Chapter 2

Screen of MicroRNA Targets in Zebrafish Using Heterogeneous Data Sources: A Case Study for Dre-miR-10 and Dre-miR-196

Based on

Yanju Zhang, Joost M. Woltering, Fons J. Verbeek. (2007). Screen of MicroRNA Targets in Zebrafish Using Heterogeneous Data Sources: A Case Study for Dre-miR-10 and Dre-miR-196 Proceedings WASET, Bangkok. Also published at International Journal of Mathematical, Physical and Engineering Sciences, Vol. 2 (1), 10 - 17.

Summary

It has been established that microRNAs (miRNAs) play an important role in gene expression by post-transcriptional regulation of messengerRNAs (mRNAs). However, the precise relationships between microRNAs and their target genes in sense of numbers, types and biological relevance remain largely unclear. Dissecting the miRNA-target relationships will render more insights for miRNA targets identification and validation therefore promote the understanding of miRNA function. In miRBase, miRanda is the key algorithm used for target prediction for zebrafish. This algorithm is high-throughput but brings lots of false positives (noise). Since validation of a large scale of targets through laboratory experiments is very time consuming, several computational methods for miRNA targets validation should be developed. In this chapter, we present an integrative method to investigate several aspects of the relationships between miRNAs and their targets with the final purpose of extracting high confident targets from miRanda predicted targets pool. This is achieved by using the techniques ranging from statistical tests to clustering and association rules. Our research focuses on zebrafish. It was found that validated targets do not necessarily associate with the highest sequence matching. Besides, for some miRNA families, the frequency of their predicted targets is significantly higher in the genomic region nearby their own physical location. Finally, in a case study of dre-miR-10 and dre-miR-196, it was found that the predicted target genes *hoxd13a*, *hoxd11a*, *hoxd10a* and *hoxc4a* of dre-miR-10 while *hoxa9a*, *hoxc8a* and *hoxa13a* of dre-miR-196 have similar characteristics as validated target genes and therefore represent high confidence target candidates.

1 Introduction

The microRNA (miRNA) field started with the discovery of lin-4 in 1993 [15] which was initially considered as an isolated case but later miRNAs have been found to be widely present in multicellular organisms, ranging from plants to human. MicroRNAs (miRNAs) are $\sim$22 nucleotide single-stranded noncoding RNA molecules that repress messenger-RNA (mRNA) translation or mediate mRNA degradation through sequence-specific base pairing [18, 7]. Several miRNAs have been found to play an important role in life and development. To name a few: miRNAs lin-4 and let-7 regulate developmental timing in *C. elegans* [15, 20]; bantam and miR-14 are involved in the gene regulation of apoptosis in *Drosophila* [2]; miR-181 modulates hematopoietic lineage differentiation in mice [5]; miR-32 regulates primate foamy virus type 1 (PFV-1) proliferation in human [14].

MiRNAs function by binding to target sites in mRNAs and thereby preventing their translation or promoting their decay. In order to better understand the biological function of miRNAs, it is of fundamental importance to identify miRNA targets. Identifying miRNA targets in animals is not as straightforward as in plants. Computational approaches have been successful in plants, where known target sites tend to be almost perfectly complementary to miRNAs [21, 28]. Whereas in animals, miRNA-target binding is loosely complementary [19]. The inexact sequence match property has complicated computational approaches for target site identification significantly.

Several computational high-throughput methods to predict miRNA targets have been described [7, 25, 16, 3]. The miRanda algorithm is one of the frequently used methods. For each miRNA, target genes are selected on the basis of three properties: sequence complementarity using a position-weighted local alignment algorithm, free energy of RNA-RNA duplexes, and conservation of target sites in related genomes [7, 9].

This computational method introduces one crucial problem, i.e., too much noise. Most likely, not all of the predicted targets for a miRNA represent true biological targets and only a few of these have been confirmed either positive or negative. For example, regarding lin-4 in *C. elegans*, 554 targets are predicted and to date only 2 are confirmed through laboratory experiments. Therefore, nowadays the challenge is to find an effective way to filter out false positive predicted targets. Accurate target prediction and validation are still major obstacles in miRNA research.

Recently, as opposed to other computational methods like miRanda, a few bottom-up approaches for high-throughput miRNA targets validation have been reported. Zhou *et al.* suggest that targets identified by multiple prediction algorithms would appear to be the better candidates for verification [32]. Stark *et al* describe an algorithm to screen targets according to sequence and free energy features shared by validated targets [26].

Unlike the above described methods, we explore a bottom-up approach which focuses on selecting targets based on genomic location and physical association on the genome.

An integrative method is presented to analyze the relationships between miRNAs and targets in order to extract high confident miRNA targets. The method consists of three layers: data retrieval, data analysis and data visualization. A panel of algorithms such as clustering and association rules are applied on different resources such as genomic location information, physical association on the genome, Gene Ontology terms as well as predicted sequence scores and p-values generated by miRanda algorithm.

Results from the analysis indicate that validated targets do not necessarily associate with highest sequence matching. For some miRNA families, the relative frequency of predicted targets is significantly higher in the genomic region surrounding their own location. The method is illustrated in a case study using two zebrafish miRNA families: dre-miR-10 and dre-miR-196. Their currently known targets can be treated as control. Finally on the basis of the method, we suggest *hoxd13a*, *hoxd11a*, *hoxd10a* and *hoxc4a* as high confident targets for dre-miR-10 and *hoxa13a*, *hoxa9a*, *hoxc8a* for dre-miR-196. Our approach is a prelude to large scale machine learning analysis for all miRNAs in zebrafish.

This chapter is structured as follows. In section 2, the material and components of the approach are introduced. Section 3 describes the results which indicate the feasibility of the method. Finally, in section 4, we conclude the results, discuss the advantages and disadvantages of our approach and prospect for our future work.

2 Material and Methods

The workflow of the method is displayed in Fig. 1. In this section the components of our approach and how these are integrated in the analysis are described.

2.1 Material

Zebrafish miRNAs and predicted targets were derived from miRBase. The most recent (release 9.2) miRBase (http://microrna.sanger.ac.uk/) contains 233 miRNAs belonging to 177 families and 23331 predicted targets. The genomic location and symbols for each gene are retrieved from Ensembl database danio_rerio_core_45_6f.

MiRBase is the repository for published miRNA sequence data, annotation and predicted gene targets [9, 8]. It consists of three parts:

- The miRBase Registry acts as an independent arbiter of miRNA gene nomenclature, assigning names prior to publication of novel miRNA sequences.
- The miRBase Sequences is the primary online repository for miRNA sequence data and annotation.
- The miRBase Targets is a comprehensive new database of predicted miRNA target genes.

Gene Ontology (GO) provides structured, controlled vocabularies and classifications that cover several domains of molecular and cellular biology; these are freely available for the community to annotate genes, gene products and sequences across all species [10]. All the genes and gene products are described in a species-independent manner using three descriptors namely biological process, cellular component and molecular function [1].

Ensembl is an information system to store, analyze, use and display genomic information. In addition to sequence information, Ensembl also incorporates other biological data such as cross-species, synteny, genes, transcripts, proteins, supporting evidences, dot-plots, protein domains and gene or protein families [12, 24].

2.2 Data retrieval

In general, there are three ways to create database access: using a public mirror database, downloading individual database tables or files, and creating one's own private mirror [23]. We assemble all relevant information in a local database by three different ways based on the consideration of speed, consuming time and space.

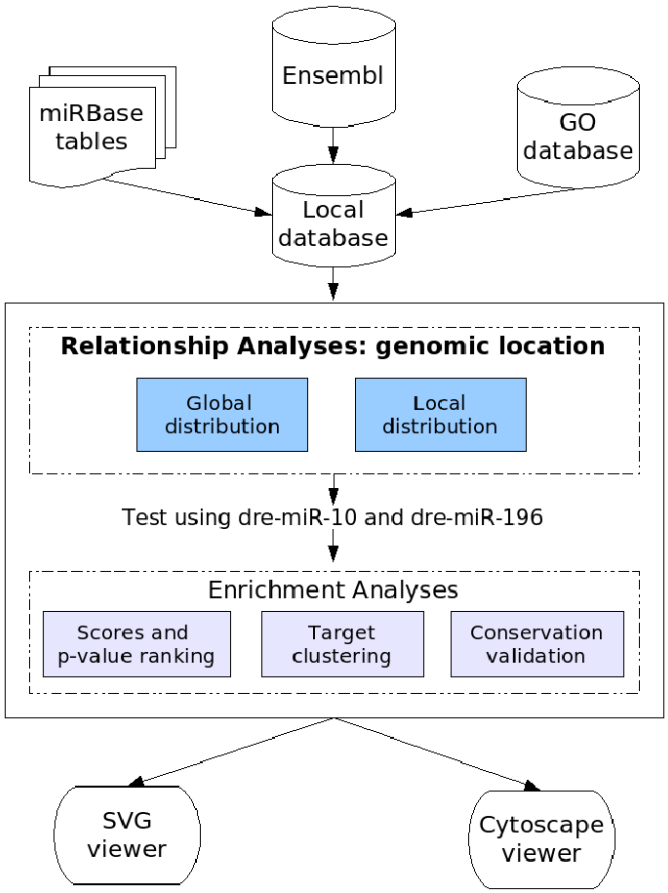


Figure 1: The workflow of the miRNA targets validation method. It consists of three stages: data retrieval, analysis and visualization.

Firstly, to access miRNAs and targets data, the sequence and target tables for zebrafish in miRBase are downloaded. Secondly, as far as the genomic information is concerned, it is retrieved from Ensembl public mirror database. In order to avoid consuming too much time and space, the Ensembl data is accessed through the Ensembl Perl Application Programming Interface (API). This API is a framework of applications for accessing or storing data in the Ensembl databases. The great advantage of using the Ensembl API is that it separates developers from the underlying structure and changes at a lower level. Without deep knowledge of the schema of the database, information can be easily fetched from database. Thirdly, the annotation is retrieved from Gene Ontology database which is directly available through our local AmiGO database.

2.3 Analysis

Due to vertebrate genome duplication, often multiple copies and isoforms of one particular miRNA exist [30]. These multiple copies and isoforms are sequence similar and function alike. In our research, the predicted targets are analyzed for each miRNA family instead of miRNA individuals. With the final aim of screening high confident targets, the genomic location relationships between miRNAs and targets are firstly investigated through a global and a local distribution analysis.

Next, the high confident targets for dre-miR-10 and dre-miR-196 are predicted on the basis of the found relationships. Moreover, the confident targets are validated by using sequence matching score and p-value ranking, targets clustering as well as conservation validation.

Global distribution analysis

We start with exploring the genomic distribution of all the targets for each miRNA family. With the results we intend to answer whether all the targets are evenly distributed over all the 25 chromosomes or more predicted targets are located in the same chromosomes as their miRNAs.

To achieve this, firstly all the targets are mapped from mRNA level to gene level and the genomic location is extracted from Ensembl. Subsequently, a t-test is used to compare the difference between the average targets number over all chromosomes and that over their miRNA located chromosomes. The alternate H_1 hypothesis is defined as follow: true difference in means between the number of target genes distributed on all zebrafish chromosomes and that on their miRNA located chromosome is not equal to 0.

Local distribution analysis

For the well characterized *hoxb8* and miR-196, it is known that the miRNA and target gene are physically located within each others close vicinity [31]. Therefore we investigate whether this represents a more common theme for miRNA-target relationships, and if there is a correlation between the genomic locations of predicted target genes and miRNAs. It is also possible that the targets near miRNAs have high probability of being true

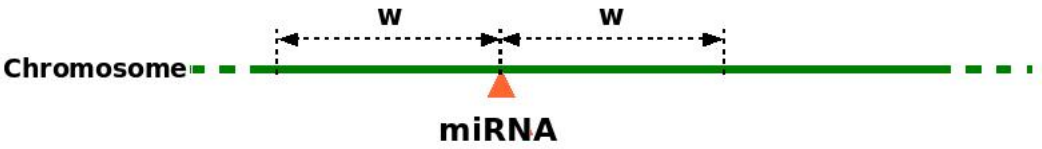


Figure 2: Window size definition

targets.

For this purpose, the targets are mapped from transcripts to genes and the genomic distance between miRNAs and their targets are calculated. The distance is calculated by genomic position subtraction when targets located on the same chromosomes as the miRNAs. For other targets, the distance is defined as infinity. Window size is defined as physical distance each centered on the position of a specified miRNA as displayed in Fig. 2. Thus, we statistically analyze the numbers of targets in 50kb to 1000kb window size. Moreover, to investigate the areas which contain more targets, Expected target number (E_{target}) and Relative Frequency (RF) are defined as follows.

$$E_{target}[w] = N_{gene}[w] \times \frac{N_{alltargets}}{N_{allgenes}} \tag{1}$$

$$RF[w] = \frac{N_{target}[w]}{N_{gene}[w]} \tag{2}$$

Where $[w]$ represents within window size w; function N_{object} gets the number of object; $E_{target}[w]$ represents the number of target genes which are expected to be present in window w. This is derived from the number of genes in window w multiplied by the proportion of target genes and genomic genes. According to this definition, the number of the expected targets and that of the miRBase predicted targets in different windows for each family are compared in order to detect in which region the predicted targets distributed regularly.

Relative frequency in a specific window $RF[w]$ is calculated using the number of predicted targets divided by the number of genes in the window w. It enables us to compare the target frequency between different areas significantly. According to the relative fre-

quency, the areas which are prone to have more targets can be deduced.

By better understanding the genomic location relationships between miRNAs and targets, the targets are ranked according to their genomic location and further validated using the following steps.

Matching scores and p-values ranking

At present, the accuracy of the miRanda algorithm predictions is unknown, whereas miRanda offers several likely outputs as predictors for target genes i.e. the sequence match score and the p-value. The match score represents the complementarity between miRNAs and their targets. The p-value represents an estimated probability of the same miRNA family hitting multiple transcripts for different species in an orthologous group [17].

In order to assess whether high sequence match score or low p-value are associated with real targets, the predicted targets are sorted by either matching scores or p-values for each miRNA family. Henceforth, we examine whether the known and the selected targets are captured in the top 50 ranked lists. In general, the number of the predicted targets for different miRNA families vary from 420 to 2016, therefore selecting 50 can cover 2.5% to 12% of the predicted targets (*cf.* Section 4).

Clustering analysis

Since a specific family of miRNAs is likely to function in specific biological processes, it is assumed that its targets also belong to functional gene groups.

Gene Ontology (GO) terms are standardized annotation for genes and gene products. Here we apply association rules to cluster targets according to GO terms. Association rules discovery technique (ARD) is a machine learning method that has been used to discover associations among subsets of items in large transaction databases. This method detects sets of elements that frequently co-occur in a database and establish relationships between them [4]. Genes which share a number of GO terms are associated to one set. Based on association rules, the similarity between target genes is defined as follow:

$$Similarity(g1, g2) = \frac{S(g1 \bigcup g2)}{S(g1 \bigcap g2)} \tag{3}$$

Where $S(g)$ is the function which calculates the number of GO terms for the gene. $g1 \bigcap g2$ represent the intersection of GO terms between gene1 and gene2. While $g1 \bigcup g2$ represent the union of GO terms for gene1 and gene2.

Conservation validation

Conservation plays an important role in targets selection. It is known that *hox* genes are expressed collinear in time and space along the anteroposterior body axis and highly conserved across species [27]. It is also verified and showed in miRBase that miR-10 and miR-196 are conserved in other vertebrates like mouse and human.

After knowing the genomic location of miRNAs and targets, the conservation of the physical location relationships between miRNAs and their targets are studied. The selection of target genes for dre-miR-10 and dre-miR-196 are checked whether they are located closely together in other species as well. For this purpose, we utilize the found miRNA-target relationships, repeat the genomic location analyses and detect the closely located targets near miR-10 and miR-196 in human and mouse.

2.4 Visualization

Scalable Vector Graphics (SVG) and the Cytoscape viewer are applied in order to visualize the results.

Scalable Vector Graphics is a language for describing two-dimensional graphics and graphical applications in XML [22]. SVG produces vector based graphics and consequently, the resulting pictures can be zoomed without degradation. In using SVG, the intention is that all the predicted targets or a set of interested targets and miRNA families can be viewed globally on all chromosomes, at the same time detail location between genes and their miRNAs can be even zoomed in to basepair scale.

Cytoscape software platform is frequently used in bioinformatics for visualizing molecular interaction networks and integrating these interactions with gene expression profiles and other state data [6]. In our application it is suitable to visualize the results of the clustering. Nodes represent targets or target genes, while edges represent the similar functions as retrieved from the GO term identity. Furthermore, the visualization of other attributes

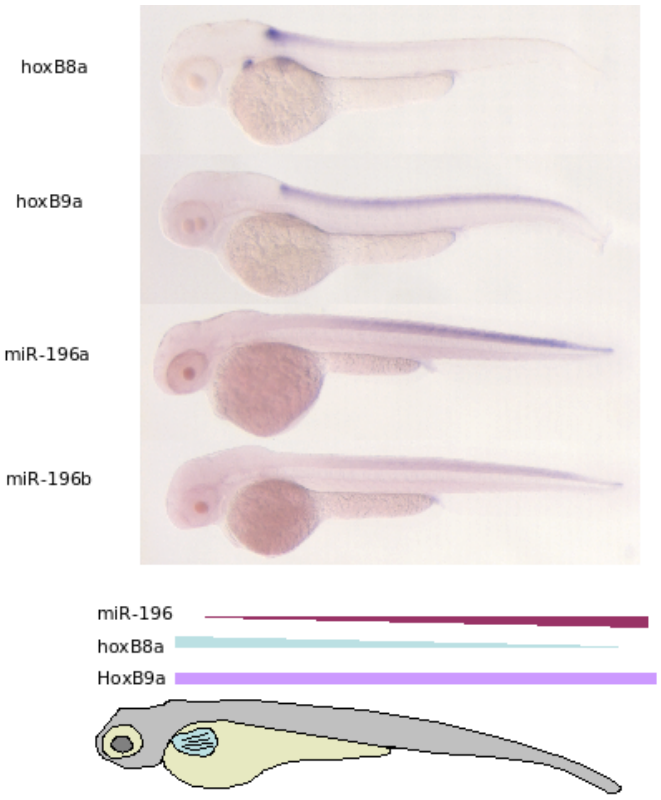
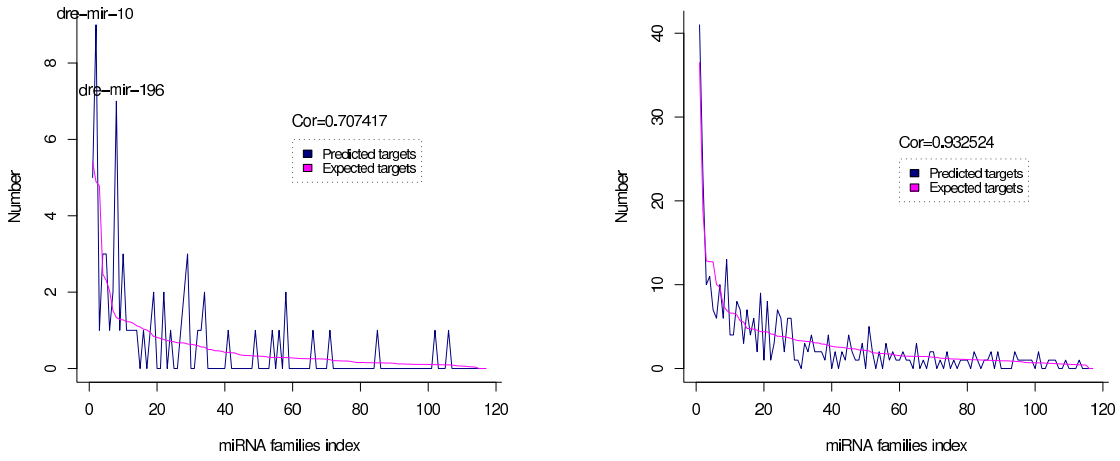


Figure 3: Expression patterns of *hoxb8a*, *hoxb9a* and dre-miR-196a and 196b. It showed the mutually excluding expression patterns for *hoxb8a* and *mir-196* and constant expression of *hoxb9a*.

such as genomic location and p-value can be supplemented and showed in a sub panel.



(a) 100kb (b) 1000kb

Figure 4: Expected vs. predicted target numbers in 100kb (a) and 1000kb (b) windows. Cor represents the correlation coefficient between expected and predicted targets curves.

3 Results

It has been validated that *hoxb8a* is the target of dre-miR-196. Fig. 3 shows an *in situ* hybridization for *hoxb8a*, *hoxb9a* and dre-miR-196a and miR-196b on 48 hpf zebrafish embryos. *Hoxb8a* is a target gene for miR-196. Obviously, this figure showed the mutually excluding expression patterns for the two genes in the spinal cord where *hoxb8a* is expressed in the anterior and miR-196 in the posterior part. However, the *hoxb9a* gene which is physically located in between miR-196a and *hoxb8a* is expressed with the same intensity throughout the spinal cord.

In the global distribution analysis, the alternate hypothesis was defined in Section 2.3. According to t-test, the average targets number over all chromosomes and over their miRNA located chromosomes equal to 32.22154 and 31.96404 respectively. The p-value which equals 0.8926 indicates that there is a 90% probability that the H_1 hypothesis occurred by chance. As a consequence, it is concluded that when studied on a chromosomal scale there is no significant difference between the target density in all chromosomes and in the chromosomes wherein the miRNA is located.

Next the targets distribution on a smaller scale were studied by comparing the numbers of targets in different windows surrounding the miRNA genomic positions. Fig. 4 shows the number of expected targets and predicted targets for 117 miRNA families showed as index in the window of 100kb Fig. 4(a) and 1000kb Fig. 4(b). The correlation coefficient for the group of expected and predicted targets in 100kb is 0.707417, which is less than 0.932524 in 1000kb. This indicated that target genes in 100kb are distributed less proportionally with the genomic genes in comparison with the one in 1000kb. From this, it is deduced that the 100kb window may be an interesting zone to be further examined.

In order to compare the targets distribution difference in 100kb and 1000kb, the relative frequency was calculated as equation (2). 35 out of 117 total number of families are found having targets in the window of 100kb. Furthermore, 85.7% of them have relative frequency in the window of 100kb greater than the one in 1000kb. Fig. 5 shows that 12 out of 13 selected families, which have highest absolute targets number in the window of 100kb, have relative frequency in the window of 100kb higher than in 1000kb. Therefore it is concluded that many miRNA families are likely to have a higher density of predicted targets located in nearby their genomic regions.

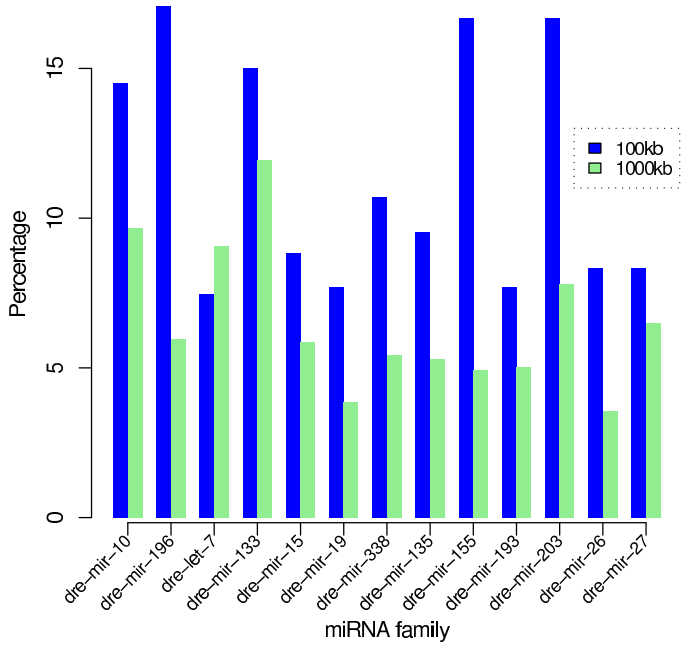
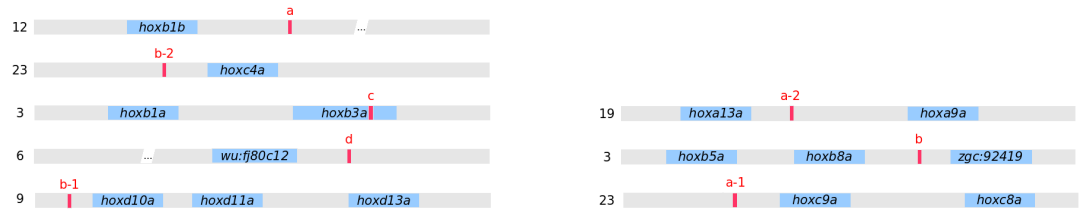


Figure 5: Relative frequency in the window of 100kb and that in 1000kb. It illustrated that 12 out of 13 families have relative frequency in the window of 100kb higher than the one in 1000kb.

According to the above findings and the fact that dre-miR-196 and its known target gene *hoxb8a* are physically close, the targets which are located within 100kb window size of their miRNAs are screened and are assumed to have high probability of being true targets. This is a so called distance criterion. In our study, we applied this distance criterion to dre-miR-10 and dre-miR-196. Fig. 6(a) and 6(b) illustrate the relative genomic location of the high ranked targets depicted in blue (*hoxb1b*, *hoxc4a*, *hoxb1a*, *hoxb3a*, *wu:fj80c12*, *hoxd10a*, *hoxd11a hoxd13a*, *hoxa13a*, *hoxa9a*, *hoxb5a*, *hoxb8a*, *zgc:92419*, *hoxc9a* and *hoxc8a*) and the miRNA genomic copies depicted in red (dre-miR-10: a, b-1, b-2, c, d and dre-miR-196: a, b-1, b-2) respectively. They are located in different chromosomes



(a) Dre-mir-10 (b) Dre-mir-196

Figure 6: The relative genomic location for the high ranked targets of dre-miR-10 and dre-miR-196. High ranked targets (blue) and miRNAs (red) are located in different chromosomes marked by the numbers in front of each line. The intervals in chromosome 6 and 12 represent the duplicated entries due to the zebrafish genomic assembly errors.

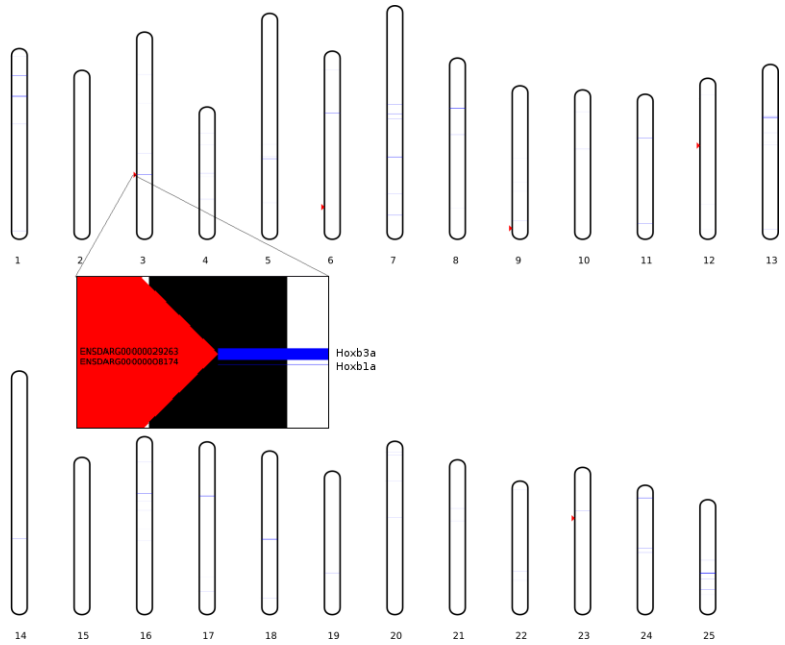


Figure 7: The overview of genomic location of top 50 predicted targets of dre-miR-10 ranked by p-value. The isoforms of dre-miR-10 (red triangles) and targets (blue lines) are displayed over 25 chromosomes (columns). The closeup view illustrated two *hox* genes *hoxb1a* and *hoxb3a* genomic located near dre-miR-10.

marked by the numbers on the left side. The length of each box is not related to the length of genes. The intervals in chromosome 6 and 12 represent the duplicated entries for dre-

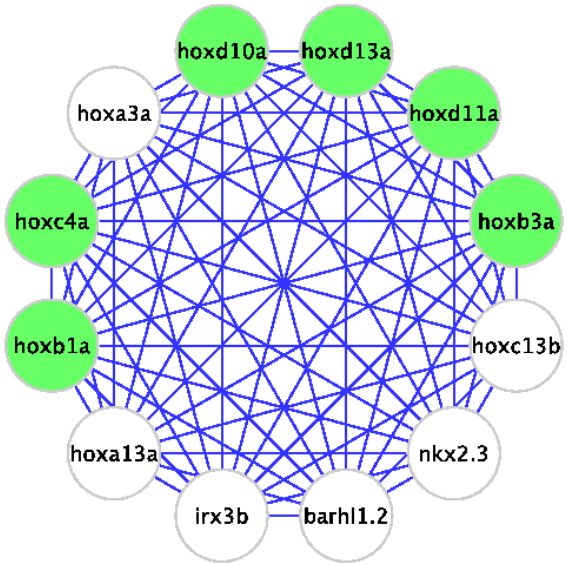


Figure 8: A group of targets for dre-miR-10 which have the same GO term descriptions (expressed by lines) as known targets *hoxb1a* and *hoxb3a*. Within 100kb distance target genes are showed in green.

miR-10a and 10d caused by the zebrafish genomic assembly errors. Since the erroneous *in silico* duplications are mapped close to each other, they do not interfere with our data analysis.

After that the analyses were enriched by using sequence matching scores or p-values ranking, targets clustering and conservation validation.

The top 50 targets for miRNAs selected by p-value were visualized using SVG viewer. Fig. 7 shows the case for dre-miR-10. Zebrafish possesses 5 genomic miR-10 copies attributed to 4 isoforms named a, b, c and d [30] (*cf.* Fig. 6). The genomic positions of the different dre-miR-10 copies are depicted by red triangles. Targets selected by p-value for dre-miR-10 are shown by the blue lines distributed over 25 chromosomes. From the detailed view, it is clear that there is also a physical association between dre-miR-10 and its confirmed targets *hoxb1a* and *hoxb3a*.

Validated targets are known for dre-miR-196 namely *hoxb8a* [31, 11] and for dre-miR-10, *hoxb1a* and *hoxb3a* [29]. These are the controls in the analysis. *Hoxb8a* is found in both top 50 lowest p-value and top 50 highest score scale for dre-miR-196. The known targets *hoxb1a* and *hoxb3a* for miRNA dre-miR-10 are in top 50 lowest p-value but not in top 50 highest score list. These results showed that real targets do not necessarily associate with the highest sequence matching. Whereas selecting good targets by p-value works well in these two miRNA families, since the known targets all have very low p-values.

Regarding to GO term clustering, in current stage the GO term similarity is set to 100% defined as clustering criterion. This means that genes which have the same GO descriptions are grouped together. Fig. 8 shows a particular set for dre-miR-10 visualized with Cytoscape viewer. This set consists of not only the known targets *hoxb1a* and *hoxb3a* but also *hoxd13a*, *hoxd11a*, *hoxd10a* and *hoxc4a* which are physically closely located with dre-miR-10 in the window of 100kb showed in green.

Except for the known target *hoxb8a*, targets *hoxa9a* and *hoxc8a* are found also conserved in mouse and human. The results of the enrichment process are listed in Table 1. The selected targets are validated by testing whether they are in top 50 lowest p-value list (abbreviated Top 50 p in Table 1) or functioning like known targets (abbreviated GO as known in Table 1) or conserved in mouse and human. The known targets are marked in boldface.

Table 1: Enrichment information for high confident targets selected by distance criterion.

Candidates	Top 50 p	GO as known	Conservation
hoxb3a	√	√	-
hoxb1a	√	√	-
hoxb1b	-	-	-
hoxc4a	-	√	-
wu:fj80c12	-	-	-
hoxd10a	-	√	-
hoxd11a	-	√	-
hoxd13a	-	√	-
hoxb8a	√	√	√
hoxa9a	√	-	√
hoxc9a	-	-	-
hoxc8a	√	-	√
hoxa13a	-	√	-
hoxb5a	-	-	-
zgc:92419	-	-	-

Finally, based on the distance criterion combined with either p-value ranking or function similarity or conservation, *hoxd13a*, *hoxd11a*, *hoxd10a* and *hoxc4a* are predicted as high confidence targets for dre-miR-10 and in similar fashion *hoxa9a*, *hoxc8a*, *hoxa13a* for dre-miR-196.

4 Conclusions and discussion

To date, still little is known on the interactions of miRNA with the transcriptome. In order to promote the understanding of these interactions and learn how to perform pattern recognition using the available resources, we presented an integrated approach to validate miRNA targets through the analyses of physical location, p-value, the function of the targets and conservation. We found that validated targets do not necessarily associate with the highest sequence matching. Such is consistent with the general idea that targets can imperfectly bind to miRNAs in animal systems [19]. An interesting phenomenon we found is that for most of miRNA families, which have predicted targets located near by, the frequency of their predicted targets is significantly higher in the genomic region nearby their own locations. This result is, to a certain extent, consistent with the findings which

report lower expression of genes near miRNA in C. elegans germline [13]. In addition, the method was validated in the case study of dre-miR-10 and dre-miR-196. For these two miRNA families, the known targets *hoxb8a*, *hoxb1a*, *hoxb3a*, which were described as control targets, are also captured in the high ranked targets scale screened by using distance criterion. This may suggest that genomic location of miRNAs and their targets also have an effect on miRNAs function. Furthermore, the enrichment analysis enhanced the confidence to some of the candidates. Target genes *hoxd10a*, *hoxd11a*, *hoxd13a* and *hoxc4a* are not only located nearby dre-miR-10 but also have the same GO descriptions as the known targets. For dre-miR-196 the closed located target genes *hoxa9a* and *hoxc8a* are conserved in mouse and human and have low p-values as well. Integrating all the results, finally *hoxd13a*, *hoxd11a*, *hoxd10a* and *hoxc4a* were predicted as high confident targets for dre-miR-10 and *hoxa9a*, *hoxc8a*, *hoxa13a* for dre-miR-196.

Nevertheless, there are still some limitations in the method. Firstly, the input data sources are from different databases, the degree of the accuracy of these databases affects the results. For example, the genomic assembly errors in Ensembl will probably affect the analysis of other miRNAs. Secondly, since the actual mechanism of miRNAs remains unclear, our assumptions may only be suitable for a selection of miRNAs. Thirdly, in the current version, we use some preset values as cutoff. This can be improved in the future by computing the cutoff values from the datasets and evaluating them through a number of computational approaches.

In general, different from other miRNA targets screen approaches, we integrated heterogeneous data sources and algorithms to screen target candidates mainly based on genomic location feature which were elucidated as playing a role in miRNA-target interaction. By using Ensembl perl API, the progress of the analysis has been greatly improved and the Ensembl data are easily updated and retrieved.

An important step in the analysis was to visualize the relations and the physical mapping so that our collaborators could grasp the underlying ideas. This was accomplished with SVG, i.e. physical location of miRNAs and targets, and Cytoscape, i.e. GO relations between targets.

In the future, this approach will be extended to other model systems and we are going to integrate miRNAs microarray analysis which can monitor the temporal and spatial expression profile of miRNAs and their targets during zebrafish embryo development.

By knowing the relationship between the expression of miRNAs and genes, the research of the biological mechanism of miRNAs can be further facilitated. Besides these, more data mining techniques are going to be applied to dissect miRNA target features. This approach is a prelude to large scale machine learning analysis for all miRNAs in zebrafish and possibly other model systems.

Acknowledgements

This research has been partially supported by the BioRange program of the Netherlands Bioinformatics Centre (NBIC, BSIK grant).

References

[1] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet*, 25(1):25–29, May 2000.

[2] J. Brennecke, D. R. Hipfner, A. Stark, R. B. Russell, and S. M. Cohen. bantam encodes a developmentally regulated microrna that controls cell proliferation and regulates the proapoptotic gene hid in drosophila. *Cell*, 113(1):25–36, April 2003.

[3] J. R. Brown and P. Sanseau. A computational view of micrornas and their targets. *Drug Discov Today*, 10(8):595–601, April 2005.

[4] P. Carmona-Saez, M. Chagoyen, A. Rodriguez, O. Trelles, J. M. Carazo, and A. Pascual-Montano. Integrated analysis of gene expression by association rules discovery. *BMC Bioinformatics*, 7, 2006.

[5] C.Z. Chen, L. Li, H.F. Lodish, and D.P. Bartel. Micrornas modulate hematopoietic lineage differentiation. *Science*, 303(5654):83–86, Jan 2004.

[6] Cytoscape. http://www.cytoscape.org/.

[7] A. J. Enright, B. John, U. Gaul, T. Tuschl, C. Sander, and D. S. Marks. Microrna targets in drosophila. *Genome Biol*, 5(1), 2003.

[8] S. Griffiths-Jones. The microrna registry. *Nucleic Acids Res*, 32, January 2004.

[9] S. Griffiths-Jones, R. J. Grocock, S. van Dongen, A. Bateman, and A. J. Enright. mirbase: microrna sequences, targets and gene nomenclature. *Nucleic Acids Res*, 34, January 2006.

[10] M. A. Harris, J. Clark, A. Ireland, J. Lomax, M. Ashburner, R. Foulger, K. Eilbeck, S. Lewis, B. Marshall, C. Mungall, and et. al. The gene ontology (go) database and informatics resource. *Nucleic Acids Res*, 32(Database issue), January 2004.

[11] Eran Hornstein, Jennifer H. Mansfield, Soraya Yekta, Jimmy K. Hu, Brian D. Harfe, Michael T. Mcmanus, Scott Baskerville, David P. Bartel, and Clifford J. Tabin. The microrna mir-196 acts upstream of hoxb8 and shh in limb development. *Nature*, 438(7068):671–674, 2005.

[12] T. J. Hubbard, B. L. Aken, K. Beal, B. Ballester, M. Caccamo, Y. Chen, L. Clarke, G. Coates, F. Cunningham, T. Cutts, and et. al. Ensembl 2007. *Nucleic Acids Res*, 35(Database issue), January 2007.

[13] Hidenori Inaoka, Yutaka Fukuoka, and Isaac S. Kohane. Lower expression of genes near microrna in c. elegans germline. *BMC Bioinformatics*, 7(1), March 2006.

[14] CH Lecellier, P Dunoyer, K Arar, J Lehmann-Che, S Eyquem, C Himber, A Sab, and O. Voinnet. A cellular microrna mediates antiviral defense in human cells. *Science*, 308(5721):795–825, April 2005.

[15] R. C. Lee, R. L. Feinbaum, and V. Ambros. The c. elegans heterochronic gene lin-4 encodes small rnas with antisense complementarity to lin-14. *Cell*, 75(5):843–854, December 1993.

[16] B. P. Lewis, I. H. Shih, M. W. Jones-Rhoades, D. P. Bartel, and C. B. Burge. Prediction of mammalian microrna targets. *Cell*, 115(7):787–798, December 2003.

[17] miRBase. http://microrna.sanger.ac.uk/targets/v4/faq.html.

[18] R. H. Plasterk. Micrornas in animal development. *Cell*, 124(5):877–881, March 2006.

[19] N. D. Rajewsky, N.and Soccib. Computational identification of microrna targets. *Developmental Biology*, 267(2):529–535, March 2004.

[20] Brenda J. Reinhart, Frank J. Slack, Michael Basson, Amy E. Pasquinelli, Jill C. Bettinger, Ann E. Rougvie, Robert H. Horvitz, and Gary Ruvkun. The 21-nucleotide let-7 rna regulates developmental timing in caenorhabditis elegans. *Nature*, 403(6772):901–906, February 2000.

[21] M. W. Rhoades, B. J. Reinhart, L. P. Lim, C. B. Burge, B. Bartel, and D. P. Bartel. Prediction of plant microrna targets. *Cell*, 110(4):513–520, August 2002.

[22] ScalableVectorGraphics. http://www.w3.org/graphics/svg/.

[23] Peter Schattner. Automated querying of genome databases. *PLoS Computational Biology*, 3(1), January 2007.

[24] J. Stalker, B. Gibbins, P. Meidl, J. Smith, W. Spooner, H. Hotz, and A.V. Cox. The ensembl web site: Mechanics of a genome browser. *Genome Res*, 14(5):951–955, May 2004.

[25] A. Stark, J. Brennecke, R. B. Russell, and S. M. Cohen. Identification of drosophila microrna targets. *PLoS Biol*, 1(3), December 2003.

[26] A. Stark, J. Brennecke, R. B. Russell, and S. M. Cohen. Identification of drosophila microrna targets. *PLoS Biol*, 1(3), December 2003.

[27] A. Tanzer, C. T. Amemiya, C. B. Kim, and P. F. Stadler. Evolution of microrna located within hox gene clusters. *Experimental Zoology*, 304B:75–85, 2005.

[28] Xiaowei Wang and Xiaohui Wang. Systematic identification of microrna functions by combining target prediction and expression profiling. *Nucleic Acids Research*, 34(5):1646–1652, 2006.

[29] J. M. Woltering and A. J. Durston. Mir-10 targets hoxb1a and hoxb3a and is required for correct migration of the xth cranial nerve. *In preparation*, 2007.

[30] Joost M. Woltering and Antony J. Durston. The zebrafish hoxdb cluster has been reduced to a single microrna. *Nature Genetics*, 38(6):601–602, 2006.

[31] S. Yekta, I. H. Shih, and D. P. Bartel. Microrna-directed cleavage of hoxb8 mrna. *Science*, 304(5670):594–596, April 2004.

[32] J Zhou, V Melfi, J Verducci, and S Lin. Composite microrna target predictions and comparisons of several prediction algorithms. *JSM 2006 Online Program*, 2006.