich.unl.edu.ar/sinc/blog/web-demo/mirna-som-ce/.

Table 1: Conserved and novel *Echinococcus multilocularis* microRNAs predicted from whole genome data.

MiRNA ID	Read counts[a]	Biological function[b]	Reference[b]
emu-bantam-3p	1184581	Regulates the growth of dendrites in sensory neurons of Drosophila melanogaster epithelial cells. Present only in protostomes	Parrish et al. (2009)
emu-miR-31-5p	88	Tumoursuppressor in humans	O'Day et al. (2010)
emu-miR-36a-3p	617	Unknown, present only in protostomes	Macchiaroli et al. (2015)
emu-miR-36b-3p	1075	Unknown, present only in protostomes	Cucher et al. (2015)
emu-miR-61-3p	578860	Promotes development in Caenorhabditis elegans. Present only in protostomes	Yoo AS et al. (2005)
emu-miR-281-3p	17958	Enhance viral replication in Aedes albopictus	Zhou et al. (2014)
emu-mir-307-3p	123277	Unknown, present only in protostomes	Cucher et al. (2015)
emu-miR-1992-3p	24	Unknown, present only in protostomes	Cucher et al. (2015)

emu-miR-2162-3p	100642	Unknown, present only in protostomes	Cucher et al. (2015)
emu-miR-10293-3p	4017	Unknown	Cucher et al. (2015)
emu-miR-3479a-3p	56603	Unknown	Cucher et al. (2015)
emu-miR-3479b-3p	63552	Unknown	Cucher et al. (2015)
emu-miR-7b-5p	1070	Controls epidermal growth factor receptor signaling and promotes photoreceptor cell differentiation in Drosophila	Jiang et al. (2010); Macchiaroli et al. (2015) (egr-miR-7b-5p)
emu-miR-new1-5p	8	Unknown	This work and Bai et al. (2014) (egr-new-48)
emu_miR-new2-3p	32	Unknown	This work
emu_miR-new3-5p	123	Unknown	This work
emu_miR-new4-5p	58	Unknown	This work and Bai et al. (2014) (egr-new -12)
emu_miR-new5-3p	1	Unknown	This work and Bai et al. (2014) (egr-new-25)
emu_miR-new6-5p	1	Unknown	This work and Bai et al. (2014) (egr-new-114)
emu-miR-new7-5p	41	Unknown	This work and Bai et al. (2014) (egr-new-7)
emu-miR-new8-3p	20	Unknown	This work and Bai et al. (2014) (egr-new-24)
emu-miR-new9-3p	246	Unknown	This work
emu-miR-new10-5p	231	Unknown	This work and Bai et al. (2014) (egr-new-29)
Total	2133125		

[a]Number of clean mapped reads without normalization. [b]Described in model species.
[c]Most relevant references for miRNA function in other organisms or studies on related *Echinococcus* species.

3.Conclusions

We applied SOM analysis for *E. multilocularis* miRNA prediction and demonstrated its effectiveness and usefulness. Although using purely computational methods for de novo miRNA prediction was a real challenge and a difficult problem to address, this analysis allow us to discover good candidates from *E. multilocularis* genome sequencing data. Most pre-miRNA prediction methods based on supervised machine learning methods, which need to artificially define the negative class, cannot handle the class imbalance existing in such genome-wide data. However, the proposed method addressed the problem effectively without requiring the artificial definition of a negative class dataset. With this approach, complete genomes containing thousands of hairpins sequences could be analyzed and only highly likely hairpin sequences can be further selected for biological validation. We found novel *E. multilocularis* pre-miRNAs from non target genomic data without the need of RNA-seq data and all of them conserved in at least one related flatworm species. These results clearly indicate that there are still several genomic sequences to be classified and ready to be analyzed deeply. We found expression of mature miRNAs derived from pre-miRNA candidates adding confidence to the predictions obtained by SOM analysis. The data obtained in this work will be useful to search for new mature miRNAs expressed in the human parasite *E. multilocularis* resulting in new tools for the diagnosis, prevention and developmental regulation of alveolar echinococcosis neglected disease.

Authors' contributions

LK, GS and DHM wrote the manuscript and designed the experiments. GS and DHM designed and implemented the SOM deep architecture and training scripts. CY developed the scripts for feature extraction and data pre-processing. LK, NM and LM analyzed data from high-throughput experiments. All authors read and approved the manuscript.

Acknowledgements

This work was supported by Agencia Nacional de Promoción Científica y Tecnológica (ANPCyT), Argentina, project PICT-CABBIO 2012 No 3044 and PICT 2014 No 2627, Universidad Nacional del Litoral (UNL) CAI+D 2011 548, and by Consejo Nacional de Investigaciones Científicas y Tecnológicas (CONICET) project PIP 114 2011 and PIP 117 2013. High-throughput analysis was performed in a local server at Instituto de Investigaciones en Microbiología y Parasitología Médicas (IMPaM) which is part of Sistema Nacional de Computación de Alto Desempeño (SNCAD) of Ministerio de Ciencia, Tecnología e Innovación Productiva (MINCyT). Thanks to Dr. Marcela Cucher for making raw data of *E. multilocularis* available.

Legends to figures

Figure 1: Flow diagram of the pipeline proposed for miRNA discovery from *Echinococcus multilocularis* genome-wide data. The folded *E. multilocularis* genome (491532 sequences) is used as input. Blue arrows indicate pre-processing and SOM analysis. Green arrows indicate pre-miRNA validation after RNA-seq data integration.

Figure 2: Architecture developed to find pre-miRNA candidates in *E. multilocularis* genome. Top: Hierarchy of SOM classifier for 10 levels (h=10). Dark blue neurons have highly likely pre-miRNA candidates, which are input to the next level SOM (black lines). Bottom: Number of pre-miRNA candidates in each level.

Figure 3: Schematic representation of the secondary structure from the conserved pre-miRNA 36b discovered by the SOM. The secondary structure predictions for pre-miRNA-36b is shown for four species of flatworms. Emul: *E. multilocularis*; Egra: *E. granulosus*; Ecan: *E. canadensis*; Hmic: *H. microstoma*. Mature miRNA sequences are underlined. Minimum free energy (MFE) is expressed as kcal/mol.

Figure 4: The secondary structure predictions of all new miRNAs from *E. multilocularis* discovered by SOM analysis. Mature miRNAs are indicated in red. Minimum free energy (MFE) is expressed as kcal/mol.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. Journal of Molecular Biology 215:403-410.

Ameres, S. and Zamore, P. (2013). Diversifying microRNA sequence and function. Nature Reviews Molecular Cell Biology, 14(8), 475–488.

Bai, Y., Zhang, Z., Jin, L., Kang, H., Zhu, Y., Zhang, L., X, X. L., Ma, F., Zhao, L., Shi, B., Li, J., McManus, D., Zhang, W., and Wang, S. (2014). Genome-wide sequencing of small RNAs reveals a tissue-specific loss of conserved microRNA families in Echinococcus granulosus. BMC Genomics, 1(15), 736.

Bartel, D. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. Cell, 116, 281–297.

Cucher, M., Prada, L., Mourglia-Ettlin, G., Dematteis, S., Camicia, F., Asurmendi, S., and Rosenzvit, M. (2011). Identification of Echinococcus granulosus microRNAs and their expression in different life cycle stages and parasite genotypes. International journal for parasitology, 41(3-4), 439–448.

Cucher, M., Macchiaroli, N., Kamenetzky, L., Maldonado, L., Brehm, K., and Rosenzvit, M. C. (2015). High-throughput characterization of echinococcus spp. metacestode mirnomes. International Journal for Parasitology, 45(4), 253–267.

de Souza Gomes M, Muniyappa MK, Carvalho SG, Guerra-S R, Spillane C. (2011) Genome-wide identification of novel microRNAs and their target genes in the human parasite Schistosoma mansoni. Genomics, 98 (2): 96-111.

Ding, J., Zhou, S., and Guan, J. (2010). MiRenSVM: towards better prediction of microRNA precursors using an ensemble SVM classifier with multi-loop features. BMC Bioinformatics, 11(11), S11.

Fromm, B., Worren, M., Hahn, C., Hovig, E., and Bachmann, L. (2013). Substantial Loss of Conserved and Gain of Novel MicroRNA Families in Flatworms. Molecular Biology and Evolution, 30(12), 2619–2628.

Gomes CPC, Cho J-H, Hood L, Franco OL, Pereira RW, Wang K. A. (2013). Review of Computational Tools in microRNA Discovery. Frontiers in Genetics., 4: 81.

Gkirtzou, K., Tsamardinos, I., Tsakalides, P., and Poirazi, P. (2010). MatureBayes: A probabilistic algorithm for identifying the mature miRNA within novel precursors. PLOS one, 5 (8), e11843.

Gudy, A., Szczeniak, M., Sikora, M., and Makalowska, I. (2013). HuntMi: an efficient and taxon-specific approach in pre-miRNA identification. BMC Bioinformatics, 14(1), 83+.

Hackenberg, M., Rodriguez-Ezpeleta, N., and Aransay, A. (2011). miRanalyzer: an update on the detection and analysis of microRNAs in high-throughput sequencing experiments. Nucleic Acids Research, 39(suppl 2), W132–W138.

Hertel, J. and Stadler, P. F. (2006). Hairpins in a Haystack: recognizing microRNA precursors in comparative genomics data. Bioinformatics, 22(14), e197–e202.

Hofacker, I., Fontana, W., Stadler, P., Bonhoeffer, S., Tacker, M., and Schuster, P. (1994). Fast Folding and Comparison of RNA Secondary Structures. Monatshefte für Chemie/*Chemical* Monthly, 125, 167–188.

Hofacker, I. L. (2003). The vienna rna secondary structure server. Nucleic Acids Research, 31, 3429–3431.

Huang, T. H., Fan, B., Rothschild, M., Hu, Z. L., Li, K., and Zhao, S. H. (2007). MiRFinder: an improved approach and software implementation for genome-wide fast microRNA precursor scans. BMC Bioinformatics, 8(1), 341.

de ON Lopes and A. Schliep and A. de Carvalho (2014). The discriminant power of RNA features for pre-miRNA recognition. BMC Bioinformatics, 15(1), 124+.

Jacobson, A. and Zuker, M. (1993). Structural Analysis by Energy Dot Plot of a Large mRNA. Journal of Molecular Biology, 233(2), 261–269.

Jiang, P., Wu, H., Wang, W., Ma, W., Sun, X., and Lu, Z. (2007). MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features. Nucleic Acids Research, 35(1), W339–W344.

Kohonen, T., Schroeder, M. R., and Huang, T. S. (2005). Self-Organizing Maps. Springer-Verlag New York, Inc.

Li, L., Xu, J., Yang, D., Tan, X., and Wang, H. (2010). Computational approaches for microRNA studies: a review. Mammalian Genome, 21(1), 1–12.

Liu, B., Li, J., and Cairns, M. (2014). Identifying mirnas, targets and functions. Briefings in Bioinformatics, 15(1), 1–19.

M. Friedlander and S. Mackowiak and N. Li and W. Chen and N. Rajewsky (2012). miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. Nucleic Acids Research, 40(1), 37–52.

Macchiaroli, N., Cucher, M., Zarowiecki, M., Maldonado, L., Kamenetzky, L., and Rosenzvit, M. C. (2015). microRNA profiling in the zoonotic parasite Echinococcus canadensis using a high-throughput approach. Parasites & Vectors, 8(1), 83.

Mendes, N., Freitas, A., Vasconcelos, A., and Sagot, M.-F. (2010). Combination of measures distinguishes pre-mirnas from other stem-loops in the genome of the newly sequenced anopheles darlingi. BMC Genomics, 11(1), 529.

Mendes, N. D., Heyne, S., Freitas, A. T., Sagot, M.-F., and Backofen, R. (2012). Navigating the unexplored seascape of pre-miRNA candidates in single-genome approaches. Bioinformatics, 28(23), 3034–3041.

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

O'Day, E. and Lal, A. (2010). MicroRNAs and their target gene networks in breast cancer. Breast cancer research : BCR, 12(2), 201.

Parrish, J., Xu, P., Kim, C., Jan, L., and Jan, Y. (2009). The microRNA bantam functions in epithelial cells to regulate scaling growth of dendrite arbors in drosophila sensory neurons. Neuron, 63(6), 788–802.

Piriyapongsa, J., Ramírez, L., and Jordan, K. (2007). Origin and evolution of human microRNAs from transposable elements. Genetics, 176(2).

Pritchard, C., Cheng, H., and Tewari, M. (2012). MicroRNA profiling: approaches and considerations. Nature reviews. Genetics, 13(5), 358–369.

Rahman ME, Islam R, Islam S, Mondal SI, Amin MR. (2012) MiRANN: a reliable approach for improved classification of precursor microRNA using Artificial Neural Network model. Genomics, 99(4):189-94.

Rosenzvit, M., Cucher, M., Kamenetzky, L., Macchiaroli, N., Prada, L., and Camicia, F. (2013). MicroRNAs in Endoparasites. Nova Science Publishers. Book, MicroRNA and Non-Coding RNA: Technology, Developments and Applications.Series: Genetics - Research and Issues. Pages: 7x10 – (NBC-R)Editors: James C. Johnson Editorial: Nova Science Publishers, Inc. Nueva York, USA. ISBN: 978-1-62618-443-5.

Saçar MD, Bağcı C, Allmer J (2014) Computational prediction of microRNAs from Toxoplasma gondii potentially regulating the hosts' gene expression. Genomics Proteomics Bioinformatics, 12(5):228-38.

Sætrom, P. and Snøve, O. (2007). Robust Machine Learning Algorithms Predict MicroRNA Genes and Targets. Methods Enzymol, 427C, 25–49.

Torgerson, P., Keller, K., Magnotta, M., and Ragland, N. (2010). The Global Burden of Alveolar Echinococcosis. PLoS Neglected Tropical Diseases, 4(6), e722.

Tsai, I., Zarowiecki, M., Holroyd, N., and et al. (2013). The genomes of four tapeworm species reveal adaptations to parasitism. Nature, 496, 57–63.

L. Wei, M. Liao, Y. Gao, R. Ji, Z. He, and Q. Zou (2014) Improved and Promising Identification of Human MicroRNAs by Incorporating a High-quality Negative Set, IEEE/ACM Trans. Comput. Biol. Bioinformatics, 11 (1): 192–201

Winter, A., Wei, W., Hunt, M., Berriman, M., Gilleard, J., Devaney, E., and Britton, C. (2012). Diversity in parasitic nematode genomes: the microRNAs of Brugia pahangi and Haemonchus contortus are largely novel. BMC Genomics, 13(1), 4.

Xu, Y., Zhou, X., and Zhang, W. (2008). MicroRNA prediction with a novel ranking algorithm based on random walks. Bioinformatics, 24(1), i50–i58.

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

Yones CA, Stegmayer G, Kamenetzky L, Milone DH. (2015) miRNAfe: A comprehensive tool for feature extraction in microRNA prediction. Biosystems, 138:1-5

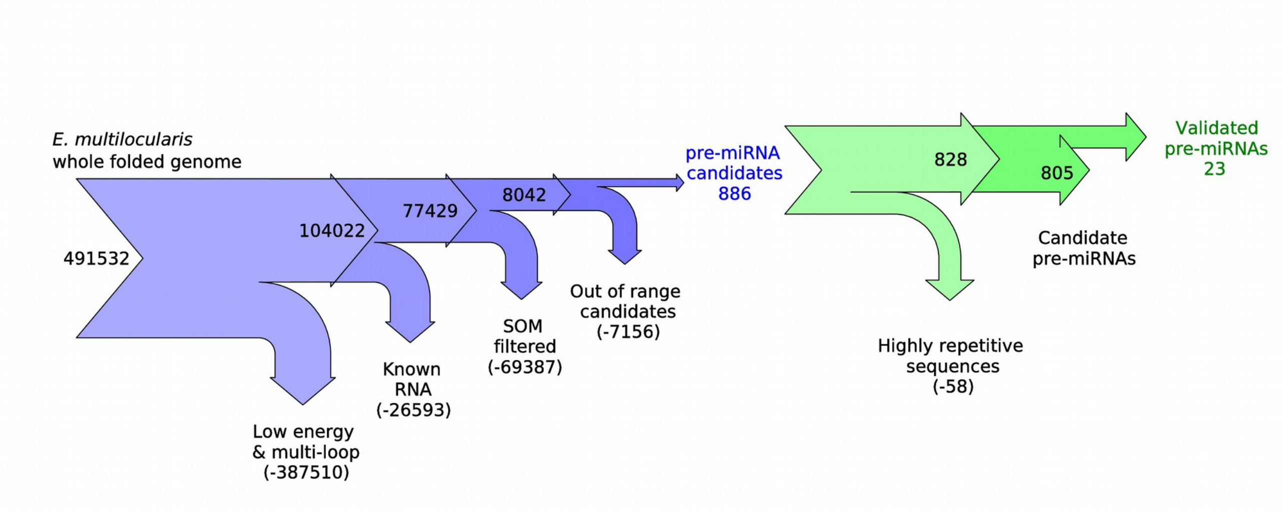
Yoo, A. and Greenwald, I. (2005). LIN-12/Notch activation leads to microRNA-mediated down-regulation of Vav in C. elegans. Science, 310(5752), 1330–1333.

M. Yousef, S. Jung, L. Showe and M. Showe (2008) Learning from Positive Examples when the Negative Class is Undetermined- microRNA gene identification, Algorithms for Molecular Biology, 3: 2.

Zhou, Y., Liu, Y., Yan, H., Li, Y., Zhang, H., Xu, J., Puthiyakunnon, S., and Chen, X. (2014). miR-281, an abundant midgut-specific miRNA of the vector mosquito Aedes albopictus enhances dengue virus replication. Parasit Vectors, 1(7), 488.

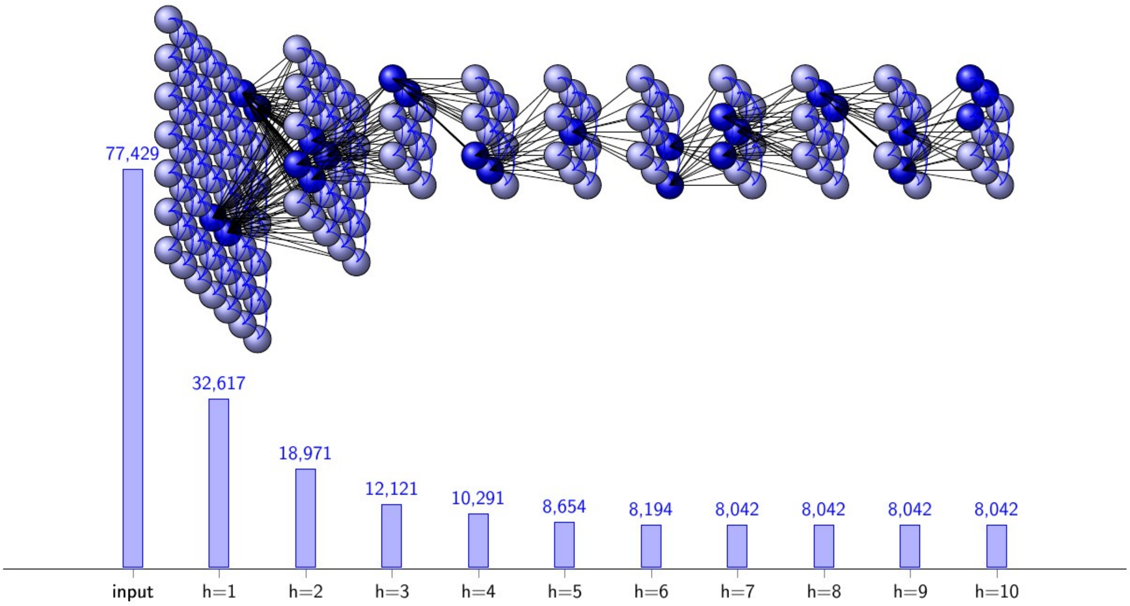
Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. Nucleic acids research, 31(13), 3406–3415.

Figure 1



E. multilocularis
whole folded genome
491532
104022
77429
8042
pre-miRNA
candidates
886
Low energy
& multi-loop
(-387510)
Known
RNA
(-26593)
SOM
filtered
(-69387)
Out of range
candidates
(-7156)
828
805
Validated
pre-miRNAs
23
Candidate
pre-miRNAs
Highly repetitive
sequences
(-58)

502 **Figure 2**



77,429
32,617
18,971
12,121
10,291
8,654
8,194
8,042
8,042
8,042
8,042
input
h=1
h=2
h=3
h=4
h=5
h=6
h=7
h=8
h=9
h=10