Heterogeneous data analysis for annotation of microRNAs and novel genome assembly

Zhang, Y.

Heterogeneous Data Analysis for Annotation of microRNAs and Novel Genome Assembly

Yanju Zhang

This work was part of the BioRange programme of the Netherlands Bioinformatics Centre (NBIC), which is supported by a BSIK grant through the Netherlands Genomics Initiative (NGI).

Printed by Ridderprint, RIDDERKERK

ISBN 978-90-5335-491-9

Heterogeneous Data Analysis for Annotation of microRNAs and Novel Genome Assembly

PROEFSCHRIFT

ter verkrijging van

de graad van Doctor aan de Universiteit Leiden,

op gezag van Rector Magnificus Prof. mr. P.F. van der Heijden,

volgens besluit van het College voor Promoties

te verdedigen op donderdag 24 november 2011

klokke 16.15 uur

door

Yanju Zhang

Geboren te Guilin, China, in 1980

Promotiecommissie

Promotor:

Prof. dr. J.N. Kok

Co-promotor:

Dr. Ir. F.J. Verbeek

Overige Leden

Prof. dr H.P. Spaink	Leiden University
Prof. dr. A.P.J.M Siebes	Utrecht University
Prof. dr. H. Blokkeel	Katholieke Universiteit Leuven, Belgium
Dr. E. Vreugdenhil	Leiden University

Contents

Chapter 1

Introduction

1 General introduction

This thesis is the collection of four published papers demonstrating annotation of genes and microRNAs with the aid of bioinformatics, in particular using heterogeneous data integration. In this thesis, the research objects are genes and microRNAs. Genes are regions of DNA that can be transcribed to messenger RNA and later on translated to proteins which are the chief actors within the cell. MicroRNAs (miRNAs) are recently discovered very short messenger RNAs which are transcribed from DNA sequences. Instead of being further translated, these short RNAs bind to messenger RNAs, and thus inhibit their target expression. The main goal of this thesis is to efficiently and accurately annotate miRNAs and coding region of a novel genome. To achieve these goals, we developed several complex workflows which integrate the current data sources and tools together. Chapter 2, 3 and 4 are about miRNA annotation, while in Chapter 5 we demonstrate genome annotation of the common carp.

The purpose of the introduction is to provide the general background of the subjects that were studied, motivations and applied methodologies and to make the connections between chapters explicit. First, the key concepts of this thesis, which are integration and annotation, are explained in Section 2 and 3. Subsequently, the biological background of the research objects is introduced in Section 4 followed by the general analysis of miRNA and carp genome annotation. The final section is an overview of the thesis.

2 Methodology: integration

Life science is a research field that elucidates the complicated and delicate biological mechanisms of living organisms. With the development of high-throughput technologies, a huge amount of system-wide biological data, e.g. genomic, transcriptomics and proteomics are produced. The capability of generating multi-omic datasets brings new challenges to Bioinformatics.

Bioinformatics is a rapidly developing area that applies computational approaches to solve biological problems. Basically, it is an interdisplinary science that utilizes computers to store and process biological data and develops and applies statistics, algorithms and pipelines to analyze biological data. The final goal is to accelerate and enhance our

understanding of biological phenomena, mechanisms and processes. Currently, a lot of computational tools and algorithms have been developed and have shown the capacity of facilitating our understanding towards biological mechanisms.

With the huge amount of multi-omic data sets and hundreds of bioinformatics tools available, there is a need for integration of heterogeneous resources. Heterogeneous data refers to the information from multiple sources and in many varying formats and structures. Currently, the huge amounts of heterogeneous data in life science are generated at relatively high speed by different organizations all over the world. It is more and more frequently required to correlate and combine the heterogeneous information as the volume and the need to share data explodes. The essence of integration is not to produce even more data by combining different data sources or types but to increase the sensitivity and/or specificity of the algorithm and system.

Data integration can be achieved by two methods: management and analysis. From the management point of view, heterogeneous data integration is the process of the standardization of data definitions and structures by using a common conceptual schema across a collection of data sources [12, 19]. This leads to the development of common databases, warehouses, software, platforms and systems that retrieve data from different sources and provide a unified view. One example is the National Center for Biotechnology Information (NCBI) database which is a U.S. government-funded national resource for molecular biology. This database provides information such as genomics, proteomics, bioinformatics tools and literature for researchers. The topic of management will not be addressed specially in this thesis.

In terms of analysis, integration correlates and combines data from several experiments and databases in an effort to extract better and more significant information than the means of a single source. This technique is widely applied in data-driven bioinformatics which requests to build a model or analysis after the data has been generated. Integration brings new insights from multi-dimensional data and therefore improves our understanding of the research. Using integration for heterogeneous data analysis is the general theme though this thesis.

In general, data can be integrated from two ends, low level and high level. Low-level integration refers to the analysis dealing with multi-factorial raw data directly. One example is prognosing a disease by combining DNA variation, gene expression and phenotypic

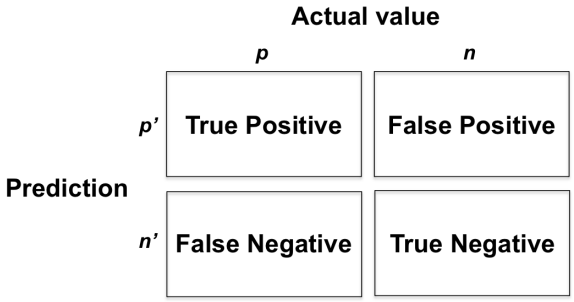


Figure 1: Definitions of true positive (TP), false positive (FP), false negative (FN) and true negative (TN) in binary classification. Positive (p) and negative (n) are the two classes, and p' and n' are the prediction outcomes. A true positive occurs when a positive instance is predicted as positive; however if the actual value is negative and prediction is positive, then it is called a false positive. False negative and true negative can be defined in a similar way.

data. High-level integration, on the other hand, means to integrate multiple same-type results from different studies [18]. For example, in the pathway analysis, the significant pathways derived from different approaches might not be identical. In this case, it will be interesting to integrate the results from different methods to arrive at some consensus that is more reliable than any of the individual results.

Whatever levels the data are integrated on, they can be integrated in either a sequential or a parallel fashion. In the sequential approach, each type of data can be used as a filter. In the analysis of differentially expressed genes in microarrays, possible candidates are first selected through statistical analysis. After that, Gene Ontology or pathway information can serve as an enrichment dataset to further screen differentially expressed genes. In the parallel approach, different raw data are treated as features or measurements and integrated by machine learning algorithms to build models with the final goal of finding patterns, trends and anomalies.

Integration will lead to the improvement of sensitivity and/or specificity which are the two measurements of system performance. Sensitivity, also known as the true positive rate, is defined as the ratio of actual positives which are correctly identified; specificity measures the probability that the negatives are correctly identified. In the case of two classes classification, as shown in Fig. 1, sensitivity and specificity are defined as equation 1 and 2 respectively.

$$Sensitivity = \frac{TP}{TP + FN} \tag{1}$$

$$Specificity = \frac{TN}{FP + TN} \quad (2)$$

where TP, FN, TN and FP denote the true positive, false negative, true negative and false positive respectively. In general, for all algorithms, it is desirable to achieve both high sensitivity and specificity. However, there is a trade-off between the measures; high sensitivity will sacrifice specificity by increasing its false positive rate and vice versa.

Many high-throughput methods sacrifice specificity for scale. Microarray is the technique which can monitor the expression patterns of thousands of genes simultaneously. Microarray analysis can predict gene function by assessing coexpression relationships in a high throughput fashion. Although gene coexpression data are an excellent tool for hypothesis generation, microarray data alone often lack the degree of specificity needed for accurate gene function prediction.

In some cases, sensitivity is sacrificed for accuracy. In epidemiologic studies, accurately diagnosing the disease of a patient outweighs finding all the potential patients. Therefore high specificity tools are the key for accurate disease diagnoses which have great impact on the consequent treatment; For the purpose of validation, specificity of an algorithm outweighs its sensitivity. When high-throughput biological validation is not available, only a few highly ranked candidates will be selected for testing in priority.

The cutoff for sensitivity and specificity are arbitrary decisions. Users can decide the cutoff to achieve a higher sensitivity or specificity according to their own requirements.

Integration normally is not a straightforward process. Multiple steps will be involved according to the heterogeneity of the data. Usually an integration strategy is represented by a workflow which is the depiction of a sequence of operations. Each operation is a model and the workflow is the collection of these models processed in a desirable order. Using workflow, the process is repeatable, therefore the same type of heterogeneous data can be integrated in the same manner. The development of workflows is faciliated by the tools such as Taverna [24]. It is a workflow management system allows bioinformaticans to build workflows using the tools and databases avaliable on the web.

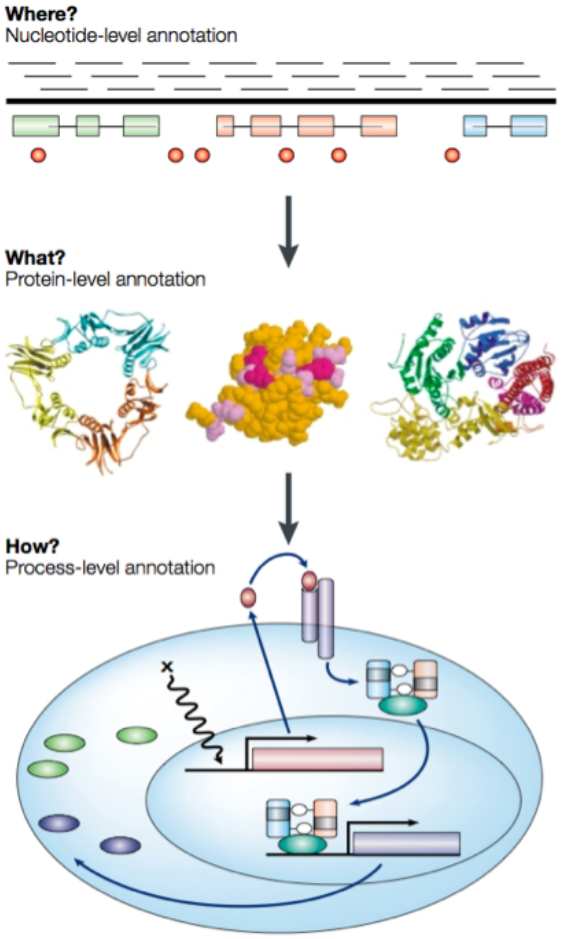


Figure 2: Three layers of genome annotation. Nucleotide-level annotation aims for identifying the physical map of the functional units. Protein-level annotation aims for identifying 3D protein configurations and protein-protein interactions. Process-level annotation aims for identifying the biological processes which the functional units are involved in. *-Lincoln Stein. Genome annotation from sequence to biology. 2001.*

3 Goal: annotation

'Genome annotation is the process of taking the raw DNA sequence and adding the layers of analysis and interpretation necessary to extract its biological significance and place it into the context of our understanding.'

-Lincoln Stein. Genome annotation from sequence to biology. 2001.

Annotation is an important and necessary analysis which bridges the gap between biological sequence and the biology of the organism. During the past decade, only a few genomes have been completely annotated, such as *Saccharomyces cervesiae* (yeast), *Caenorhabdi-*

tis elegans (worm), *Drosophila melanogaster* (fruitfly) and *Arabidopsis thaliana* (mustard weed). Many other genomes are on the way, including mouse, rat, zebrafish, pufferfish and human [31]. Genome annotation is a complex process which can be achieved from three levels: nucleotide, protein and process as displayed in Fig. 2.

The task of nucleotide-level annotation is to identify the physical map, e.g. the start and end position of the functional units. In this phase, the most important analysis is gene finding, i.e. to determine structures for the protein coding genes. In prokaryotes, gene finding is comparatively easy since most of the genome is comprised by the coding region. However in eukaryotes, the case is more complicated. Firstly, the genome size is relatively big. Secondly, less than 25% of the genome is a coding region [31]. And thirdly, splicing and alternative splicing events take place during transcription. All these factors complicate the gene finding. One branch of algorithms predicts gene structures using a data mining strategy which trains a model with currently available genes and predicts structures for the novel sequences. Another branch is the homolog gene prediction that derives a complete gene model according to the sequence similarities of other species. The sequence alignment tool BLASTX [20] can be used for this purpose. Due to the complexity of gene structures, the current trend in gene prediction is the combination of the above-mentioned *ab initio* and comparative methods.

On the protein level, the main goal is to detect protein structures and protein interactions. Proteins are the essential functional units within a cell. They comprise sequences of amino acids folded in 3D structures carrying specified information encoded in the gene. Most cellular processes are carried out by protein-protein interactions, such as forming a complex or signal transduction. In practice, protein structures can be predicted by searching for similarities using BLASTP against several protein sequences databases such as SWISS-PROT [3], or by searching against functional domain databases such as PFAM [7]. Protein-protein interactions can be simulated using protein docking tools such as STRING [32].

The last and most challenging part of the annotation is called functional annotation, the process in which the genes and proteins are linked to different biological processes, e.g. cell cycles and apoptosis. At process-level annotation, the Gene Ontology (GO) [11] and pathway database are the main resources. GO is a standardized vocabulary for describing the functions of eukaryotic genes categorized in molecular functions, biological processes

and cellular components. Pathway databases such as KEGG [14] and BioCarta [23] are widely utilized at this stage.

4 Biological Background

In our studies, functional annotation of miRNA is mostly performed on zebrafish and human data. The zebrafish, which is small in size, easily cultured and has transparent embryos, is a model organism used in molecular genetics and developmental biology. As for human, many studies have been performed and databases on human biology are the most complete.

De novo genome assembly and annotation are applied to the common carp. The common carp is becoming a serious candidate model organism for very high throughput screens of pharmaceutical compound libraries and we have been participating in a recently initialized common carp genome project. In this section, a brief introduction of miRNA and key components in a genome project will be given.

4.1 MicroRNAs

For a long time, researches have been working on unraveling the function of DNA coding sequences which are responsible for the expression of proteins, the functional units in the cell. The scientists also wonder why the non-coding sequences, sometimes called 'junk DNA' (since no known biological function was previously found in this region), are conserved through evolutionary selection. New light was shed on this problem. In 1993 the first miRNA lin-4 was identified in the 'junk DNA' of *C. Elegans* [15]. It was found that lin-4 encodes a 22-nucleotide non-coding RNA that negatively regulates the expression of the lin-14 gene in a temporal control of post-embryonic development [1]. In 2000, another non-coding RNA let-7 was discovered [26]. Since then, an abundant amount of these gene regulators have been identified in a variety of plants, animals and viruses. The discovery of miRNAs revealed a new mechanism of gene regulation and inspired a series of molecular and biochemical studies in this area.

Mature miRNAs are $\sim$22 nucleotide single-stranded noncoding RNA molecules. They are transcribed from miRNA genes. The process of biogenesis and function of miRNAs

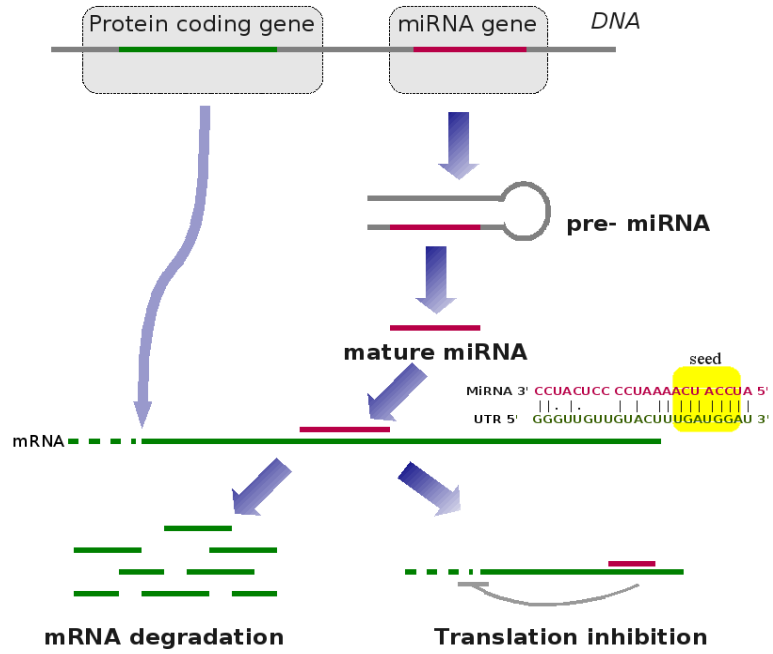


Figure 3: Simplified illustration of miRNA biogenesis and function. miRNA genes are first transcribed to pre-miRNA, and then processed to mature miRNAs. Upon binding to these miRNAs through sequence complementarity, the messenger RNAs (mRNAs), which are called the targets of miRNAs, will be either degraded or the translation of the targets will be inhibited.

are illustrated in Fig. 3. For reasons of simplification the auxiliary protein complexes are not included in the picture. First, a miRNA gene is transcribed to primary miRNA transcripts, which are between a few hundred or a few thousand base pairs long. Subsequently, this primary miRNA is processed into hairpin precursors, called pre-miRNA, which have a length of approximately 70 nucleotides, by the protein complex consisting of the nuclease Drosha and the double-stranded RNA binding protein Pasha. The pre-miRNA is then transported to cytoplasm and cut into small RNA duplexes of approximately 22 nucleotides by the endonuclease Dicer. Finally, either the sense strand or antisense strand functioning as a template gives rise to mature miRNA. Upon binding to the active RISC complex, mature miRNAs interact with the target mRNA molecules through base pair complementarity, therefore inhibit translation or sometimes induce mRNA degradation [6].

The main functional characterization method of miRNAs is based on the loss-of-function mutation of miRNA genes. Using this technique, fly miR-14 was identified as an inhibitor of apoptotic cell death [34]; worm lsy-6 was found to promote specific cell fates [13];

the miR-34 family was discovered in the p53 pathway in which p53 genes are tumor suppressors [5]. Many studies suggest that a miRNA can have the capacity of regulating hundreds of genes and in total miRNAs could regulate about 30% of the gene expression in humans [16].

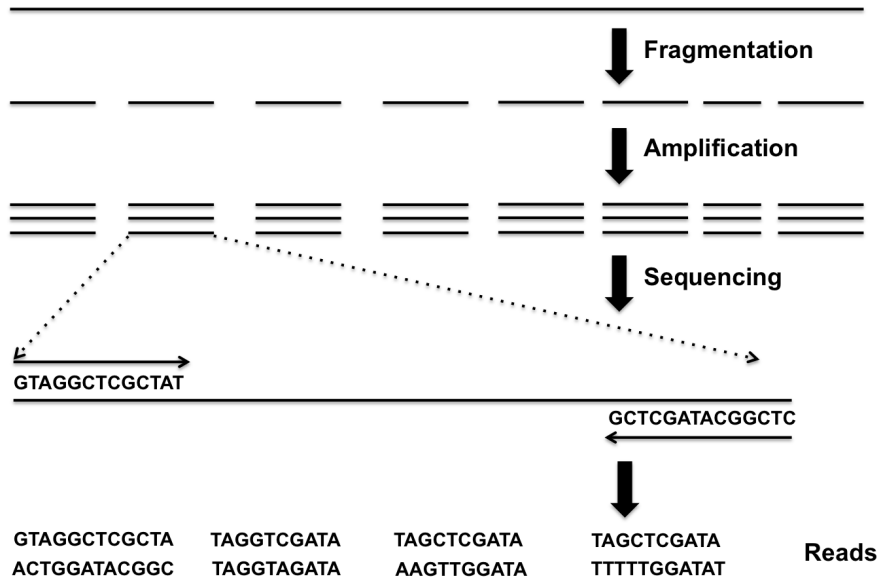
The miRNAs are also found to be involved in the pathogenesis of infectious diseases and cancer. It was discovered that miR-107 is associated with Alzheimer's disease [33]; miR-133b is related to Parkinson's disease; miR-1 plays a role in the development of cardiovascular diseases [9]. These findings have resulted in miRNAs becoming drug target candidates in many pharmaceutical research projects.

4.2 Genome project

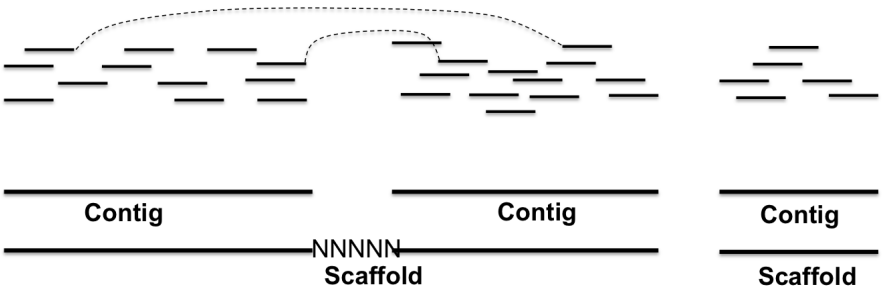
The human genome project, initialized in 1993, released a draft and a complete genome assembly in 2000 and 2003 respectively. These groundbreaking results showed that scientists are capable of decoding the full set of DNA that make a human. Since then, many genome projects of different species, such as zebrafish and mouse, have been initiated. Aiming to determine the complete genome sequence of an organism, a genome project, in general, consist of three stages: sequencing, assembly and annotation. The procedure of sequencing and assembly are briefly explained in Fig. 4.

Genome sequencing is the process of determining the order of nucleotides over the whole genome. In the 1970's, most DNA sequencing was performed using the chain termination method, developed by Fred Sanger [27]. In the last couple of years, remarkable technological innovations have emerged that allow the cost-effective sequencing of complex samples at an unprecedented scale and speed [25]. These techniques are referred to as next-generation sequencing or high-throughput sequencing since they are based on principles different from the classical Sanger-based method (first generation). They can produce thousands or millions of sequences at once with a fraction of the cost of traditional sequencing. The new sequencing platforms include Roche 454, Genome Analyzer (Illumina/Solexa) and ABI-SOLiD (Applied Biosystems).

The development of next-generation sequencing technologies poses numerous computational challenges for bioinformatics. High speed and scale of data generation challenge the efficient and effective way of data storage and processing. *De novo* assembly is one



(a) Sequencing



(b) Assembly

Figure 4: Principle of sequencing and assembly. At the sequencing stage, as shown in (a), first DNA molecules are extracted and then sheared into short fragments. Later on adaptors are attached to one or both ends. With or without amplification, each fragment is then sequenced by the sequencer to obtain short sequences from one end or both ends resulting in single-end or paired-end reads. Genome assembly is the process that constructs the original continuous DNA sequences from millions of short DNA reads. The concepts are illustrated in (b). Contigs represent the contiguous pieces of DNA, while a scaffold refers to the joint contigs according to the pairing information

of the steps that is computationally extremely expensive, i.e. time, memory and CPU consuming. It is a process of piecing millions or billions of short reads together to form a set of continuous sequences (contigs) representing the DNA in the sample. Previously, *de novo* assembly was achieved using overlapping computation strategies, while currently the *de Bruijn* [22] graph representation is prevalent in assemblers. Some of the most frequently used assemblers are Velvet [35], ABySS [30], Phusion [21], CLC Bio genomic

workbench [2], Curtain [28] and SOAPdenovo [17].

The analysis after a genome has been sequenced and assembled is genome annotation, which refers to finding the protein coding genes and other functional units such as miRNAs, and then further attaching biological functions, biochemical functions and expression patterns to these elements. Annotation is the goal of this thesis and has been introduced in Section 3.

5 Challenges in annotation of miRNAs and carp genome

5.1 Annotation of miRNAs

In the last few decades, 851 mature miRNAs in human and 233 in zebrafish have been identified (miRBase http://microrna.sanger.ac.uk/). But due to lack of high throughput experiments, functional studies have only touched upon a small fraction of miRNAs [8]. Thus, the main challenge in miRNA studies is to unravel the function of miRNAs. One crucial aspect is to identify the targets with which they directly interact.

For most of the miRNAs, functional characterization can benefit from bioinformatics by predicting miRNA target genes. In plants, miRNA target predictions have proven to be straightforward because miRNAs bind to their targets by nearly perfect sequence complementarity. In contrast in animals, the degree of sequence complementarity in miRNA-target pairing can be flexible leaving the mechanism of how miRNAs interact with the target unclear. Currently, bioinformatics prediction algorithms are built relying on rules that are derived from a few known miRNA-target interactions. These rules are 1) high sequence complementarity between 3'UTR of the target and miRNAs; 2) perfect match between 3'UTR of the target and seed region of miRNAs, in which the seed region, also called the nucleus, is the sequence from position 2 to position 8; 3) favorable structural and thermodynamic formation between RNA-RNA duplexes; 4) evolutionary conservation of miRNA target sites.

Many public databases have been built to facilitate miRNA studies. miRBase [10] is the integrated repository for the miRNAs as well as their predicted targets. TarBase [29] records all the experimentally validated targets collected from the published literature. These databases provide valuable information and have triggered the development

of some data mining algorithms which predict candidates based on miRNA-target interaction models built from known targets.

5.2 Annotation of carp genome

Common carp (*Cyprinus carpio*) is one of the most important fresh water cultured fish species. It is widely used in fish biology research [4]. A single female is capable of producing up to a few hundred thousand eggs that can be efficiently fertilized in vitro, which enables hundreds of thousands of pharmaceutical drug candidates to be tested with a relatively small genetic diversity. Thus, common carp is a relevant model system for high throughput screens of pharmaceutical compound libraries.

Currently, there are 32046 carp EST and 2136 carp nucleotide sequences recorded in Genbank, but there is no carp genome assembly available. Using the next-generation sequencing technology, we have generated a huge amount of sequence reads from the carp genome and transcriptome with which we aim to identify all the carp genes. Since zebrafish is evolutionarily close to the common carp (both are cyprinids) and the zebrafish genome is relatively well covered and annotated in the Ensembl database, we used the zebrafish genome to facilitate the annotation of the carp genes.

We currently focus on discovering the carp genes involved of the innate immune response as a pilot study. The innate immune system is the first line of defense against infectious diseases and cancer by identifying and killing pathogens and detrimental cells. Understanding of the gene structures and their expressions will benefit the testing of hundreds of thousands of pharmaceutical drug candidates.

6 Structure of the thesis

This thesis is composed of two parts categorized by the research objects. In Chapter 2, 3 and 4, we focus on the functional annotation of miRNAs via target predictions. While in Chapter 5, we will describe the aspects of *de novo* genome assembly and annotation for a new candidate model system, the common carp.

In Chapter 2, we focus on the discovery of miRNA targets in zebrash. An integrative method is described to investigate several aspects of the relationships between miRNAs

and their targets with the nal purpose of extracting high condent targets from the target pool predicted by miRanda. This is achieved by using techniques ranging from statistical tests to clustering and association rules. In this chapter, we found that validated targets do not necessarily associate with the highest sequence matching scores. Besides, for some miRNA families, the frequency of their predicted targets is signicantly higher in the genomic region close to their own physical location. Finally, in a case study of dre-miR-10 and dre-miR-196, it was found that seven candidate target genes, all of which belong to hox gene family, have similar characteristics as validated target genes and therefore represent high confidence target candidates.

In Chapter 3, we present an approach that analyzes miRNA-miRNA relationships and utilizes them for target predictions in human. We have developed a pipeline which integrates machine learning techniques to reveal the feature patterns between known miRNAs. Different data setups are evaluated and compared to achieve the best performance. Furthermore, the derived rules are applied to miRNAs of which the targets are not yet known so as to see if new targets could be predicted. Our method contributes to the improvement of target identification by predicting targets with high specificity and without conservation limitations.In the analysis of functionally similar miRNAs, we found that genomic distance and seed similarity between miRNAs are dominant features in the description of a group of miRNAs binding the same target. Application of one specific rule resulted in the prediction of targets for several unannotated miRNAs. Some of these targets were also detected by the existing methods.

In Chapter 4, we evaluate the performance of different target prediction algorithms and use integration methods to improve prediction accuracy. Both high-level integration approaches, e.g. algorithm combinations and ranking aggregation, and low-level integration approaches, e.g. a Bayesian Network classification, are performed. All of the methods are tested on miRNA-target interactions that are experimentally validated and several compiled negative control data sets. The results reveal that each individual prediction algorithm has its own advantages, as was shown using different test datasets. Moreover, we inspected on the characteristics of miRNA-target site interactions and discovered a novel feature: i.e. miRNAs have binding preference at the end of the 3' UTR sequence of their target. Finally, we concluded that among different integration strategies, the application of the Bayesian Network classifier on the features calculated from multiple prediction

methods significantly improved target prediction accuracy. Further research is directed towards the categorization of miRNA-target interaction into subtypes, i.e. to discriminate the targets for degradation and for translation inhibition.

In Chapter 5, we focus on the assembly and functional annotation of the carp genome. The common carp is a candidate model system that can be used for high throughput screens of pharmaceutical compound libraries. In this chapter, we develop a genome assembly and an annotation pipeline with the final aim of identifying immune response genes, especially Toll/Interleukin-1 receptor (TIR) domain-containing genes, using next generation sequencing data. The genome assembly pipeline consists of data cleaning, pre-assembly and assembly using CLCBio, ABySS and SOAPdenovo. A basic annotation pipeline of these low coverage genomes is obtained by using simple gene prediction based on protein-based gene model prediction as well as comparative annotation to other genomes which is a prediction of ortholog with respect to zebrafish. The preliminary assembly was achieved with an N50 contig length of 2260 bp and from our data it is estimated that the carp genome is about 1.23 Gbp. Compared to zebrafish immuno genes, we estimated that there are 39 TIR domain-containing genes and transcripts in the common carp.

In Chapter 6, the techniques used in the previous chapters will be summarized. Moreover, the lessons we learned from the studies will be discussed.

References

[1] S. Bagga, J. Bracht, S. Hunter, K. Massirer, J. Holtz, R. Eachus, and A. E. Pasquinelli. Regulation by let-7 and lin-4 miRNAs results in target mRNA degradation. *Cell*, 122(4):553–563, August 2005.

[2] CLC Bio. http://www.clcbio.com/.

[3] Brigitte Boeckmann, Amos Bairoch, Rolf Apweiler, Marie claude Blatter, Anne Estreicher, Elisabeth Gasteiger, Maria J. Martin, Karine Michoud, Isabelle Phan, Rine Pilbout, and Michel Schneider. The swiss-prot protein knowledgebase and its supplement trembl in 2003. *Nucleic Acids Res*, 31:365–370, 2003.

[4] A. B. J. Bongers, M. Sukkel, G. Gort, J. Komen, and C. J. J. Richter. Development and use of genetically uniform strains of common carp in experimental animal research. *Lab Anim*, 32(4):349–363, 1998.

[5] Tsung-Cheng C. Chang, Erik A. Wentzel, Oliver A. Kent, Kalyani Ramachandran, Michael Mullendore, Kwang Hyuck H. Lee, Georg Feldmann, Munekazu Yamakuchi, Marcella Ferlito, Charles J. Lowenstein, Dan E. Arking, Michael A. Beer,

Anirban Maitra, and Joshua T. Mendell. Transactivation of miR-34a by p53 broadly influences gene expression and promotes apoptosis. *Molecular cell*, 26(5):745–752, June 2007.

[6] C. Z. Chen. MicroRNAs as oncogenes and tumor suppressors. *N Engl J Med*, 353(17):1768–1771, October 2005.

[7] Robert D. Finn, John Tate, Jaina Mistry, Penny C. Coggill, Stephen John Sammut, Hans rudolf Hotz, Goran Ceric, Kristoffer Forslund, Sean R. Eddy, Erik L. L. Sonnhammer, and Alex Bateman. The pfam protein families database. *Nucleic Acids Res*, 36:281–288, 2008.

[8] Dimos Gaidatzis, Erik van Nimwegen, Jean Hausser, and Mihaela Zavolan. Inference of miRNA targets using evolutionary conservation and pathway analysis. *BMC bioinformatics*, 8:69+, March 2007.

[9] Michela Garofalo, Gerolama Condorelli, and Carlo Maria Croce. Micrornas in diseases and drug response. *Current Opinion in Pharmacology*, 8(5):661–667, 2008.

[10] Sam Griffiths-Jones, Russell J. Grocock, Stijn van Dongen, Alex Bateman, and Anton J. Enright. mirbase: microrna sequences, targets and gene nomenclature. *Nucleic Acids Research*, 34(Database-Issue):140–144, 2006.

[11] C. J. Harris. The gene ontology (GO) database and informatics resource – gene ontology consortium 32 (supplement 1): 258 – nucleic acids research. *Nucleic Acids Res.*, 1(32):D258–D261, January 2004.

[12] Dennis Heimbigner and Dennis Mcleod. A federated architecture for information management. *ACM Trans. Inf. Syst.*, 3(3):253–278, July 1985.

[13] Robert J J. Johnston Jr and Oliver Hobert. A novel c. elegans zinc finger transcription factor, lsy-2, required for the cell type-specific expression of the lsy-6 microRNA. *Development*, November 2005.

[14] M. Kanehisa and S. Goto. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30, January 2000.

[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] Benjamin P. Lewis, Christopher B. Burge, and David P. Bartel. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, 120(1):15–20, January 2005.

[17] Ruiqiang Li, Hongmei Zhu, Jue Ruan, Wubin Qian, Xiaodong Fang, Zhongbin Shi, Yingrui Li, Shengting Li, Gao Shan, Karsten Kristiansen, Songgang Li, Huanming Yang, Jian Wang, and Jun Wang. De novo assembly of human genomes with massively parallel short read sequencing. *Genome Research*, 20(2):265–272, December 2009.

[18] Shili Lin and Jie Ding. Integration of ranked lists via cross entropy monte carlo with applications to mRNA and microRNA studies. *Biometrics*, 65(1):9–18, March 2009.

[19] Witold Litwin and Leo Mark. Nick roussopoulos: Interoperability of multiple autonomous databases. *ACM Computing Surveys*, 1990.

[20] Scott Mcginnis and Thomas L. Madden. Blast: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res*, 32:20–25, 2004.

[21] James C. Mullikin and Zemin Ning. The phusion assembler. *Genome Research*, 13(1):81–90, January 2003.

[22] Eugene W Myers. The fragment assembly string graph. *Bioinformatics*, 21 Suppl 2:ii79–85, 2005.

[23] BioCarta Charting Pathways of Life. http://www.biocarta.com/genes/index.asp.

[24] Tom Oinn, Matthew Addis, Justin Ferris, Darren Marvin, Martin Senger, Mark Greenwood, Tim Carver, Kevin Glover, Matthew R. Pocock, Anil Wipat, and Peter Li. Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics*, 20(17):3045–3054, November 2004.

[25] Mihai Pop. Genome assembly reborn: recent computational challenges. *Brief Bioinform*, 10(4):354–366, July 2009.

[26] 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.

[27] F. Sanger, S. Nicklen, and A. R. Coulson. DNA Sequencing with Chain-Terminating Inhibitors. *PNAS*, 74(12):5463–5467, 1977.

[28] Michael C. Schatz, Arthur L. Delcher, and Steven L. Salzberg. Assembly of large genomes using second-generation sequencing. *Genome Research*, 20(9):1165–1173, September 2010.

[29] Praveen Sethupathy, Benoit Corda, and Artemis G. Hatzigeorgiou. TarBase: A comprehensive database of experimentally supported animal microRNA targets. *RNA (New York, N.Y.)*, 12(2):192–197, December 2005.

[30] Jared T. Simpson, Kim Wong, Shaun D. Jackman, Jacqueline E. Schein, Steven J. Jones, and Inanç Birol. ABySS: a parallel assembler for short read sequence data. *Genome research*, 19(6):1117–1123, June 2009.

[31] L. Stein. Genome annotation: from sequence to biology. 2:493–503+, 2001.

[32] Damian Szklarczyk, Andrea Franceschini, Michael Kuhn, Milan Simonovic, Alexander Roth, Pablo Minguez, Tobias Doerks, Manuel Stark, Jean Muller, Peer Bork, Lars J. Jensen, and Christian von Mering. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic acids research*, 39(Database issue):D561–D568, January 2011.

[33] Wang-Xia Wang, Bernard W Rajeev, Arnold J Stromberg, Na Ren, Guiliang Tang, Qingwei Huang, Isidore Rigoutsos, and Peter T Nelson. The expression of microrna mir-107 decreases early in alzheimers disease and may accelerate disease progression through regulation of beta-site amyloid precursor protein-cleaving enzyme 1. *Journal of Neuroscience*, 28(5):1213–1223, 2008.

[34] P. Xu. The drosophila MicroRNA mir-14 suppresses cell death and is required for normal fat metabolism. *Current Biology*, 13(9):790–795, April 2003.

[35] Daniel R. Zerbino and Ewan Birney. Velvet: algorithms for de novo short read assembly using de bruijn graphs. *Genome research*, 18(5):821–829, May 2008.

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.

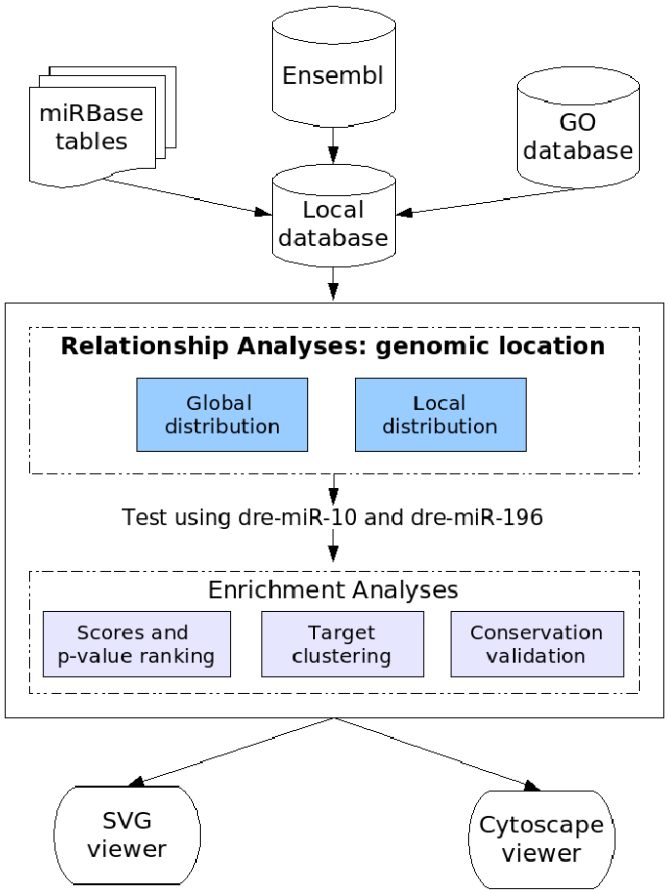


Figure 1: The workflow of the miRNA targets validation method. It consists of three stages: data retrieval, analysis and visualization.

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

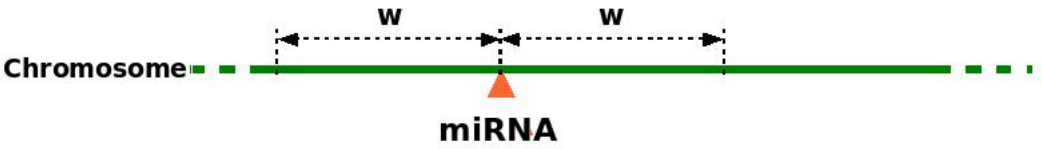


Figure 2: Window size definition

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

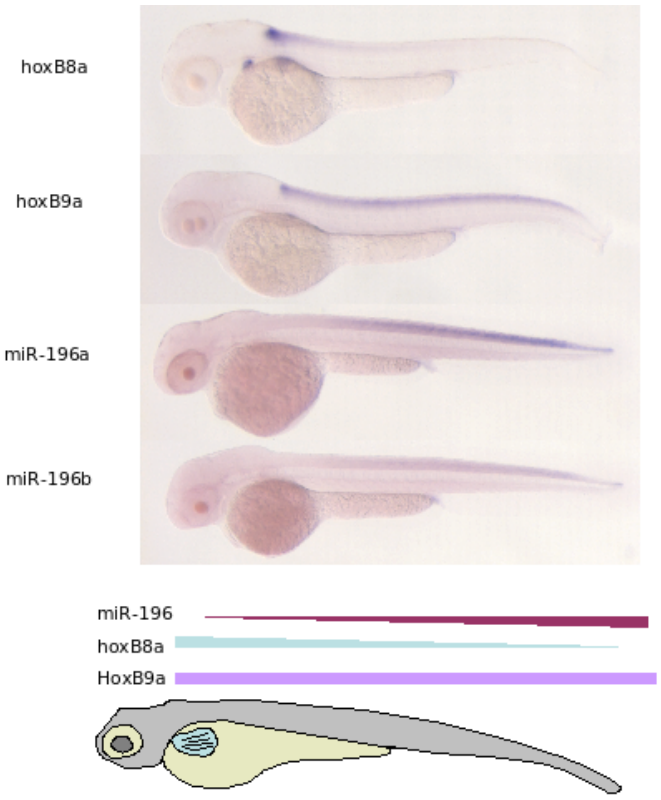
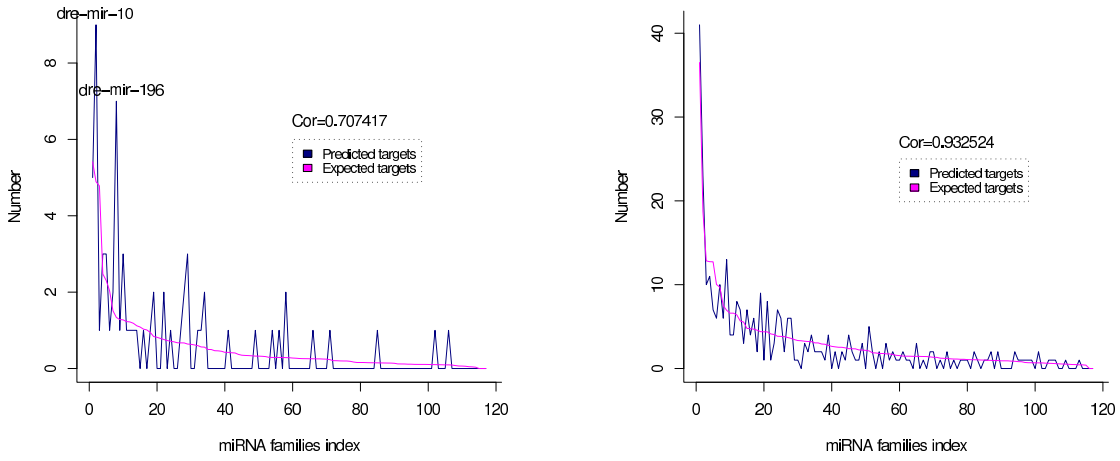


Figure 3: Expression patterns of *hoxb8a*, *hoxb9a* and dre-miR-196a and 196b. It showed the mutually excluding expression patterns for *hoxb8a* and *mir-196* and constant expression of *hoxb9a*.

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.

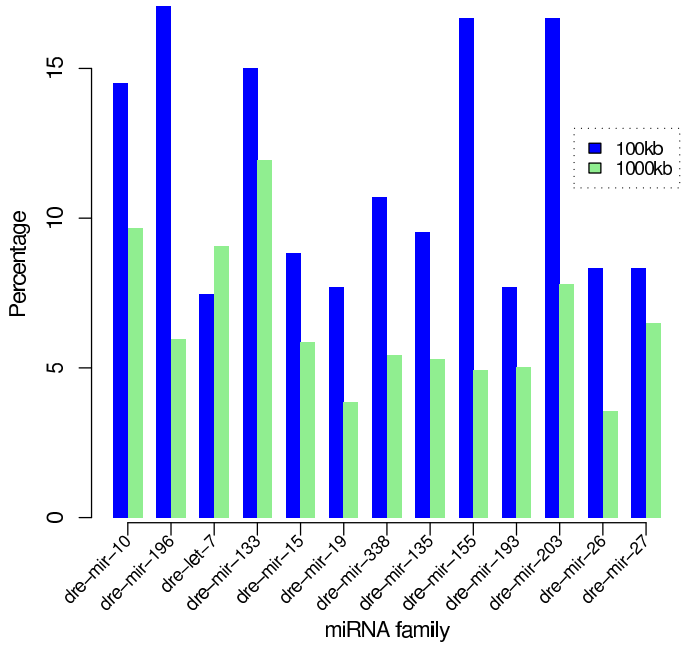
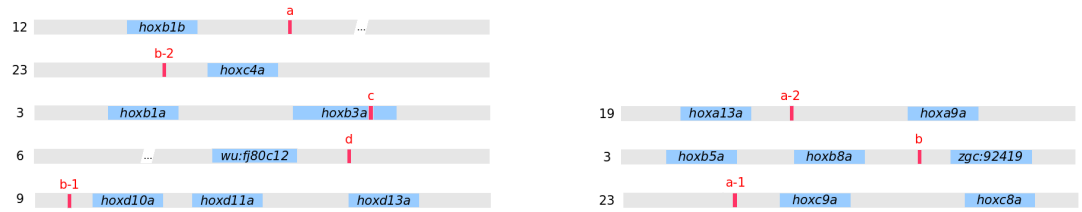


Figure 5: Relative frequency in the window of 100kb and that in 1000kb. It illustrated that 12 out of 13 families have relative frequency in the window of 100kb higher than the one in 1000kb.

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.

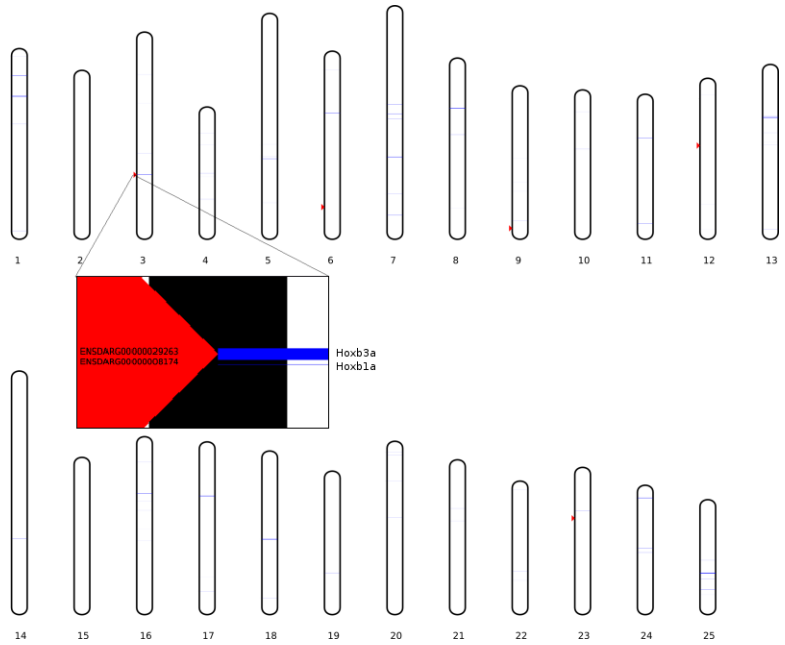


Figure 7: The overview of genomic location of top 50 predicted targets of dre-miR-10 ranked by p-value. The isoforms of dre-miR-10 (red triangles) and targets (blue lines) are displayed over 25 chromosomes (columns). The closeup view illustrated two *hox* genes *hoxb1a* and *hoxb3a* genomic located near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

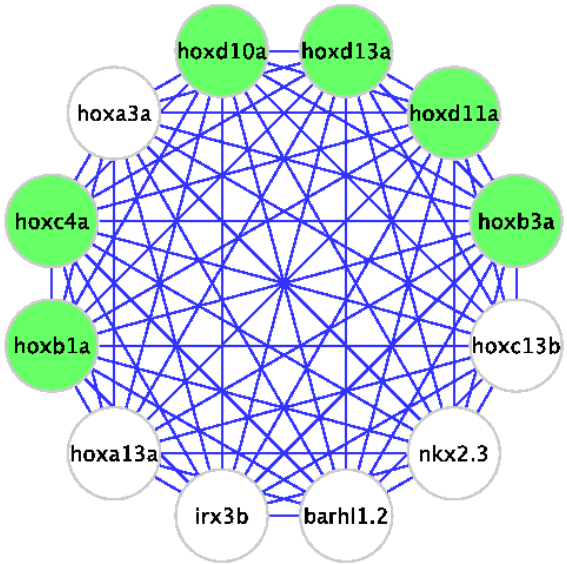


Figure 8: A group of targets for dre-miR-10 which have the same GO term descriptions (expressed by lines) as known targets *hoxb1a* and *hoxb3a*. Within 100kb distance target genes are showed in green.

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.

Chapter 3

miRNA Target Prediction through Mining of miRNA Relationships

Based on

Yanju Zhang, Jeroen S. de Bruin, Fons J. Verbeek. (2010). Specificity enhancement in microRNA target prediction through knowledge discovery. Chapter 20. In: Machine Learning, Eds. Yagang Zhang, ISBN: 978–953–307–033–9, INTECH

Summary

miRNAs are small regulators that mediate gene expression and each miRNA regulates specific target genes. In animals, target prediction of the miRNAs is accomplished through several computational methods, i.e. miRanda, TargetScan and PicTar. Typically, these methods predict targets from features of miRNA-target interaction such as sequence complementarity, free energy of RNA duplexes and conservation of target sites. They are constructed for high throughput and also result in a large amount of predictions and a high estimated false-positive rate. To date, specific rules to capture all known miRNA targets have not been devised. We observed that miRNAs sometimes share targets. Therefore, in this chapter we present an approach which analyzes miRNA-miRNA relationships and utilizes them for target prediction. We use machine learning techniques to reveal the feature patterns between known miRNAs. Different data setups are evaluated and compared to achieve the best performance. Furthermore, the derived rules are applied to miRNAs of which the targets are not yet known so as to see if new targets could be predicted. In the analysis of functionally similar miRNAs, we found that genomic distance and seed similarity between miRNAs are dominant features in the description of a group of miRNAs binding the same target. Application of one specific rule resulted in the prediction of targets for seven miRNAs for which the targets were formerly unknown. Some of these targets were also predicted by other existing methods. Our method contributes to the improvement of target identification by predicting targets with high specificity and without conservation limitation.

1 Introduction

In this chapter we explore and investigate a range of methods in pursue of improving target prediction of microRNA. The currently available prediction methods produce a large output set that also includes a rather high amount of false positives. Additional strategies for target prediction are necessary and we elaborate on one particular group of microRNAs; i.e. those that might bind to the same target. We intend to transfer our approach to other groups of microRNAs as well as the broader application to the important model species.

microRNAs (miRNAs) are a novel class of post-transcriptional gene expression regulators discovered in the genome of plants, animals and viruses. The mature miRNAs are about 22 nucleotides long. They bind to their target messengerRNA (mRNA) and therefore induce translational repression or degradation of target mRNAs [6, 1]. Recent studies have elucidated that these short molecules are highly conserved between species indicating their fundamental roles conserved in evolutionary selection. They are implicated in developmental timing regulation [26], apoptosis [3] and cell proliferation [19]. Some of them even act as potential tumor suppressors [14], potential oncogenes [13] and might be important targets for drugs [23].

The identification of large number of miRNAs existing in different species has increased the interest in unraveling the mechanism of this regulator. It has been proven that more than one miRNA regulates one target and vice versa [6]. Therefore understanding this novel network of regulatory control is highly dependent on identification of miRNA targets. Due to the costly, labor-intensive nature of experimental techniques required, currently, there is no large-scale experimental target validation available leaving the biological function of the majority completely unknown [5]. These limitations of the wet experiments lead to the development of computational prediction methods.

It has been established that the physical RNA interaction requires sequence complementarity and thermodynamic stability. Unlike plant miRNAs, which bind to their targets through near-perfect sequence complementarity, the interaction between animal miRNAs and their targets is more flexible. Partial complementarity is frequently found [6] and this flexibility complicates computation. Lots of effort has been put into characterizing functional miRNA-target pairing. The most frequently used prediction algorithms are

miRanda, TargetScan/TargetScanS, RNAhybrid, DIANA-microT, picTar, and miTarget.

MiRanda [6] is one of the earliest developed large-scale target prediction algorithm which was first designed for Drosophila then adapted for human and other vertebrates. It consists of three steps: First, a dynamic programming local alignment is carried out between miRNAs and 3' UTR of potential targets using a scoring matrix. After filtering by threshold score, the resulting binding sites are evaluated thermodynamically using the Vienna RNA fold package [35]. Finally, the miRNA pairs that are conserved across species are kept.

TargetScan/TargetScanS [22, 21] have a stronger emphasize on the seed region. In the standard version of TargetScan, the predicted target-sites first require a 7-nucleotide (nt) match to the seed region of miRNA, i.e., nucleotides 2-8; second, conservation in 4 genomes (human, mouse, rat and puffer fish), and third, thermodynamic stability. TargetScanS is the new and simplified version of TargetScan. It extends the cross-species comparison to 5 genomes (human, mouse, rat, dog and chicken) and requires a seed match of only 6-nt long (nucleotides 2-7). Through the requirement of more stringent species conservation it leads to more accurate predictions even without conducting free energy calculations.

RNAhybrid [25] was the first method which integrated powerful statistical models for large-scale target prediction. Basically, this method finds the energetically most favorable hybridization sites of a small RNA in a large RNA string. It takes candidate target sequences and a set of miRNAs and looks for energetically favorable binding sites. Statistical significance is evaluated with an extreme value statistics of length normalized minimum free energies for individual hits, a Poisson approximation of multiple hits, and the calculation of effective numbers of orthologous targets in comparative studies of multiple organisms. Results are filtered according to p-value thresholds.

DIANA-microT identified putative miRNA-target interaction using a modified dynamic programming algorithm with a sliding window of 38 nucleotides that calculated binding energies between two imperfectly paired RNAs. After filtering by an energy threshold, the candidates are examined by the rules derived from mutation experiments of a single let-7 binding site. Finally, those which were conserved between human and mouse were further considered for experimental verification [12, 28].

PicTar takes sets of co-expressed miRNAs and searches for combinations of miRNA binding sites in each 3' UTR [17]. And miTarget is a support vector machine classifier for miRNA target-gene prediction, which utilizes a radial basis function kernel to characterize targets by structural, thermodynamic, and position-based features [16].

Among the algorithms discussed previously, miRanda and TargetScan/TargetScanS belong to the sequence-based algorithms which evaluate miRNA-target complementarity first, then calculate the binding site thermodynamics to further prioritize; in contrast, DIANA-microT and RNAhybrid are based on algorithms that are rooted in thermodynamics, thus using thermodynamics as the initial indicator of potential miRNA binding site.

Until now, it remains unclear whether sequence or structure is the better predictor of a miRNA binding site [23]. All of the above mentioned methods produce a large set of predictions and include a relatively high false positive ratio; all in all this indicates that these methods are promising methods but still far away from perfect. The estimated false-positive rate (FPR) for PicTar, miRanda and TargetScan is about 30%, 24-39% and 22-31% respectively [2, 30, 22]. It has been reported that miTarget has a similar performance as TargetScan [16]. In addition to the relatively high FPR, *Enright et al.* observed that many real targets are not predicted by these methods and this seems to be largely due to requirements for evolutionary conservation of the putative miRNA target-site across different species [6, 8]. In general we also notice that in all of these algorithms, the target prediction is based on features that consider the miRNA-target interaction such as sequence complementarity and stability of miRNA-target duplex.

Through the observations in the population of confirmed miRNAs targets we became aware that some miRNAs are validated as binding the same target. For example, in human miR-17 and miR-20a both regulate the expression of E2F1; while miR-221 and miR-222 both bind to KIT. Subsequently, we considered that this observation would allow target identification from the analysis of functionally similar miRNAs.

Based on this idea, we present an approach which analyzes miRNA-miRNA relationships and utilizes them for target prediction. Our aim is to improve target prediction by using different features and discovering significant feature patterns through tuning and combining several machine learning techniques. To this respect, we applied feature selection, principle component analysis, classification, decision trees, and propositionalization-

based relational subgroup discovery to reveal the feature patterns between known miRNAs. During this procedure, different data setups were evaluated and the parameters were optimized. Furthermore, the derived rules were applied to functionally unknown miRNAs so as to see if new targets could be predicted. In the analysis of functionally similar miRNAs, we found that genomic distance, seed and overall sequence similarities between miRNAs are dominant features in the description of a group of miRNAs binding the same target. Application of one specific rule resulted in the prediction of targets for five functionally unknown miRNAs which were also detected by some of the existing methods. Our method is complementary to the existing prediction approaches. It contributes to the improvement of target identification by predicting targets with high specificity and without conservation limitation. Moreover, we discovered that knowledge discovery especially the propositionalization-based relational subgroup discovery, is suitable for this application domain since it can interpret patterns of similar function miRNAs with respect to the limited features available.

The remainder of this chapter is organized as follows. In Section 2, miRNA biology and databasing as well as the background of the machine learning techniques which are the components of our method are explained: i.e., miRNA biogenesis and function, related databases, feature selection, principle component analysis, classification, decision trees and propositionalization-based relational subgroup discovery. Section 3 specifies the proposed method including data preparation, algorithm configuration and parameter optimization. The results are summarized in Section 4. Finally, In Section 5, we discuss the strengths and the weaknesses of the applied machine learning techniques and feasibility of the derived miRNA target prediction rules.

2 Background

The first two subsections are devoted to the exploration of miRNA biology whereas the latter two subsections have a computational nature.

2.1 microRNA biogenesis and function

The mature miRNAs are ∼22 nucleotide single-stranded noncoding RNA molecules. They are derived from miRNA genes. First, miRNA gene is transcribed to primary miRNA transcripts (pri-miRNA), which is between a few hundred or a few thousand base pair long. Subsequently, this pri-miRNA is processed into hairpin precursors (pre-miRNA), which has a length of approximately 70 nucleotides, by the protein complex consisting of the nuclease Drosha and the double-stranded RNA binding protein Pasha. The pre-miRNA then is transported to cytoplasm and cut into small RNA duplexes of approximately 22 nucleotides by the endonuclease Dicer. Finally, either the sense strand or antisense strand can function as templates giving rise to mature miRNA. Upon binding to the active RISC complex, mature miRNAs interact with the target mRNA molecules through base pair complementarity, therefore inhibit translation or sometimes induce mRNA degradation [4].

It is suggested that miRNAs tend to bind 3' UTR (3' Untranslated Region) of their target mRNAs [20]. Further studies have discovered that position 2-8 of miRNAs, which is called 'seed' region, has been described as a key specificity determinant of binding, requires good or perfect complementarity [22, 21]. The process of biogenesis and function of miRNAs are illustrated in Fig. 3, Chapter 1. A detailed miRNA-target interaction is also showed with a highlighted seed region.

2.2 miRNA databases

miRBase: MiRBase is the primary online repository for published miRNA sequence data, annotation and predicted gene targets [9, 10]. It consists of three parts:

The miRBase Registry acts as an independent authority of miRNA gene nomenclature, assigning names prior to publication of novel miRNA sequences.

The miRBase Sequences is a searchable database for miRNA sequence data and annotation. The latest version (Release 13.0, March 2009) contains 9539 entries representing hairpin precursor miRNAs, expressing 9169 mature miRNA products, in 103 species including primates, rodents, birds, fish, worms, flies, plants and viruses.

The miRBase Targets is a comprehensive database of predicted miRNA target genes. The

core prediction algorithm currently is miRanda (version 5.0, Nov 2007). It searches over 2500 animal miRNAs against over 400 000 3' UTRs from 17 species for potential target sites. In human, the current version predicts 34788 targets for 851 human miRNAs.

Tarbase: Tarbase is a comprehensive repository of a manually curated collection of experimentally supported animal miRNA targets [29, 24]. It describes each supported target site by the miRNA which binds it, the target genes which includes this binding site, the direct and indirect experiments that were conducted to validate it, binding site complementarity and etc. The latest version (Tarbase 5.0, Jun 2008) records more than 1300 experimentally supported miRNA target interactions for human, mouse, rat, zebrafish, fruitfly, worm, plant, and virus. As machine learning methods become more popular, this database provides a valuable resource to train and test for machine learning based target prediction algorithms.

2.3 Pattern recognition

Pattern recognition is considered a sub-topic of machine learning. It concerns with classification of data either based on a priori knowledge or based on statistical information extracted from the patterns. The patterns to be classified are usually groups of measurements, features or observations, which define data points in an appropriate multidimensional space. Our pattern recognition proceeds in three different stages: feature reduction, classification and cross-validation.

Feature reduction: Feature reduction includes feature selection and extraction. Feature selection is the technique of selecting a subset of relevant features for building learning models. In contrast, feature extraction seeks a linear or nonlinear transformation of original variables to a smaller set. The reason why not all features are used is because of performance issues, but also to make results easier to understand and more general. Sequential backward selection is a feature selection algorithm. It starts with entire set, and then keeps removing one feature at a time so that the entire subset so far performs the best. Principle component analysis (PCA) is an unsupervised linear feature extraction algorithm. It derives new variables in decreasing order of importance that are a linear combinations of the original variables, uncorrelated and retain as much variation as possible [33].

Classification: Classification is the process of assigning labels on data records based on their features. Typically, the process starts with a training dataset that has examples already classified. These records are presented to the classifier, which trains itself to predict the right outcome based on that set. After that, a testing set of unclassified data is presented to the classifier, which classifies all the entries based on its training. Finally, the classification is being inspected. The better the classifier, the more good classifications it has made. Linear discriminant classifier (LDC) and quadratic discriminant classifiers (QDC) are two frequently used classifiers which separate measurements of two or more classes of objects or events by a linear or a quadric surface respectively.

Cross-validation: Cross-validation is the process of repeatedly partitioning a dataset in a training set and a testing set. When the dataset is partitioned in *n* parts we call that *n*-fold cross-validation. After partitioning the set in *n* parts, the classifier is trained with *n-1* parts, and tested on the remaining part. This process is repeated n times, each time a different part functions as the training part. The n results from the folds then can be averaged to produce a single estimation of error.

2.4 Knowledge discovery

Knowledge discovery is the process which searches large volumes of data for patterns in order to find understandable knowledge about the data. In our knowledge discovery strategy, decision tree and relational subgroup discovery are applied.

Decision tree: The decision tree [34] is a common machine learning algorithm used for classification and prediction. It represents rules in the form of a tree structure consisting of leaf nodes, decision nodes and edges. This algorithm starts with finding the attribute with the highest information gain which best separates the classes, and then it is split into different groups. Ideally, this process will be repeated until all the leaves are pure.

Relational subgroup discovery: Subgroup discovery belongs to descriptive induction [36] which discover patterns described in the form of individual rules. Relational subgroup discovery (RSD) is the algorithm which utilizes relational datasets as input, generates subgroups whose class-distributions differ substantially from the complete dataset with respect to the property of interest [18]. The principle of RSD can be simplified as follows; first, a feature is constructed through first-order feature construction and the features

covering empty datasets are retracted. Second, rules are induced using weighted relative accuracy heuristics and weighted covering algorithm. Finally, the induced rules are evaluated by employing the combined probabilistic classifications of all subgroups and the area under the receiver operating characteristics (ROC) curve [7]. The key improvement of RSD is the application of weighted relative accuracy heuristics and weighted covering algorithm, i.e.

$$WRAcc(H \leftarrow B) = p(B) \cdot (p(H \mid B) - p(H)) \quad (1)$$

The weighted relative accuracy heuristics is defined as equation 1. In rule $H \leftarrow B$, H stands for Head representing classes, while B denotes the Body which consists of one or a conjunction of first-ordered features. p is the probability function. As shown in the equation, weighted relative accuracy consists of two components: weight $p(B)$, and relative accuracy $p(H \mid B) - p(H)$. The second term, relative accuracy, is the relative accuracy gain between the conditional probability of class H given that features B is satisfied and the probability of class H. A rule is only interesting if it improves over this default rule $H \leftarrow true$ accuracy [36].

In the weighted covering algorithm, the covered positive examples are not deleted from the current training set which is the case for the classical covering algorithm. Instead, in each run of the covering loop, the examples are given decreasing weights while the number of iterations is increasing. In doing so, it is possible to discover more substantial significant subgroups and thereby achieving to find interesting subgroup properties of the entire population.

3 Experimental setups, methods and materials

3.1 Data collection

In the interest of including maximally useful data, human miRNAs are chosen as the research focus. The latest version of TarBase (TarBase-V5 released at 06/2008) includes 1093 experimentally confirmed human miRNA-target interactions. Among them, 243 are supposed by direct experiment such as in vitro reporter gene (Luciferase) assay, while the rest are validated by an indirect experimental support such as microarrays. Considering the fact that the indirect experiments could induce the candidates which are in the

downstream of the miRNA involved pathways, it is uncertain whether these can virtually interact with miRNA or not. Thus they are excluded and only the miRNAs-target interactions with direct experiment support are used in this study.

We observed that some miRNAs are validated as binding the same target. According to this observation, we pair the miRNAs as positive if they bind the same target, and randomly couple the rest as the negative data set. In total, there are 93 positive pairs. After checking the consistency of the name of miRNAs and removing the redundant data (for example, miR-26 and miR-26-1 refer to the same miRNA), 73 pairs are kept and thus another 73 negative pairs are generated. For quality control reasons, the data generation step is repeated 10 times and each set is tested individually in the following analysis.

Here we clarify two notions; known miRNAs are those whose function is known and have been validated for having at least one target, unknown miRNAs refer to those for which the targets are unknown.

3.2 Feature collection

In the study of miRNA-target interaction, it has been established that this physical binding requires sequence complementarity and thermodynamic stability. Here some of miRNA-target interaction features are transformed to the study of functionally similar miRNA pairs.

We predefine four features: overall sequence ($\sim$22 nt) similarity, seed (position 2-8) similarity, non-seed (position 9-end) similarity and genomic distance. Seed has been proven to be an important region in miRNA-target interaction which display an almost perfect match to the target sequence [15], thus we suggest that seed similarity between miRNAs is a potentially important feature. Additionally, including non-seed and sequence similarity features enables us to investigate the property behaviors of these two regions. Genomic distance is not a well investigated feature which is defined as base pair distance between two genes. The idea of investigating genomic distance between miRNAs is derived from our former study. Previously, through statistical methods and heterogeneous data support, we demonstrated that the genomic location feature plays a role in miRNA-target interaction for a selection of miRNA families [38]. Here we induce this idea to the study of miRNAs relationships based on the genomic distance.

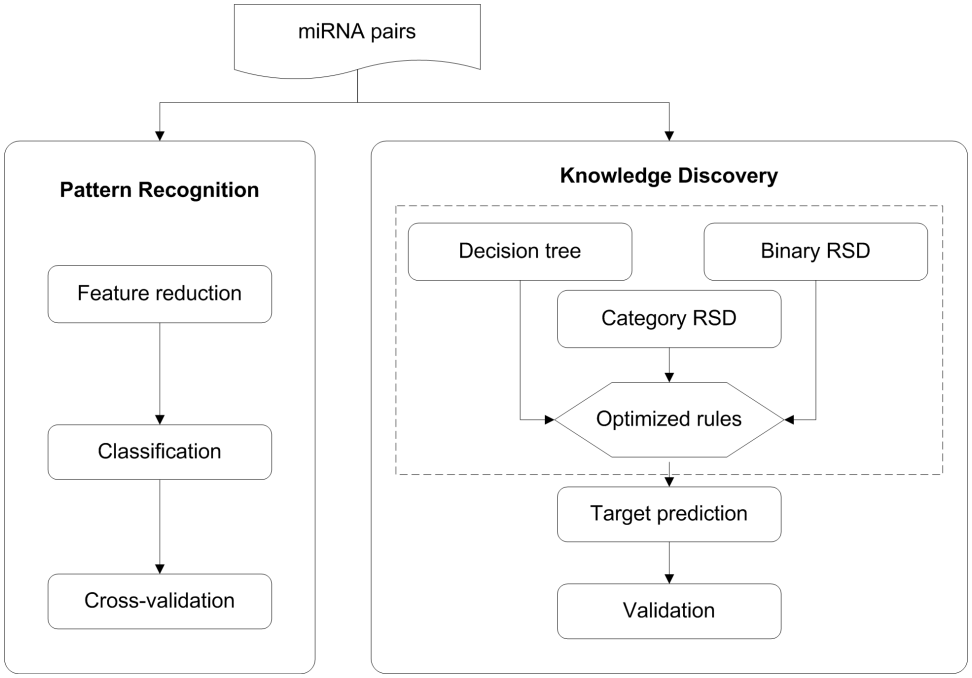


Figure 1: Workflow. miRNA pairs are analyzed by both pattern recognition and knowledge discovery strategies.

In the data preparation, sequence similarity is calculated using the EBI pairwise global sequence alignment tool: i.e. Needle [27]. Genomic sequence and location are retrieved from the miRBase Sequence Database. The distance between two miRNAs is calculated by genomic position subtraction when they are located on the same chromosome; otherwise it is set to undefined.

3.3 Workflow

As showed in Fig. 1, we use two strategies to discover miRNA-miRNA relationships. In pattern recognition strategy, different classifiers are applied to preprocessed dataset in order to discriminate positive and negative miRNA pairs. Then the performance of each classifier is evaluated by cross-validation. In knowledge discovery, rules are first discovered from three methods with respect to decision tree and relational subgroup discovery techniques. Through combining the results, the optimized rules describing functionally alike miRNAs are generated which are used for final targets prediction and validation.

Pattern recognition: In this strategy, the first step is feature reduction. Features are selected by sequential backward elimination algorithm and extracted by principle com-

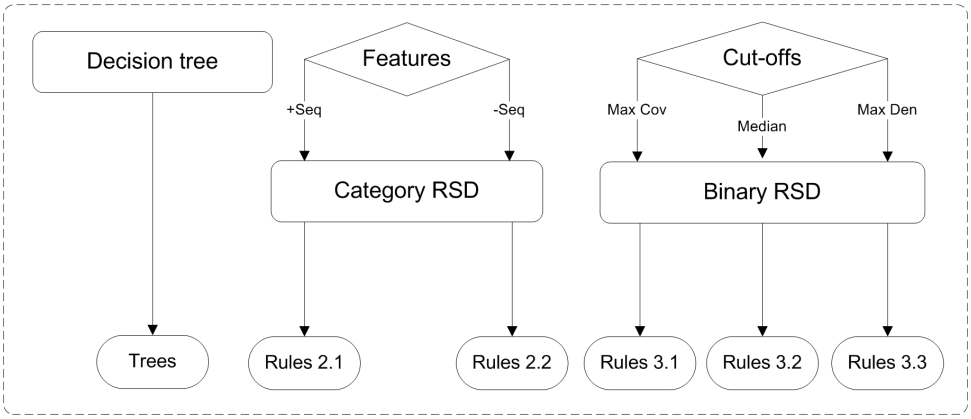
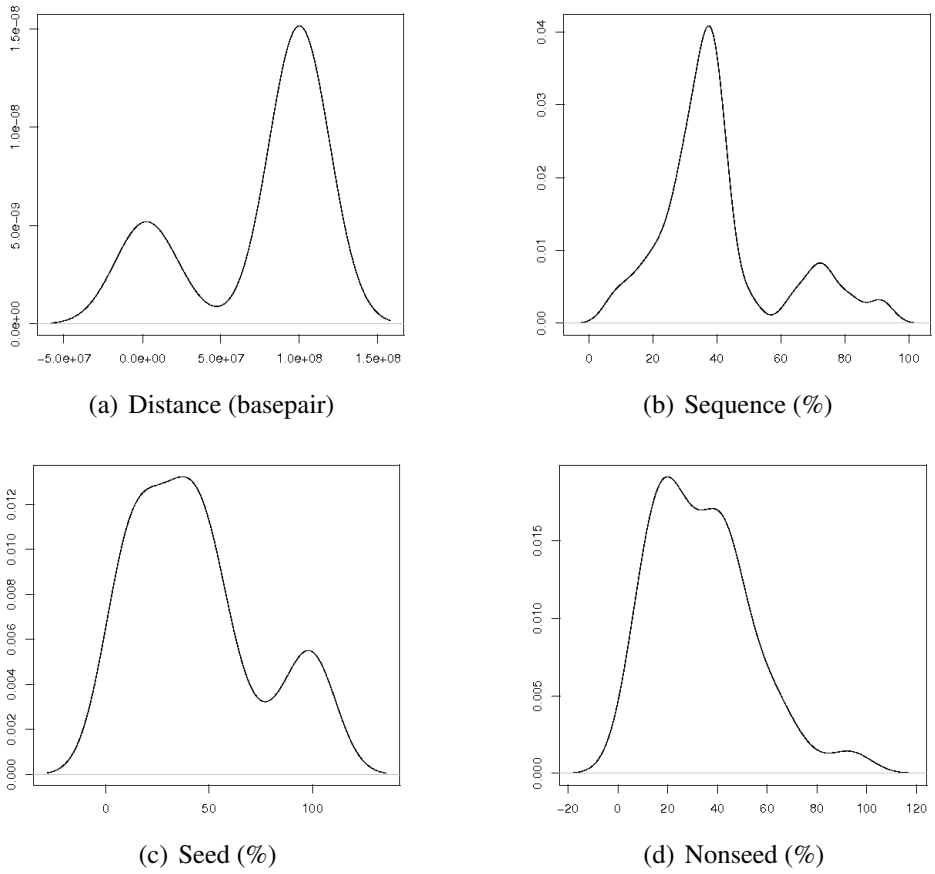


Figure 2: Detailed experimental design in rule generation stage. Three methods are applied which are Decision tree, Category RSD and Binary RSD. In Category RSD, datasets are first categorized into groups. Subsequently, data with two feature sets, which are with and without overall sequence similarity, are used as the input to RSD algorithm. In Binary RSD, feature values are binariezed using decision tree. Due to the fact that data are sampled 10 times, the cut-offs are then established using max coverage (Max Cov), median and max density (Max Den). Finally, RSD is applied to all 3 conditions in order to find out the feature cut-offs, which lead to the most significant rule sets.

ponent analysis. As it is known that sequential forward selection adds new features to a feature set one at a time until the final feature set is reached [33]. It is simple and fast. The reason it is not applied in our experiment is due to the limitation that the selected features could not be deleted from the feature set once they have been added. This could lead to local optimum. After dimension reduction, classification is performed by both linear and quadratic classifiers. Finally, the performance is examined by 5-fold cross-validation with 10 repetitions. This part was implemented with PRtools [32] a plugin for the MatLab platform.

Knowledge discovery: As contrasted to the pattern recognition which classifies miRNA pairs by complicated statistical models, knowledge discovery describes data patterns which allow us gain knowledge about the data. This could promote our understanding of functionally similar miRNAs. Furthermore, integration of this knowledge could finally promote target prediction. In this strategy, there are three phases: rule generation illustrated in the framework (dashed) of Fig. 1, target prediction and validation. In the first step, rules are discovered using decision trees and relational subgroup discovery. With the aim to discover the most significant rules, different data structures and feature thresholds are evaluated and compared. Details are explained in the following sections and an overview of this methodology is shown in Fig. 2.



(a) Distance (basepair)

(b) Sequence (%)

(c) Seed (%)

(d) Nonseed (%)

Figure 3: Density plot for the four features. The plots of distance (a) and seed similarity (c) match bimodal distribution indicating two main groups in each feature. However it is not straightforward to judge sequence (b) and nonseed similarity (d) distributions.

Decision tree learning is utilized as a first step in order to build a classifier discriminating two classes of miRNA pairs. In our experiments, we used the decision tree from the Weka software platform [34]. The features were tested using the J48 classifier and evaluated by 10 fold cross-validation.

Due to the fact that not all the determinant features are known at this stage, we are interested in finding rules for subgroups of functionally similar miRNAs with respect to our predefined features. In our experiments, we used the propositionalization based relational subgroup discovery algorithm [36]. We prefer rules that contain only the positive pairs and portray high coverage. Consequently, the repetitive rules are selected, if their E-value is greater than 0.01 and at the same time the significance is above 10.

Both the Category RSD and the Binary RSD reveal feature patterns by utilizing the rela-

tional subgroup discovery algorithm. The main difference is that the former analyzes the data in a categorized format, whereas in later algorithm the data is transformed to a binary form.

As a pilot experiment for RSD, data is first categorized as follows: the similarity percentage is evenly divided into 5 groups: very low (0-20%], low (20-40%], medium (40-60%], high (60-80%], very high (80-100%]; Distance is categorized into 5 regions: 0-1kb[1], 1-10kb, 10-100kb, 100kb-end, undef (if miRNAs that are paired are located on a different chromosome). Two relational input tables, which are with and without the overall sequence similarity feature, are constructed and further tested with the purpose of verifying whether the sequence has a global effect or only contributes as the combination of seed and non-seed parts.

Through the observation of density graphs of the features, as depicted in Fig. 3, we concluded that distance and seed similarity feature densities match a bimodal distribution. The same conclusion can, however, not be drawn easily for overall and non-seed sequence similarities. Therefore, in this method, we apply a decision tree algorithm to discriminate 4 feature values into binary values. Each feature is calculated individually and only the root classifier value in the tree is used for establishing the cut-off. After that, binary tables are generated according to three criteria:

- Maximum coverage where the value covers the most positive pairs. Max coverage (distance, sequence, seed, non-seed) = 8947013 b, 56.5%, 71.4%, 53.3%

- Median. Median (distance, sequence, seed, non-seed) = 3679 b, 65.2%, 71.4%, 60.65%

- Maximum density which is the region with the highest positive pair density. Max density (distance, sequence, seed, non-seed) = 3679 b, 69.6%, 75%, 64.7%

[1]Distance unit is base pair abbreviated as b, kb = kilo base pairs.

4 Results

4.1 Classification

After application of sequential backward feature selection, features including genomic distance, seed similarity and non-seed similarity are selected as the top 3 informative features. Sequence similarity is the least informative feature because it is highly correlated to seed and non-seed similarities. Scatter plots of two classes of miRNA pairs in the selected feature space are depicted in Fig. 4. As can be seen in the four sub-graphs of Fig. 4, the majority of positive and negative miRNA pairs are overlapping which is an indication for the complexity of the classification. The distribution of negative class is more compact. We observed that the majority of this class located in the area of non-seed$<$60%, seed$<$70% and distance is infinite. Furthermore, we noticed that for those functionally similar miRNAs, seed similarity vary from 0 to 100%. This implies that miRNAs with the same or different seed sequence can bind the same targets. This is due to the fact that miRNAs can bind to the same targets at the same binding site which leads to high similarity and at different binding site resulting low similarity. The evaluation of the classifier performance shows that the average error and standard deviation for the quadratic classifier are 0.29739 and 0.01082, and for the linear classifier are 0.30987 and 0.0131.

In Fig. 5 the dataset is plotted in 2-dimensional PCA space in combination with the linear and quadratic classifiers. In this projected 2D space, the average error and standard deviation for the quadratic classifier are 0.3029 and 0.00721, and for the linear classifier are 0.31657 and 0.00871.

With around 30% of classification errors, this means two classes are difficult to separate using features currently available. Furthermore, although the classifiers provide a statistical explanation and meaning, no biological insight is gained from them in order to be able to interpret the miRNA mechanism(s).

4.2 Rule discovery

In the decision tree analysis, several different tree structures are generated from 10 replications of the training data. Among them, the root attribute or the first depth of the tree is

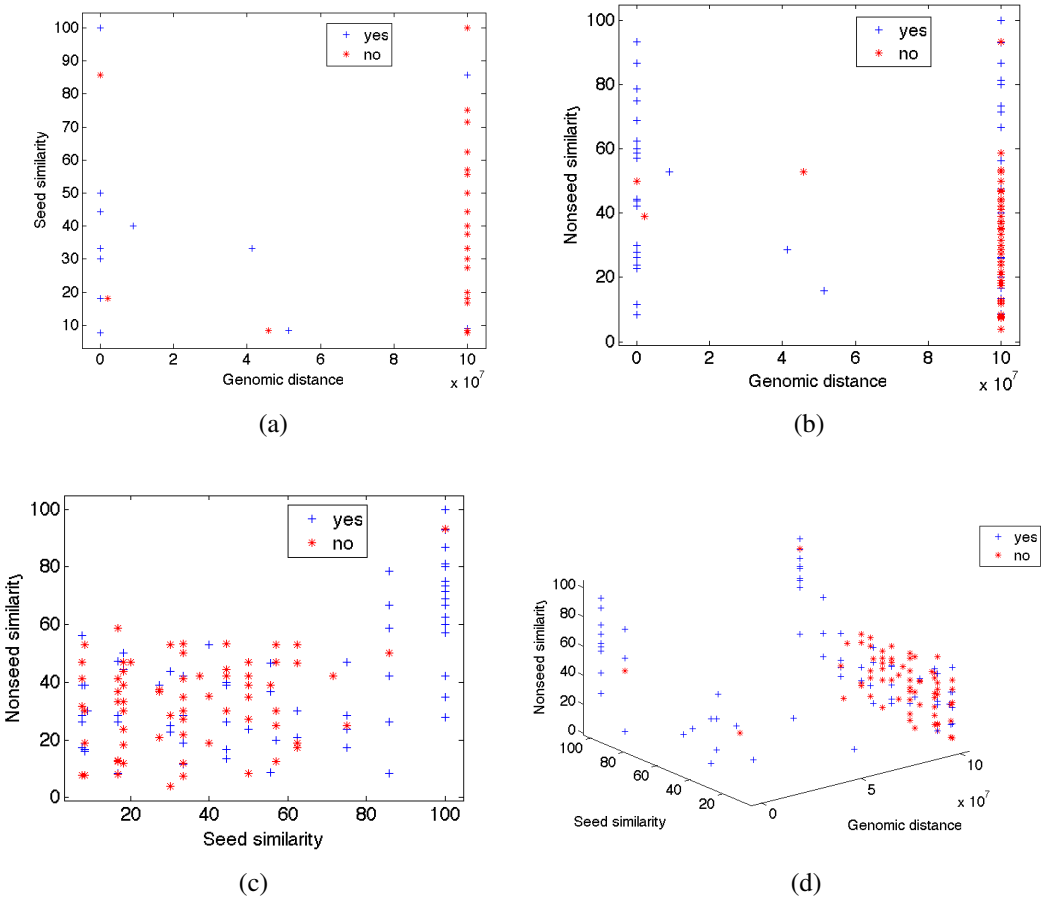


Figure 4: Scatter plots of two classes of miRNA pairs in the selected feature spaces. Positive pairs are denoted using a token of plus (blue); negatives are demonstrated by asterisk (red).

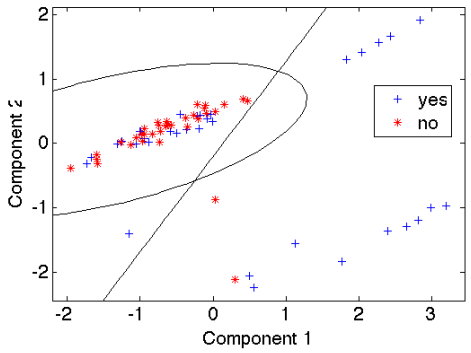


Figure 5: Scatter plot of two classes of miRNA pairs in a 2D PCA space together with a linear discriminant and a quadratic discriminant classifier showed by a line and an arc respectively.

Table 1: Category RSD results. Rules generated from two data structures: considering overall sequence, seed, non-seed similarities as well as distance (a) and only seed, non-seed similarities and distance (b).

(a)

Label	-Overall sequence : YES Rules 2.1	Significance
A	Seed>80%	26.7
A.1	Dis=undef & Seed>80%	14.3
B	Dis≤1 kb	14.1
A.2	Seed>80% & Nonseed=(60%,80%]	12.6
C	Dis=(1 kb,10 kb]	11

(b)

Label	+Overall sequence : YES Rules 2.2	Significance
A	Seed>80%	26.7
A.1	Dis=undef & Seed>80%	14.3
B	Dis≤1 kb	14.1
A.2	Seed>80% & Nonseed=(60%,80%]	11.2
C	Dis=(1 kb,10 kb]	11

mainly associated with distance, sequence and seed similarity properties, while non-seed feature appeared only near the leaf nodes. This inconsistency in the tree structures indicated that none of the predefined features, or any combination of them, can significantly classify miRNAs.

The feature patterns discovered from Category RSD are listed in Table 1 where the rules in Table 1(b) take overall sequence into account but those in Table 1(a) do not. 'YES'-rules describe functionally similar miRNAs characterized by our predefined features. 'Significance' denotes the average significance over 10 replications. Further inspection of Table 1 shows that both rule sets consist of 3 main groups with features being Seed>80%, Dis≤1 kb and Dis=(1 kb,10 kb] labeled by A, B, C respectively. The remainder is the subset of these groups. Considering overall sequence in the rule generation results only the fourth rule (A.2) in Table 1(a) and 1(b) to be different. These results indicate that genomic location and seed similarity between miRNAs are probably dominant features when deciding which miRNAs bind to the same target. Sequence information may be relevant but it is

Table 2: Binary RSD results. Rules generated from 3 sets of parameters are shown in a sequence of Max coverage (a), Median (b) and Max density (c).

(a)

Label	Max coverage: YES Rules 3.1	Significance
A.1	Seed>71.4% & Seq>56.5%	30
A	Seed>71.4%	27.2
A.2	Nonseed>53.3% & Seed>71.4% & Seq>56.5%	21.6
B	Dis≤8947013 b	19.8
A.3	Nonseed>53.3% & Seed>71.4%	18.2
A.4	Dis>8947013 b & Seed>71.4% & Seq>56.5%	13.5
A.5	Dis>8947013 b & Seed>71.4%	12.3

(b)

Label	Median: YES Rules 3.2	Significance
A	Seed>71.4%	27.2
A.1	Seed>71.4% & Seq>65.2%	23.3
B	Dis≤3679 b	23.3
B.1	Dis≤3679 b & Nonseed≤60.65%	15.9
A.2	Dis>3679 b & Seed>71.4%	14.9
A.3	Nonseed>60.65% & Seed>71.4% & Seq>65.2%	13.7
A.4	Nonseed>60.65% & Seed>71.4%	13.7
C.1	Nonseed>60.65% & Seq>65.2%	13.7
C	Seq>65.2%	12.2

(c)

Label	Max density: YES Rules 3.3	Significance
A	Seed>75%	26.7
B	Dis≤3679 b	23.3
A.1	Seed>75% & Seq>69.6%	20.8
C	Seq>69.6%	20.8
B.1	Dis≤3679 b & Nonseed≤64.7%	18
B.2/C.1	Dis≤3679 b & Seq≤69.6%	14.1
A.2	Dis>3679 b & Seed>75%	11.5
A.3/C.2	Nonseed>64.7% & Seed>75% & Seq>69.6%	11
C.3	Nonseed>64.7% & Seq>69.6%	11

not as strong as seed and distance features.

Table 2 shows the rules generated by Binary RSD, thereby using three cutoff criteria: Max coverage (a), Median (b) and Max density (c). As can be seen, three rule sets have similar structures but different feature cutoffs which lead to different significance. The main feature groups derived using max coverage, median and max density criteria respectively are Seed>71.4% (A) and Dis≤8947013 b (B) in rule set 3.1; Seed>71.4% (A), Dis≤3679 b (B) and Seq>65.2% (C) in rule set 3.2; and Seed>75% (A), Dis≤3679 b (B) and Seq>69.6% (C) in rule set 3.3. Others are the subsets of these groups.

Furthermore, the rules with similar features but different feature values are compared. The decision on final cut-off is based on the value which results in the highest significance. Therefore the final optimized rules are:

Rule 1: IF distance between two miRNAs ≤3679 b,
Rule 2: IF seed similarity between two miRNAs > 71.4%,
Rule 3: IF sequence similarity between two miRNAs > 69.6%
THEN they bind the same target.

To evaluate our methods, as a reference, a permutation test is performed. We repeat the learning procedure for each training set with the labels randomly shuffled. Using Max coverage as a cutoff criterion, we obtained that all the rules have the max significance lower than 8. This test therefore demonstrates that the rules derived from the original data are more significant compared to the random situation.

4.3 Target prediction

We apply the above rules searching for miRNAs which serve similar functions as the known miRNAs. Rule 1, 2 and 3 discovered 75, 655 and 150 miRNA pairs respectively in each subgroup which highly extends our previous findings [37] based on the similar methodology. Among them, 23 miRNA predicted targets which are covered by all of the 3 rules are selected for further validation. Since this group has relative small pairs which are easy to validate. Furthermore, as they involve more constraints, it is considered to be more reliable.

Further observation of these 23 miRNA pairs, we found that it consists of 3 confirmed pairs in which both miRNAs from each pair are well studied, 15 pairs with both members

Table 3: Informatic validation of confirmed and predicted miRNA pairs. miRNA1 and miRNA2 are the partners in one pair. Target column shows the validated targets for the known miRNAs (in italic) and the predicted targets for the unknown miRNAs (in boldface). m1 and m2 columns denote whether the targets are predicted by the existing methods for miRNA1 (m1) and miRNA2 (m2) respectively.

miRNA1	Our prediction miRNA2	Target	Targets predicted by TargetScan		MiRanda		Pictar		miTarget		RNAhybrid-mfe kcal/mol	
(m1)	(m2)		m1	m2	m1	m2	m1	m2	m1	m2	m1	m2
hsa-miR-15a	*hsa-miR-16*	*BCL2*	√	√	×	×	√	√	×	√	-24.3	-24.1
hsa-miR-17	*hsa-miR-20a*	*E2F1*	√	√	√	×	√	√	√	√	-26.8	-24.6
hsa-miR-221	*hsa-miR-222*	*KIT*	√	√	×	×	×	×	√	√	-24.9	-26.4
hsa-miR-17	**hsa-miR-18a**	*E2F1*	√	×	√	×	√	×	√	√	-26.8	-26.8
		AIB1	-	-	-	-	-	-	√	√	-26.3	-26.6
hsa-miR-106a	**hsa-miR-18b**	*RB1*	√	×	×	×	√	×	×	×	-23.2	-28.3
hsa-miR-106a	**hsa-miR-20b**	*RB1*	√	√	×	×	√	√	×	×	-23.2	-27.2
hsa-miR-132	**hsa-miR-212**	*RICS*	×	×	-	-	√	√	-	-	-	-
hsa-miR-141	**hsa-miR-200c**	*Clock*	×	×	√	√	×	×	√	×	-22.1	-20.1

from the same family which are supposed to have the same targets, and 5 new pairs which have one well-studied miRNA and one functional unknown partner. Therefore, we induce the targets for these 5 unknown miRNAs hsa-miR-18a/ 18b/ 20b /212 /200c from their known partner. Their predicted targets are listed in Table 3.

Informatic validation is performed to check the prediction consistency with the existing methods. Table 3 shows validation for the 3 confirmed and 5 predicted miRNA pairs. The miRNAs with confirmed targets are indicated in italic, while the miRNAs in boldface are the unknown ones for which the targets are predicted from their known partners. All of their targets are validated by examining whether they are predicted by TargetScan, miRanda, Pictar, miTarget and RNAhybrid. For example the table can be read as following: whether the target (BCL2) is predicted by the existing methods (TargetScan) for m1 (hsa-miR-15a) or m2 (hsa-miR-16). Consequently, we discover that among our prediction, Retinoblastoma 1 (RB1) for hsa-miR-20b are predicted by TargetScan and Pictar; Circadian Locomoter Output Cycles Kaput (Clock) for hsa-miR-200c is captured by miRanda; Rho GTPase activating protein (RICS) for hsa-miR-212 is detected by Pictar; E2F transcription factor 1 (E2F1) and AIB1 for hsa-miR-18a are identified by miTarget.

5 Conclusions and discussion

Machine learning is widely used in commercial businesses which produce vast amounts of data. The life-sciences, molecular oriented research in particular, is a rapidly growing field which has gained a lot of attention lately especially now that the genomes of the major research model species have been sequenced and are publicly available. With the development of more and more large-scale and advanced techniques in biology, the need to discover hidden information triggered the application of machine learning in the field of the life-sciences. But these applications bear a risk, since, first of all, because most biological mechanisms are not yet fully understood, and second, some techniques produce too little experimental data due to the limitations of these techniques, thereby making machine learning unreliable. In this chapter, we explained how we integrated different machine learning algorithms and tuned and optimized experimental setups to a growing but not yet mature research field, miRNA target prediction. The innovation of this approach is not only integration and optimization of machine learning algorithms, but also the prediction through new features in miRNA relationship instead of widely studied features of miRNA-target interaction. Existing methods for analysis have shown to be insufficient in identifying targets from this perspective.

As illustrated in the methods and results sections, pattern recognition generates models enabling class descriptions. In this case, a rather high misclassification error around 30% is surfacing. In contrast, subgroup discovery aims at discovering statistically unusual patterns of interesting classes [36]. It discovers three main groups describing only the positive miRNA pairs.

One of the disadvantages of pattern recognition method is that the model is not biologically interpretable. Consisting of linear or quadratic transformations of features, the classifiers tell nothing about the mechanisms of miRNA-target binding. However decision tree and relational subgroup discovery are descriptive induction algorithms which discover patterns in the form of rules. With these discovered rules, we gain knowledge about miRNA-target interaction which can be used to predict more targets.

We compared two main algorithmic approaches used in knowledge discovery. Given the circumstances that not all the targets and useful features are known in advance, the classification of miRNA data using decision trees is not recommended. However, the

relational subgroup discovery, an advanced subgroup discovery algorithm, has shown to be suitable for this application domain since it can discover the rules for subgroups of similar function miRNAs with respect to our predefined features. During the rule mining, we also noticed that feature threshold optimization is a crucial procedure which helps revealing the significant rules.

We have established that distance, seed and sequence similarities are determinants. The question is whether it makes sense from the biological point of view. It has been reported that many miRNAs appear in clusters on a single polycistronic transcript [31]. They are transcribed together in a long primary transcript, yielding one or more hairpin precursors and finally are cut to multi-mature miRNAs. *Tanzer et al.* reported that the human mir-17 cluster contains six precursor miRNA (mir-17/ 18/ 19a/ 20/ 19b-1/ 92-1) within a region of about 1kb on chromosome 13 [31]. These observations are similar with the feature embedded in Rule 1 (cf. Section 4.2). Besides the fact that clustered miRNAs can be transcribed together, we further showed that miRNAs that are in close proximity to each other can bind to the same target so as to serve as the regulators for the same goal. In this study, we showed that the genomic location also contributes to miRNA target identification.

As for seed similarity, Rule 2 (cf. Section 4.2) describes that the miRNAs with seed similarity above 71.4% share the same targets. This means only a perfect match or one mismatch in the seed is allowed in the process of binding the same targets. This is consistent with the idea that seed is a specific region, in particular it requires a nearly perfect match with the target [15]. Moreover, TargetScanS also only requires a 6-nt seed match comprising nucleotides 2-7 of the miRNA. Thus, the rule requiring at least 6 out of 7 nucleotides to be similar in seed region can be considered reasonable.

Overall sequence similarity is also a predictor but not as decisive as seed and genomic distance. This means that not only the seed region is important; sometimes two miRNAs with generally similar sequences can also bind to the same target. This is consistent with the finding that some miRNA-target interaction bindings have a mismatch or wobble in the 5' seed region but compensate through excellent complementarity at the 3' end, which leads to high average sequence complementarity [23].

In order to support our findings, we validated the results using five existing algorithms presented in Table 3. Not all of the predicted targets are identified by TargetScan, miRanda,

Pictar, miTarget and RNAhybrid, whereas this is the same case for the known targets. Most of the candidates are predicted by at least one of these methods. Both miTarget and our method are based on machine learning techniques; miTarget uses a support vector machine and considers sequence and structure features of miRNA-target duplexes whereas we focus the integration of several machine learning algorithms on the genomic location and sequence features between miRNAs. Moreover, we noticed that miRanda has a relatively low performance for target prediction in human. This may be due to the fact that miRanda was initially developed to predict miRNA targets in *Drosophila melanogaster*, and later adapted to vertebrate genomes [6]. In the application of RNAhybrid tool, a predefined threshold of the normalized minimum free energy (mfe) is lacking, we therefore decided to list the original values. We found that most of our predicted miRNA-target duplexes are more stable illustrated by the lower minimum free energy relative to the known ones.

In addition to these encouraging results, we also noticed that only groups of miRNA relationships are discovered by our method. Some miRNAs which are located far apart and whose seed similarity is low still have the same target. This indicated that besides genomic distance, seed and sequence similarities, more features need to be included in order to find more and better patterns shared by functionally alike miRNAs. *Grimson et al.* uncovered five general features of target site context beyond seed pairing that boost site efficacy [11]. In future research we will explore the site context in the miRNA relationship analysis. Additionally, we also consider taking into account miRNA co-expression patterns.

In summary, we conclude that genomic distance, seed and sequence similarities are the determinants for describing the relationships of functionally similar miRNAs. Our method is complementary to the approaches that are currently used. It contributes to the improvement of target identification by predicting targets with high specificity. Moreover, it does not require conservation information for classification, so it is free from the limitations of some of the existing methods. In future research, with more biologically validated targets and features available, more rules can be generated from a large dataset, and consequently more targets can be identified to the functionally unknown miRNAs. The methodology can be transferred to a broad range of other species as well.

Acknowledgements

We would like to thank Dr. Erno Vreugdenhil for discussing some biological implications of the results and Peter van de Putten for suggestions on the use of WEKA. This research has been partially supported by the BioRange program of the Netherlands BioInformatics Centre (BSIK grant).

References

[1] David P. Bartel. Microrna: Genomics, biogenesis, mechanism, and function. *Cell*, 116(2):281–297, January 2004.

[2] I. Bentwich. Prediction and validation of microrna and their targets. *FEBS Lett*, 579(26):5904–5910, October 2005.

[3] 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.

[4] C. Z. Chen. MicroRNAs as oncogenes and tumor suppressors. *N Engl J Med*, 353(17):1768–1771, October 2005.

[5] A. J. Enright and S. Griffiths-Jones. mirbase: a database of microrna sequences, targets and nomenclature. In Krishnarao Appasani, Sidney Altman, and Victor R. Ambros, editors, *MicroRNAs:From Basic Science to Disease Biology*, pages 157–171. Cambridge University Press, 2007.

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

[7] T. Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, June 2006.

[8] Martin G, Schouest K, Kovvuru P, and Spillane C. Prediction and validation of microrna targets in animal genomes. *J Biosci*, 32(6):1049–1052, September 2007.

[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(Database issue), January 2006.

[10] Sam Griffiths-Jones. The microRNA registry. *Nucleic acids research*, 32(Database issue), January 2004.

[11] Andrew Grimson, Kyle Kai-How K. Farh, Wendy K K. Johnston, Philip Garrett-Engele, Lee P P. Lim, and David P P. Bartel. Microrna targeting specificity in mammals: Determinants beyond seed pairing. *Mol Cell*, 27(1):91–105, July 2007.

[12] D. Grun and N. Rajewsky. Computational prediction of microrna targets in vertebrates, fruitflies and nematodes. In Krishnarao Appasani, Sidney Altman, and Victor R. Ambros, editors, *MicroRNAs:From Basic Science to Disease Biology*, pages 172–186. Cambridge University Press, 2007.

[13] Lin He, Michael M. Thomson, Michael T. Hemann, Eva Hernando-Monge, David Mu, Summer Goodson, Scott Powers, Carlos Cordon-Cardo, Scott W. Lowe, Gregory J. Hannon, and Scott M. Hammond. A microrna polycistron as a potential human oncogene. *Nature*, 435(7043):828–833, 2005.

[14] S. M. Johnson, H. Grosshans, J. Shingara, M. Byrom, R. Jarvis, A. Cheng, E. Labourier, K. L. Reinert, D. Brown, and F. J. Slack. Ras is regulated by the let-7 microrna family. *Cell*, 120(5):635–647, March 2005.

[15] Fedor V. Karginov, Cecilia Conaco, Zhenyu Xuan, Bryan H. Schmidt, Joel S. Parker, Gail Mandel, and Gregory J. Hannon. A biochemical approach to identifying microrna targets. *Proceedings of the National Academy of Sciences*, pages 19291–19296, November 2007.

[16] Sung-Kyu Kim, Jin-Wu Nam, Je-Keun Rhee, Wha-Jin Lee, and Byoung-Tak Zhang. mitarget: microrna target-gene prediction using a support vector machine. *BMC Bioinformatics*, 7:411+, September 2006.

[17] Azra Krek, Dominic Grün, Matthew N. Poy, Rachel Wolf, Lauren Rosenberg, Eric J. Epstein, Philip Macmenamin, Isabelle d. da Piedade, Kristin C. Gunsalus, Markus Stoffel, and Nikolaus Rajewsky. Combinatorial microrna target predictions. *Nature Genetics*, 37(5):495–500, April 2005.

[18] Nada Lavrac, Filip Zelezny, and Peter A/ Flach. Rsd: Relational subgroup discovery through first-order feature construction. In *Proceedings of the 12th International Conference on Inductive Logic Programming*, pages 149–165. Springer-Verlag, July 2003.

[19] 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.

[20] 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.

[21] B. P. Lewis, C. B. Burge, and D. P. Bartel. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microrna targets. *Cell*, 120(1):15–20, January 2005.

[22] Benjamin P. Lewis, I-Hung Shih, Matthew W. Jones-Rhoades, David P. Bartel, and Christopher B. Burge. Prediction of mammalian microrna targets. *Cell*, 115(7):787–798, December 2003.

[23] Pierre Mazière and Anton J. Enright. Prediction of microRNA targets. *Drug discovery today*, 12(11-12):452–458, June 2007.

[24] Giorgos L. Papadopoulos, Martin Reczko, Victor A. Simossis, Praveen Sethupathy, and Artemis G. Hatzigeorgiou. The database of experimentally supported targets: a functional update of TarBase. *Nucleic acids research*, 37(Database issue):gkn809+, January 2009.

[25] M. Rehmsmeier, P. Steffen, M. Hochsmann, and R. Giegerich. Fast and effective prediction of microrna/target duplexes. *RNA*, 10(10):1507–1517, October 2004.

[26] 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.

[27] David Sankoff and Joseph Kruskal. *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*. Center for the Study of Language and Inf, December 1999.

[28] P. Sethupathy, M. Megraw, and A. G. Hatzigeorgiou. Computational approaches to elucidate mirna biology. In Krishnarao Appasani, Sidney Altman, and Victor R. Ambros, editors, *MicroRNAs:From Basic Science to Disease Biology*, pages 187–198. Cambridge University Press, 2007.

[29] Praveen Sethupathy, Benoit Corda, and Artemis G. Hatzigeorgiou. Tarbase: A comprehensive database of experimentally supported animal microrna targets. *RNA*, 12(2):192–197, February 2006.

[30] Praveen Sethupathy, Molly Megraw, and Artemis G. Hatzigeorgiou. A guide through present computational approaches for the identification of mammalian microrna targets. *Nature Methods*, 3(11):881–886, October 2006.

[31] Andrea Tanzer and Peter F. Stadler. Molecular evolution of a MicroRNA cluster. *Journal of Molecular Biology*, 339(2):327–335, May 2004.

[32] Ferdinand van der Heijden, Robert Duin, Dick de Ridder, and David M. J. Tax. *Classification, Parameter Estimation and State Estimation: An Engineering Approach Using MATLAB*. Wiley, 1 edition, November 2004.

[33] Andrew R. Webb. *Statistical Pattern Recognition, 2nd Edition*. John Wiley & Sons, October 2002.

[34] Ian H. Witten and Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, October 1999.

[35] S. Wuchty, W. Fontana, I. L. Hofacker, and P. Schuster. Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers*, 49(2):145–165, February 1999.

[36] Filip Zelezny and Nada Lavrac. Propositionalization-based relational subgroup discovery with rsd. *Machine Learning*, 62(1-2):33–63, February 2006.

[37] Yanju Zhang, J. S. de Bruin, and Fons J. Verbeek. mirna target prediction through mining of mirna relationships. *BioInformatics and BioEngineering*, pages 1–6, 2008.

[38] Yanju Zhang, Joost M. Woltering, and Fons J. Verbeek. Screen of microrna targets in zebrafish using heterogeneous data sources: A case study for dre-mir-10 and dre-mir-196. *International Journal of Mathematical, Physical and Engineering Sciences*, 2(1):10–18, November 2007.

Chapter 4

Comparison and Integration of Target Prediction Algorithms for microRNA Studies

Based on

Yanju Zhang and Fons J. Verbeek. (2010). Comparison and Integration of Target Prediction Algorithms for microRNA Studies. Journal of Integrative Bioinformatics, 7(3):127.

Summary

microRNAs are short RNA fragments that have the capacity of regulating hundreds of target gene expression. Currently, due to lack of high-throughput experimental methods for miRNA target identification, a collection of computational target prediction approaches have been developed. However, these approaches deal with different features or factors are weighted differently resulting in diverse range of predictions. The prediction accuracy remains uncertain. In this chapter, three commonly used target prediction algorithms are evaluated and further integrated using algorithm combination, ranking aggregation and Bayesian Network classification. Our results revealed that each individual prediction algorithm displays its advantages as was shown on different test data sets. Among different integration strategies, the application of Bayesian Network classifier on the features calculated from multiple prediction methods significantly improved target prediction accuracy.

1 Introduction

microRNAs (miRNAs) are a class of novel post-transcriptional gene expression regulators which are involved in a variety of developmental, physiological or disease-associated cellular processes [1]. They bind to their targets, messenger RNAs (mRNAs). This binding marks the targets for degradation or translation inhibition [3, 2]. In the past years, around 500 miRNA genes have been discovered in human. Functional annotations are, however, available only for a small fraction of these miRNAs [5]. This fact leaves the mechanism of miRNA-mediated gene regulation largely unknown.

One crucial aspect of the functional annotation of miRNAs is the identification of the miRNA targets with which they directly interact [15]. Due to the limitations of the current techniques, high-throughput target validation via biological experiments is not practical. Given these circumstances, a lot of algorithms for computational target prediction have been developed.

Each algorithm has a key focus. On the basis of this the prediction algorithm can be categorized into three groups: i.e. sequence-based, energy-based and machine learning-based groups. In the first group, the degree of sequence complementarity is considered as primary. This principle is used in e.g. miRanda [3] and TargetScanS [9]. miRanda first calculates sequence complementarity score with a weighting scheme; TargetScanS mainly takes the complementarity of the seed region, i.e. nucleotides 2-7, of miRNAs into account. Algorithms in the second group utilize thermodynamics as the main criterion. RNAhybrid [19] belongs to this group. It predicts the hybridization sites that are energetically most favourable as the binding sites. In the third group, algorithms such as NBmiRTar [26], miTarget [8] and *Zhang et al.* [27] collect different types of features and utilize machine learning techniques to find the feature patterns shared by true miRNA-target interactions.

Recently, the number of miRNA target prediction algorithms has been significantly increased. They do facilitate target identification, however, so far none of them could capture all true targets. Moreover, these computational approaches differ in algorithmic style; i.e., they use various features or factors are weighted differently. The lack of systematic verification and justification on the algorithms leaves the prediction accuracy and consistency unclear. To that end, generating a common criterion and test sets to analyze their

prediction performance and then integrating these algorithms to improve prediction accuracy will be very beneficial.

Lin et al. [11] mentioned that data integration can, in general, be approached from two routes; the "low-level" which deals with multi-factorial raw data directly and the "high-level" which combines multiple same type results from different studies.

In this study, we evaluate the performance of different target prediction algorithms and use integration methods to improve prediction accuracy. Both high-level integration approaches, e.g. algorithm combinations and ranking aggregation and low-level integration approach, e.g. the application of Bayesian Network classification, are performed. All of the methods are tested against miRNA-target interactions that are experimentally supported and several compiled negative control data sets. Our methods revealed that the system performance measured by the product of sensitivity and specificity provides a good criterion for algorithm comparisons. Algorithms categorized in the same group have similar prediction patterns. Algorithms categorized in different groups demonstrate their own advantages on different data sets. We inspected on the characteristics of miRNA-target site interactions and discovered that miRNAs have binding preference at the end of their target. We utilized three different integration strategies and demonstrated that the Bayesian Network classification results in best prediction accuracy.

2 Materials and methods

2.1 Materials

In the past years, the number of validated miRNA targets has been increased. Tarbase is a comprehensive repository recording a collection of experimentally supported miRNA targets in animal species, plants and viruses [15, 20]. The latest version, Tarbase 5.0, extracts data from a total of 203 scientific papers resulting in 1333 experimentally supported miRNA target gene interactions. For each interaction, it also provides direct evidences, such as reporter gene assay, and/or indirect experiment evidences such as microarray.

In this study we focused on human miRNAs as, to date, these are the best studied and also a large number of experimentally validated targets is available. To that end, a collection of 1093 experimentally confirmed human miRNA-target interactions from Tarbase is down-

True set	False sets
True: miRNAs and 3'UTR of true target interactions (157)	False_ori: miRNAs and 3'UTR of false targets (28)
	Shuffled: miRNAs and shuffled 3'UTR of true targets (157)
	Coding: miRNAs and coding region of true targets (157)

Table 1: Data sets

loaded. Only the direct interactions, which have the strongest experimental evidence, have been selected.

True and False_ori sets. In further inspection of initial data, we found 5 interactions as ambiguous since they are reported as both true and false based on different forms of evidence. After removing redundant, ambiguous and sequence unavailable entries, finally, 157 and 28 miRNA-target gene interactions are kept to serve as true and false examples, respectively. We refer to the original false targets as false_ori set. A small number of false samples is insufficient for data mining algorithms, and therefore, we compiled two more false sets. All data sets are listed in Table 1.

Shuffled set. Using 3'UTR sequences of 157 true targets as templates, we randomly shuffled the order of the nucleotides in these sequences, i.e. the frequencies of the nucleotides A, C, G and U are the same as in the original true target sequences; the order of the nucleotides, however, is random. In our experiments, this data set is registered as the shuffled set. In the analysis stage, we shuffle the sequences 20 times resulting in 20 sets of the random strings which are analyzed individually and over which the averages are computed.

Coding set. It has been established that miRNAs tend to bind their targets at the 3' UTR region. Given this feature, coding sequences are not supposed to contain binding sites. Therefore they are potential false target sequences. For our experiments, we used the coding sequences of the 157 true targets as a negative set.

2.2 Definition of predicted miRNA-target interactions

Tarbase provides experimentally validated miRNA-target gene interactions. However, computational algorithms predict putative binding sites, also referred to as predicted miRNA-target site interaction where miRNAs and their target mRNAs interact. In order to connect

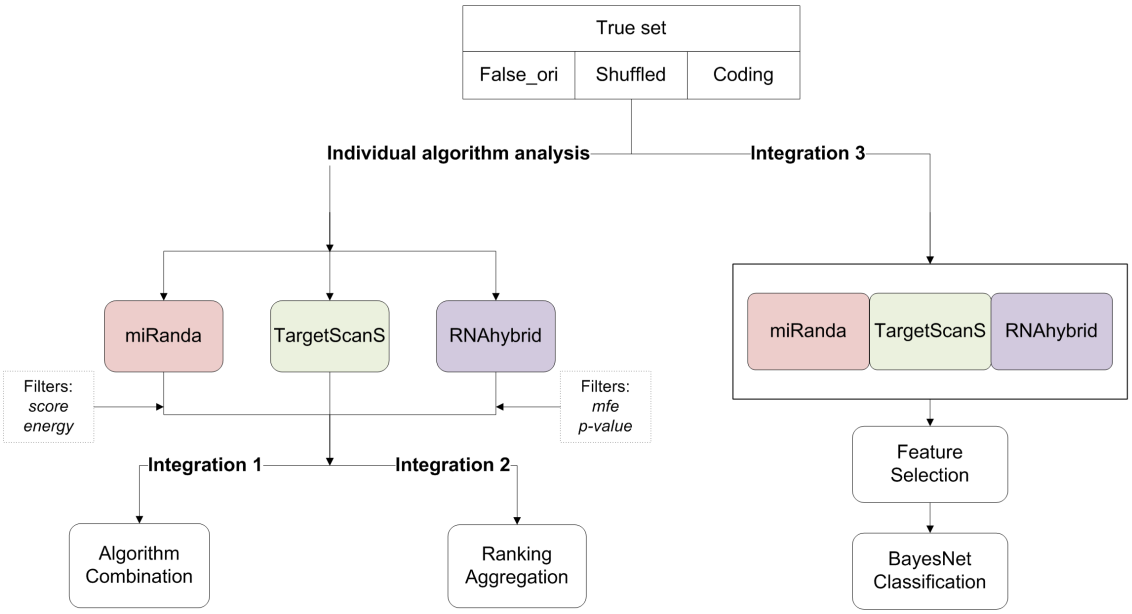


Figure 1: Schema. True dataset together with three false sets are processed through individual and integrated algorithm analysis.

experimental and computational results, we define predicted miRNA-target gene interactions as follows. A miRNA-target gene interaction is predicted only if there is at least one binding site where this miRNA interacts with at least one of the transcripts of this gene. The scale of the interactions increases in a sequence of miRNA-target site interactions, miRNA-target mRNA interactions and miRNA-target gene interactions.

2.3 Methods

In this study, we analyse the efficiency of three target prediction algorithms i.e. miRanda, TargetScanS and RNAhybrid and of a selection of integration strategies on these algorithms using multiple data sets. The motivation for choosing these three as the objects for comparison and integration is that they are the most frequently used target prediction algorithms and that they are open source which allow us to execute them locally and adapt them to different data sets and extract new self-defined features. In miRNA-target prediction, conservation is used but not always fully understood when applied over a multitude of distant species. Moreover, calculation of binding site conservation involves multiple sequence alignment over the multitude of species. This considerably contributes to computational load; in order to reduce this load, the conservation filter in each algorithm was

not applied in our experiments.

The computational procedure of our method is illustrated in Fig. 1. In the following paragraphs, we will briefly explain the components in the diagram followed by the experiment set-ups and how these algorithms are integrated.

MiRanda. miRanda [3] is one of the earliest developed large-scale target prediction algorithms for vertebrates. The standard version of miRanda selects target genes based on three properties: sequence complementarity using a position-weighed local alignment algorithm, free energies of RNA-RNA duplexes using the Vienna RNA fold package [25], and conservation of targets in related genomes. These features are weighed in a decreasing order. In this application, only the first two filtering layers, i.e. sequence and energy scores are applied to restrict the predictions.

TargetScanS. TargetScanS [9] is the new and simplified version of TargetScan [10] and it has a stronger emphasis on the seed region. In the standard version, the predicted target-sites require first a 6-nucleotide (nt) match to the seed region of miRNA, i.e., nucleotides 2-7; second, a binding site conservation in 5 genomes (human, mouse, rat, dog and chicken). Each binding site is associated with a site-type, which is either ”1a” or ”8mer” or ”m8”. In the application of local TargetScanS, only seed complementarity is required.

RNAhybrid. RNAhyrbid [19] finds the energetically most favourable hybridization sites between miRNAs and their target mRNAs using integrated powerful statistical models. It takes candidate target sequences and a set of miRNAs and looks for energetically favourable binding sites. In our practice, we first apply the RNAcalibrate tool to estimate distribution parameters, and then use the RNAhybrid tool to find the minimum free energy hybridization. The RNAeffective tool which calculates the effective numbers across species is not performed.

Ranking aggregation. Ranking aggregation is a strategy for optimization problems. In theory, it combines several individual ranked lists to produce a super list which will be as close as possible to all individual lists simultaneously. In our application, we use RankAggreg [17], an R package for weighted rank aggregation. It was illustrated by *Lin et al.* [11] that the utility of ranking aggregation leads to satisfactory simulation results when combining miRNA target lists from different algorithms. Further to these findings, we use

ranking aggregation as one integration option and test its performance. Our experimental set-up is using the tau distance function to measure distance and the Cross-Entropy Monte Carlo method [16] for aggregation.

Feature selection and Bayesian Network. Feature selection is the technique of selecting a subset of relevant features for building learning models. A Bayesian Network is a probabilistic model for classification. It is represented as a directed acyclic graph in which nodes represent attributes and edges represent conditional dependencies. The probability of any variable of a joint distribution can be calculated from conditional probabilities using the chain rules in probability theory [7]. This strategy is implemented in the Weka software environment [24]. CfsSubsetEval [6] and BayesNet [24] are applied for the purpose of feature subset selection and target classification. The error is estimated by 10-fold cross-validation.

2.3.1 Individual analysis

Running miRanda and RNAhybrid locally, one needs to decide cut-offs for several features. This is not the case for TargetScanS. The default settings of miRanda and RNAhybrid are to ensure detecting targets as much as possible. They also lead to many false positive predictions. In this case, we tune the parameters to find the best trade-off of true positive and false positive predictions. It is known that miRanda associates each predicted target site with a score which represents sequence complementarity degree between miRNA and its target as well as a free energy which measures the thermodynamics of the duplex. RNAhybrid predicts the targets with a minimum free energy (*mfe*) value and a *p*-value which represents the binding significance. The optimum cut-offs are those which achieve the performance (PERF) with highest combination of sensitivity (SENS) and specificity (SPEC), as defined by equations:

$$SENS = \frac{TP}{TP + FN} \tag{1}$$

$$SPEC = 1 - \frac{FP}{FP + TN} \tag{2}$$

$$PERF = SENS \times SPEC \tag{3}$$

where TP, FN, TN and FP represent true positive, false negative, true negative and false positive respectively. Sensitivity is also referred as to the true positive rate (TPR) which is defined as the ratio of experimentally supported miRNA-target gene interactions predicted by an algorithm. Specificity is equal to 1- false positive rate (FPR) which is defined as the ratio of false miRNA-target gene interactions detected by an algorithm as being true. We define performance as the product of sensitivity and specificity as written in equation 3. This performance is used to optimize the parameters and serves as a common reference in comparing the different integration strategies.

For model comparison, several performance measures have been described. In machine learning, the area under ROC Curve (AUC) [4] is often applied. This number, however, does not give a clue for parameter optimization. Alternatively, accuracy (ACC) [22] or F1_score [23] could be used and give similar results to our performance measure as they are derived from sensitivity and specificity as well. Our motivation to use the performance as given in equation 3 is that it reflects the requirement to the system to achieve both a high sensitivity and specificity. The value of performance is therefore logical and intuitive. And the performance differences are amplified at high values of two variables when comparing to the linear calculations.

2.3.2 Integration

After analysis of individual algorithms, three integration strategies are performed. The first is combining three individual approaches using various unions, intersections and majority vote. The second integration method is ranking aggregation which combines several ordered predicted target lists to generate a super list. In our practice, targets are first ranked according to the major feature of each prediction algorithms which are sequence score in miRanda, site-type in TargetScanS and *mfe* in RNAhybrid. After that, three ranked lists are integrated to a final list via a Cross-Entropy Monte Carlo method. The third integration approach is the application of a Bayesian network classifier. In our approach, the Bayesian network classifier is applied to the features measured by the individual target prediction approaches in order to discriminate two classes of targets. For each miRNA-target gene interaction a maximum and a minimum value of feature sets are registered. These features are (1) from miRanda: complementarity score, free energy, length of 3' UTR, relative binding position, number of hits; (2) from TargetScanS: site-type, length

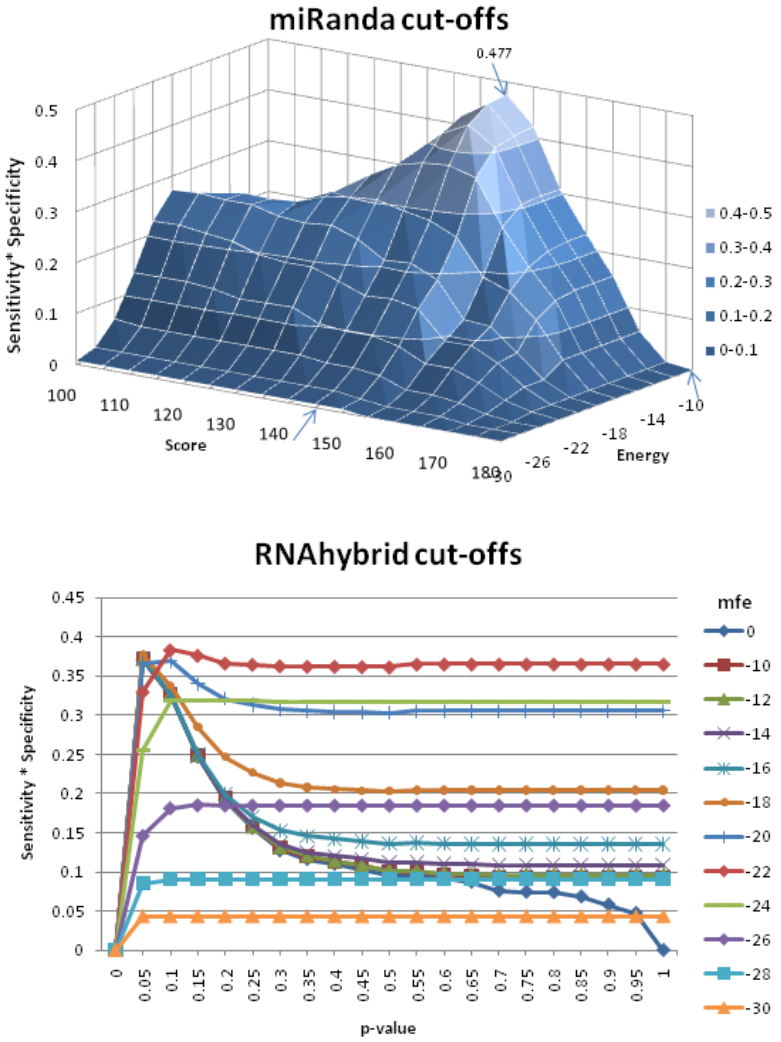


Figure 2: Filter optimizations for miRanda and RNAhyrbid. The combination of sequence score and free energy which achieves the best performance is set as thresholds for miRanda (left). The best combination of *mfe* and *p*-value is set as thresholds for RNAhybrid (right).

of 3' UTR, relative binding position, number of hits; (3) from RNAhybrid: *mfe*, *p*-value, length of 3' UTR, relative binding position. Considering both the maximum and the minimum values, there are 26 features in total. Subsequently, these features are then selected by feature selection and further classified by a Bayesian network.

	Cut-offs	*SENS*	*SPEC*			$\overline{PERF}$
		True	False_ori	Shuffled	Coding	
miRanda	score>=145 energy<=-10	0.662	0.429	0.72	0.822	0.414
TargetScanS	-	0.815	0.286	0.66	0.79	0.438
RNAhybrid	*mfe*<=-22 *p*-value<=0.1	0.578	0.454	0.654	0.629	0.313

Table 2: Performance of individual algorithms. The last column shows the average performance over the different sets. In order to assure that the results can be compared to that in Table 3,4,5, the shuffled set is listed but not used in the calculation of the average performance in this table.

3 Results

Before comparing prediction accuracy, feature cut-offs of miRanda and RNAhybrid are optimized. In order to achieve this, miRanda sequence complementarity score is tested from 100 to 180 with step=5 and energy is set from -10 kcal/mol to -30 kcal/mol with step=-2 kcal/mol; In RNAhybrid, minimum free energy is tuned from -10 kcal/mol to -30 kcal/mol with step=-2 kcal/mol and the *p*-value is tuned from 0 to 1 with step=0.05. Fig. 2 shows the optimization results for miRanda and RNAhybrid tested on true and shuffled sets. On the left, it is a landscape plot of sequence score, energy and system performance for miRanda. As can be seen, score=145 and energy=-10 kcal/mol lead to best performance represented by sensitivity * specificity =0.477. On the right, the graph shows the relationships between *mfe*, *p*-value and system performance. Free energy is represented by each line. X-axis shows the *p*-value changing from 0 to 1. Performance is depicted in y-axis. Further inspecting the graph, we found that when *mfe*=-22 kcal/mol, the system has very good performance overall. The system reaches the highest performance when *mfe*=-22 kal/mol and *p*-value=0.1.

3.1 Individual performance

After feature optimization, a peak performance for miRanda and RNAhybrid could be accomplished. Subsequently, three algorithms are compared. The average performance for each of the individual is summarized in Table 2. We found that TargetScanS has the high-

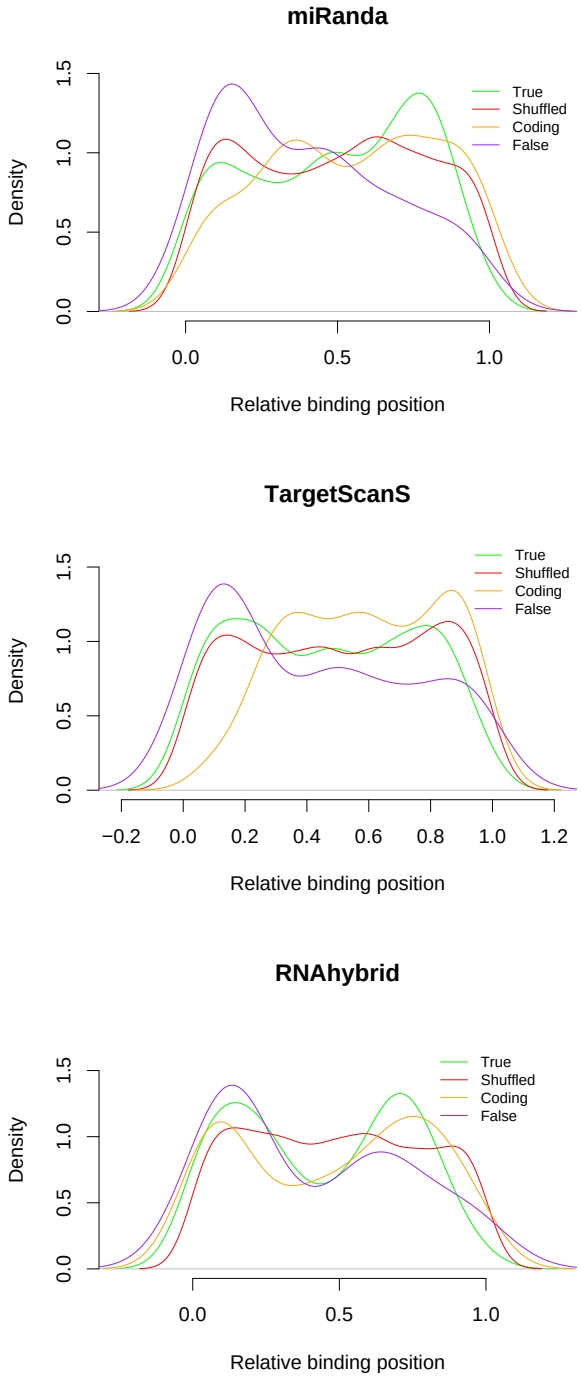


Figure 3: Characteristics of relative binding position in miRanda (left), TargetScanS (middle) and RNAhybrid (right). The density plot of relative target-binding position of true, false_ori, shuffled and coding sets are depicted in green, purple, red and orange respectively.

est sensitivity; miRanda has the highest specificity when testing on shuffled and coding sequences; RNAhyrbid has the highest specificity when testing on validated false target set. Moreover, we observed that miRanda and TargetScanS have similar patterns on different data sets. The specificity on coding set drop around 10% and 40-50% when comparing to that of shuffled and false_ori set, respectively. RNAhybrid, however, did not follow this pattern. A possible reason for this is that miRanda and TargetScanS are sequence-based algorithms which respond similarly on different types of sequences; whereas RNAhybrid is energy-based. In general, all three exhibit either a relative low specificity or/and sensitivity indicating that their prediction accuracy cannot yet be considered satisfactory.

In addition, we found one interesting target-binding site feature that is consistently displayed in three methods. Fig. 3 shows the distribution of relative binding position of each data set predicted by each method. The densities are estimated with a Gaussian kernel using R stats package [18]. Relative position is calculated as the position of a binding site divided by the length of target sequence. In the true data set, the predicted target sites in miRanda have location bias at the end; they have slightly a higher density at the two ends of the sequences in TargetScanS. This two ends binding preference is more obvious from RNAhybrid. In contrast, the target binding sites of false_ori set appeared more often at the beginning. While, the shuffled set shows nearly the uniform distribution. In summary, we conclude that the potential true target sites are enriched at the end of the binding sequences. The reason is probably that the binding sites are close to polyA tails which are the known factor effecting translation efficiency [5, 14].

3.2 Integration 1: combination

Our first strategy for integration is combining individual algorithms through various unions and intersections. The average performance of each combination over different sets is displayed in Table 3. It can be seen that majority vote is the best combination strategy, since it has the highest prediction performance. It is also higher than that of each individual algorithm. In the intersection part, we observed that targets predicted by miRanda and TargetScanS have higher degree of overlap than the other intersections. This is also because both of them weigh sequence complementarity as a main factor in their algorithms. We also suggest that, for the study of finding the networks involved by all the targets of a miRNA, using the union of these three is an option. This solution will cover a high

	SENS	*SPEC*		$\overline{PERF}$
	True	False_ori	Coding	
Unions				
miRanda, TargetScanS, RNAhybrid	0.879	0.179	0.573	0.331
Intersections				
miRanda, TargetScanS, RNAhybrid	0.452	0.607	0.879	0.336
miRanda, TargetScanS	0.643	0.464	0.866	0.428
miRanda, RNAhybrid	0.471	0.571	0.841	0.333
TargetScanS, RNAhybrid	0.516	0.536	0.841	0.335
Majority Vote				
miRanda, TargetScanS, RNAhybrid	0.79	0.357	0.79	0.453

Table 3: Performance of various unions and intersections of the individual algorithms.

range of true targets, as a trade-off, it will also cover a large number of false targets. This high false prediction rate can be reduced by further functional annotation analysis, e.g. targets can be further screened according to annotations with pathway, disease and gene ontologies.

3.3 Integration 2: ranking aggregation

Three ranked target lists from miRanda, TargetScanS and RNAhybrid are generated by sorting sequence score, binding site-type and energy respectively. For the miRNA-target gene interaction with multiple binding sites, the best values are selected to represent the whole interaction, i.e. highest sequence score, stringent binding site and lowest minimum free energy. After that, three lists are integrated to one via the RankAggreg package. The symbolic results are displayed in Table 4. Ranking for top to the end are displayed in a

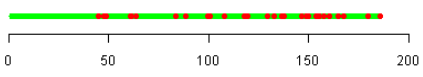
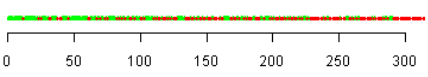
True vs False_ori	True vs Coding	$\overline{PERF}$
		
SENS=0.687 *SPEC*=0.687	*SENS*=1 *SPEC*=0	0.237

Table 4: Performance of ranking aggregation and symbolic plots of ranked lists. In the plot, true and negative targets are displayed in green and red respectively. Axis shows the ranking index.

Features				False_ori		Coding		BayesNet
RNAhybrid	miRanda	TargetScanS		*SENS*	*SPEC*	*SENS*	*SPEC*	$\overline{PERF}$
0	0	1	Feature Selection	1	0	0.815	0.79	0.322
0	1	0		1	0	0.809	0.611	0.247
1	0	0	&	0.917	0.643	0.828	0.49	0.498
0	1	1	BayesNet Classification	1	0	0.815	0.822	0.335
1	0	1		0.987	0.607	0.815	0.809	0.629
1	1	0	$\longrightarrow$	0.987	0.607	0.834	0.739	0.608
1	1	1		0.987	0.607	0.815	0.815	0.632

Table 5: Performance of Bayesian Network classification on different features. In the Features column, 1 represents that the features from this algorithm are selected for the machine learning.

direction from left to right. Green represents true targets; red represents false targets of coding (on the right) and false_ori sets (on the left) respectively. The average performance value is 0.237 indicating that ranking aggregation cannot precisely detect the true targets. A conceivable explanation is that the majority of true miRNA target does not always have very high sequence complementarity or has low free energy scores. Therefore, they are not always found in the top ranking list when using the key factor exclusively as the ranking criterion.

3.4 Integration 3: Bayesian Network classification

Feature sets are first processed through a feature selection procedure and then classified by a Bayesian Network. Their average performances are listed in Table 5. It shows that discriminating true and false targets based on the features from all different algorithms achieves the best performance. Furthermore, the classification performance on the features from RNAhybrid together with either miRanda or TargetScanS is also relatively high. This indicates that features from miRanda and TargetScanS are highly correlated and therefore could be redundant. Evaluation using this machine learning approach puts RNAhybrid as the best algorithm of the three individuals. As a comparison to the integration method 1 and 2, the Bayesian Network classification method based on the features from all three algorithms results the best overall performance and therefore can be con-

sidered as an optimal integration strategy.

4 Conclusions and Discussion

The increasing interest in miRNA regulatory function triggered the development of many computational approaches for miRNA target prediction. However, the large amount of approaches and the low degree of prediction overlap between them might leave the users confused. In this study, we demonstrated that the performance of current target prediction algorithms is by no means perfect. However, a proper integration of these prediction algorithms can significantly improve the prediction accuracy.

One of our contributions to the study of miRNA is to measure the performance of miRNA target prediction algorithms using both the true-positive and false-positive rate. Measuring target prediction performance has been recently addressed in few literature reviews. Most of these reviews compared target prediction approaches either from algorithmic point of view [1, 13], or using the estimated false positive rates [12] or using small numbers of experimentally validated miRNA targets [21]. However, using only false positive or true positive rates is not sufficient to indicate the prediction performance.

In our method, we generated the negative sets in a different way compared to previous studies. Current research focuses on finding the true targets, and consequently, only a small number of false targets are identified as by-products. This complicates calculation of false positive rates. The most common way to generate negative data set is sequence-shuffling. Besides that, we also used coding sequences as potential negative classes since most binding sites are not located in this region. Interestingly, error rates approximated on different type of negative sets have similar patterns. False positive rates on coding set are smaller than those on random set in general; while false positive rates on 3' UTR of real false target set are larger than those on random set. This indicates that all three prediction algorithms predict relatively more binding sites at 3'UTR.

The challenge of integration is to combine available data in a proper and efficient manner. In this chapter, we present three ways to integrate miRNA prediction algorithms. Algorithm combination and ranking aggregation are high-level integration methods, and application of a Bayesian Network classifier to the features measured by multiple prediction methods is a novel low-level integration method. Testing on common data sets, in-

tegration through Bayesian Network significantly improves prediction performance. This proves that, although high-level integration methods are easy and direct to apply, they lose information as not all data is passed to the integration stages. Moreover, low-level analysis which models raw data from different sources is complicated but, on the other hand, higher accuracy can be achieved. We also chose the proper classifier. *Yousef et al.* used Naive Bayes to classify targets [26]. The Naive Bayes classifier is based on the assumption of strong independence between the features. In our case, we found the Bayesian Network classifier did outperform the Naive Bayes since some of the features are not independent. In the future, for the functionally unknown miRNAs, of which the targets are unclear, we suggest the application of Bayesian Network classifier for target prediction.

From our computational analysis, we discovered one significant feature of miRNA target interaction. We observed that miRNAs have potential binding preference at the end of the target sequences. In paper [5], it is claimed that miRNAs have location bias at the beginning and at the end of 3' UTR. We found similar patterns. However, after further inspection of these patterns, we also observed many false targets at the beginning of 3' UTR.

Although Tarbase is a valuable resource for machine learning algorithms, the number of validated true targets and especially validated false targets is too small. We expect that with more validated targets available, the prediction accuracy of our proposed integration methods using Bayesian Network classification will increase. More improvement can be achieved by including more relative features such as binding site conservation. One of our further research will direct towards categorization of miRNA-target interaction to subtypes: once the target is validated, it is interesting to understand and establish whether it is the target for degradation or translation inhibition.

Acknowledgements

We thank Amalia Kallergi, Joris Slob and Kuan Yan for their critical discussions on methodology and results. This research has been partially supported by the BioRange program of the Netherlands Bioinformatics Centre (NBIC, BSIK grant).

References

[1] Christian Barbato, Ivan Arisi, Marcos Frizzo, Rossella Brandi, Letizia Da Sacco, and Andrea Masotti. Computational challenges in mirna target predictions: to be or not to be a true target? *Journal of biomedicine & biotechnology*, 2009.

[2] David P. Bartel. Micrornas: Genomics, biogenesis, mechanism, and function. *Cell*, 116(2):281–297, Jan 2004. http://dx.doi.org/10.1016/S0092-8674(04)00045-5.

[3] A. J. Enright, B. John, U. Gaul, T. Tuschl, C. Sander, and D. S. Marks. Microrna targets in drosophila. *Genome Biol*, 5(1), 2003. 1465-6914.

[4] Tom Fawcett. An introduction to roc analysis. *Pattern Recogn. Lett.*, 27(8):861–874, June 2006.

[5] Dimos Gaidatzis, Erik van Nimwegen, Jean Hausser, and Mihaela Zavolan. Inference of mirna targets using evolutionary conservation and pathway analysis. *BMC bioinformatics*, 8:69, Mar 2007.

[6] M. Hall. *Correlation-based Feature Selection for Machine Learning*. PhD thesis, 1998.

[7] D. Heckerman, D. Geiger, and D. M. Chickering. Learning bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20(3):197–243, 1995.

[8] Sung Kyu Kim, Jin Wu Nam, Je Keun Rhee, Wha Jin Lee, and Byoung Tak Zhang. mitarget: microrna target-gene prediction using a support vector machine. *BMC Bioinformatics*, 7, Oct 2006. 1471-2105.

[9] Benjamin Lewis, Christopher Burge, and David Bartel. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microrna targets. *Cell*, 120(1):15–20, Jan 2005.

[10] Benjamin P. Lewis, I. Hung Shih, Matthew W. Jones-Rhoades, David P. Bartel, and Christopher B. Burge. Prediction of mammalian microrna targets. *Cell*, 115(7):787–798, Dec 2003. http://dx.doi.org/10.1016/S0092-8674(03)01018-3.

[11] Shili Lin and Jie Ding. Integration of ranked lists via cross entropy monte carlo with applications to mrna and microrna studies. *Biometrics*, May 2009.

[12] G. Martin. Prediction and validation of microrna targets in animal genomes. *J Biosci*, 32(6):1049–1052, Oct 2007.

[13] P. Maziere and A. J. Enright. Prediction of microrna targets. *Drug Discov Today*, 12(11-12):452–458, Jun 2007. 1359-6446.

[14] B. Mazumder, V. Seshadri, and P. L. Fox. Translational control by the 3'-utr: the ends specify the means. *Trends in biochemical sciences*, 28(2):91–98, Feb 2003.

[15] Giorgos L. Papadopoulos, Martin Reczko, Victor A. Simossis, Praveen Sethupathy, and Artemis G. Hatzigeorgiou. The database of experimentally supported targets:

a functional update of tarbase. *Nucleic Acids Research*, 37(Database issue):D155–D158, 2008. http://dx.doi.org/10.1093/nar/gkn809.

[16] V. Pihur and S. Datta. Weighted rank aggregation of cluster validation measures: a monte carlo cross-entropy approach. *Bioinformatics*, 23(13):1607–1615, Jul 2007.

[17] Vasyl Pihur, Susmita Datta, and Somnath Datta. Rankaggreg, an r package for weighted rank aggregation. *BMC Bioinformatics*, 10(1):62, 2009.

[18] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. ISBN 3-900051-07-0.

[19] M. Rehmsmeier, P. Steffen, M. Hochsmann, and R. Giegerich. Fast and effective prediction of microrna/target duplexes. *RNA*, 10(10):1507–1517, 2004. 1355-8382.

[20] Praveen Sethupathy, Benoit Corda, and Artemis G. Hatzigeorgiou. Tarbase: A comprehensive database of experimentally supported animal microrna targets. *RNA*, 12(2):192–197, Feb 2006. http://dx.doi.org/10.1261/rna.2239606.

[21] Praveen Sethupathy, Molly Megraw, and Artemis G. Hatzigeorgiou. A guide through present computational approaches for the identification of mammalian microrna targets. *Nature Methods*, 3(11):881–886, 2006. 1548-7091.

[22] Wikipedia. Accuracy and precision— Wikipedia, the free encyclopedia. [Online].

[23] Wikipedia. F1 score — Wikipedia, the free encyclopedia. [Online].

[24] Ian H. Witten and Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, 1999.

[25] S. Wuchty, Fontana W., Hofacker I. L., and Schuster P. Complete suboptimal folding of rna and the stability of secondary structures. *Biopolymers*, 49:145–165, 1999.

[26] Malik Yousef, Segun Jung, Andrew V. Kossenkov, Louise C. Showe, and Michael K. Showe. Naive bayes for microrna target predictions machine learning for microrna targets. *Bioinformatics*, Oct 2007.

[27] Yanju Zhang, Jeroen S. de Bruin, and Fons J. Verbeek. mirna target prediction through mining of mirna relationships. *BioInformatics and BioEngineering*, pages 1–6, Jul 2008.

Chapter 5

Identification of Common Carp Innate Immune Genes with Whole-Genome Sequencing and RNA-Seq Data

Based on

Yanju Zhang, Elia Stupka, Christiaan V. Henkel, Hans J. Jansen, Herman P. Spaink and Fons J. Verbeek. (2011). Identification of Common Carp Innate Immune Genes with Whole-Genome Sequencing and RNA-Seq Data. Journal of Integrative Bioinformatics, 8(2):169.

Summary

The common carp is a candidate model system for immunology research. Using next-generation sequencing technology, we have generated a huge amount of sequence reads from the carp genome and transcriptome. Currently, our aim is to identify carp genes, particularly genes involved in the development of the innate immune response, given the circumstance that the carp genome assembly is not yet completed. To achieve this, we developed a comprehensive genome annotation pipeline. This analysis allowed us to estimate that the carp has approximately 39 TIR domain-containing transcript isoforms and genes.

1 Introduction

Common carp (*Cyprinus carpio*) is one of the most important freshwater cultured fish species that has been widely used in fish biology research [3]. A single female is capable of producing up to a few hundred thousand eggs that can be efficiently fertilized *in vitro*. Since the innate immune response is already active in developing embryos, common carp can be a relevant model for studying its mechanisms. The innate immune response is the first line of defence against infectious disease and cancer by identifying and killing pathogens and detrimental cells. This innate immune response relies heavily on signaling by pattern recognition receptors. The best-studied pattern recognition receptors of the vertebrate innate immune system are the Toll-like receptors (TLRs). All the TLRs, some Interleukin receptors (IL-Rs) and downstream adaptor proteins contain a Toll/Interleukin-1 receptors (TIR) domain which is a highly conserved functional unit mediating the protein-protein interactions between the receptors and the adaptors thus relaying the signal.

TIR domain-containing genes therefore play important roles in immunity signalling pathways. In zebrafish (*Danio rerio*), this family has been studied using microarray technology [23]. However, microarrays have a number of shortcomings, i.e. low sensitivity and specificity, low consistency across platforms, and, above all, they rely on a fixed definition of the transcriptome for their design.

Next-generation sequencing (NGS) is a recently developed, high-throughput sequencing technology, with which one can produce millions of sequence reads in a few days at a low cost and without the need for a priori knowledge of the sequences [19]. Applying such technology to the entire genome of a particular organism is referred to as whole-genome sequencing. Another application, RNA-Seq, is to sequence cDNA for transcriptome profiling. In comparison to microarrays, RNA-Seq has a much higher dynamic range, base-level resolution, richer splicing information and the ability to detect previously unknown transcripts.

Our ultimate goal is to study how the transcriptome, especially the expression of the innate immune response genes, changes upon pathogen infection using NGS. Since common carp is not a model system and no reference genome assembly is available, both the whole carp genome and several transcriptomes are sequenced. Currently, as a pilot study, we focus on discovering the innate immune response genes, especially TIR domain-containing genes.

In this Chapter, we present a gene identification strategy that integrates whole genome sequencing data, RNA-Seq data and relevant data obtained from public databases in order to identify TIR domain-containing genes and transcripts in carp. With limited data available, different data sources and methods are compared and integrated in order to maximize the likelihood of detecting the target sequences.

2 Background

2.1 Common carp

The common carp is a serious candidate model system for very high throughput screens of pharmaceutical compound libraries. The closely related cyprinid zebrafish has been developed into a medium throughput capacity model for drugs related to cancer and immune-related diseases. The advantage of carp is that a single female is capable of producing up to a few hundred thousand eggs that can be efficiently fertilized in vitro. Therefore, the genomic homogeneity of carp eggs is easier to control than for zebrafish that provide smaller clutches of 150 to 200 eggs. Thus, large clutches of embryos derived from a small group of common carps enable hundreds of thousands of pharmaceutical drug candidates to be tested with less genetic diversity in the screening model.

2.2 TLR pathway and TIR domain containing genes

The innate immune system is the first line of defence that protects the host against infectious disease and cancer [23]. This system relies heavily on pattern recognition receptors mediating immune responses to pathogenic microorganisms. The Toll-like receptors (TLRs) and interleukin-1 receptors (IL-1Rs) are probably the most essential pattern recognition receptors of the vertebrate immune system.

Stimulation of TLRs by their ligands leads to the recruitment of adaptor proteins to the receptors. Differential utilization of the adaptor molecules by the TLRs causes specific activation of a range of transcription factors such as NF-kB, activator protein 1 (AP-1), and IFN regulatory factors (IRF) 3, 5, and 7 through distinct signalling pathways, eventually leading to the downstream activation of proinflammatory cytokines [13]. The details of the TLR signalling pathway are depicted in Fig. 1.

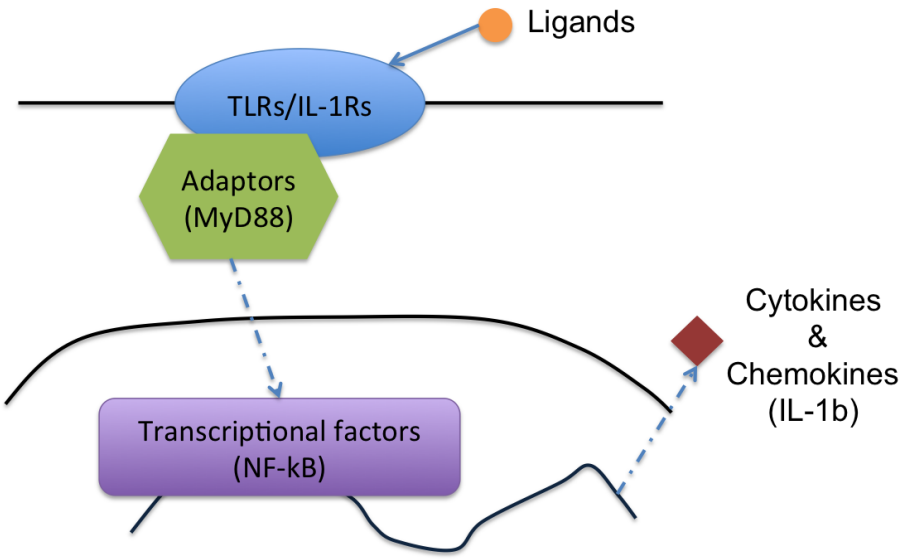


Figure 1: TLR signalling pathway.

Upon binding to Interleukin-1 ligand, IL-1R interacts with IL-1R accessory proteins. With the aid of adaptor proteins, this signal is transduced and leads to the activation of NF- kB. The whole process is very similar to the TLR pathway.

The Toll/Interleukin-1 receptor (TIR) domain is the conserved intracellular signalling domain shared by TLR and IL-1R families. It is also found in some of the adaptors, e.g. MyD88, which connect the Toll-like or Interleukin receptors with downstream transcriptional factors. The TIR-domain containing proteins in the human genome comprise ten members of the TLR family, eight members of the IL-1/IL-18 receptor group (IL-1R/IL-18R) and five adaptor proteins. Zebrafish has one or more counterparts for the human TLR1, TLR2, TLR3, TLR4, TLR5, TLR7, TLR8, TLR9, IL-1R and IL-18R genes and one copy of the adaptor genes MyD88, MAL, TRIF and SARM [13].

2.3 Next generation sequencing

With the advances of second generation sequencing technologies, including Solexa/Illumina Genome Analyzer, ABI SOLiD, Roche/454 and Helicos, massive genome sequencing projects ranging from prokaryote to eukaryote have been carried out. These technologies are now widely applied in advanced research such as genome sequencing, transcriptome sequencing, exome sequencing, microRNA expression profiling and DNA methylation studies [10]. Unlike traditional the Sanger sequencing technology [18], these new

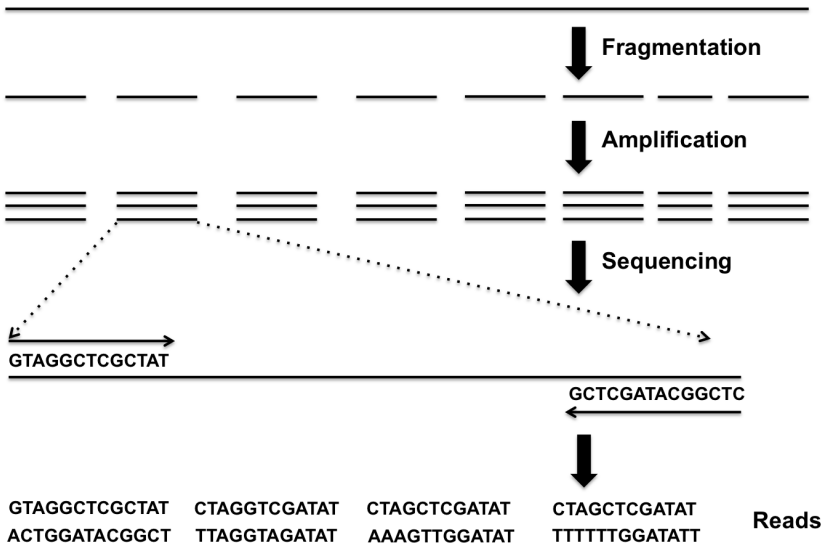


Figure 2: Using next-generation sequencing technology to sequence genome (the whole genome sequencing).

sequencing technologies produce shorter reads with higher genome coverage at very low cost in, and moreover, a short period of time [15]. In the mean time, some of the advantages, i.e. fast and high-throughput data generation, also lead to computational challenges in the genome assembly as well as the subsequent genome annotation analysis.

2.3.1 Whole genome sequencing and RNA Sequencing (RNA-Seq)

Completed genome sequences are immensely useful in biological researches. A systematic transcriptome analysis reveals the dynamic activities in the cell. Currently, the whole genome sequencing and RNA-Seq are the promising technologies which apply next generation sequencing to sequence the whole genome and transcriptome. Fig. 2 illustrates the principle of the whole genome sequencing technology. First DNA molecules are extracted and then sheared into short fragments. Later on adaptors are attached to one or both ends. With or without amplification, each fragment is then sequenced by the sequencer to obtain short sequences from one end or both ends resulting single-end or paired-end reads respectively. In principle, RNA-Seq technology is similar to whole genome sequencing. The main difference is at the sample preparation stage. Instead of extracting DNA, RNA-Seq extracts RNAs which are characterized with polyA tails and converts them to cDNAs. The output of next generation sequencing experiments are short reads which are typically 30 to 400 bp depending on the sequencing technology.

2.3.2 Assembly

Genome assembly is the process that constructs the original continuous DNA sequences from millions of short DNA reads. It can be accomplished from two directions: a *de novo* approach which constructs genomes from scratch and comparative approach that uses a closely related organism as a guide to produce contiguous genome sequences [15]. Currently, the genomes of several species, e.g microbes, yeast, worm, fruitfly and human have been completed sequenced and assembled.

A *de novo* genome assembly approach is to reconstruct genomes without using any previously sequenced organisms. This method is capable of capturing the sequence diversity for each individual, however, it is extremely hard, complicated and memory consuming. The main computational strategies applied in *de novo* assembly are overlap computation and Eulerian path. Overlap computation calculates all pair-wise alignments between a set of reads and constructs the contigs according to the read overlap. It includes the greedy algorithm based tools such as TIGR assembler [24] and overlap-layout-consensus (OLC) based assemblers such as Arachne [1]. Eulerian path utilizes graph theory and the de Bruijn graph [14] based data structures for assembling a genome. Most recently developed second-generation assemblers applied the Eulerian strategy such as SOAPdenovo [11], Velvet [27], ABySS [20] and Allpaths [4].

The principle of comparative assembly is different. First, the reads are aligned to a highly related genome called reference genome. This alignment is then used directly to compute the consensus sequence of the new genome. In practice, comparative assembly is easier since it utilizes alignment instead of expensive overlapping step. An example in this category is AMOScmp [16].

2.4 Gene annotation

Genome annotation is an important downstream analysis after an organisms genome has been completely sequenced and assembled. The main goal of genome annotation is to add the biological context to the sequence, to identify the key features of the genome in particular genes and their products. One of the analyses is gene finding that predicts the precise start and stop position of a gene and the splicing pattern of its exons [22]. Similar to genome assembly, this analysis can be approached using *ab initio* gene prediction and comparative prediction.

In general, the *ab initio* gene prediction algorithms, e.g. Augustus [21], utilize the existing gene structures as training set and then build up statistical models to predict the gene structures for the functionally unknown sequences. In contrast, comparative gene prediction induces a gene structure from close related species. A high similarity between a sequence in one species and a gene in another species is good evidence that this sequence contains a gene with similar structure. BLAST [12] and BLAT [9] are two fast sequence similarity search algorithms often applied in this area.

3 Methodology

The genome of a fully homozygous common carp, obtained in a single generation without inbreeding [3] and a heterozygous carp genome were sequenced using Illumina Genome Analyzer IIx sequencing technology. For the homozygous carp, we generated a paired-end sequencing library with insert sizes of about 200 base pairs (bp), from which we obtained approximately 40 Gbp of usable sequences with a read length of 76 bp. For the heterozygous carp genome, three DNA libraries were constructed: one single-end library with read length 51 bp; one paired-end library with 200 bp insert size and 51 bp read length; one mate-pair library with insert sizes of 5 Kbp and 36 bp read length. In total, we generated about 10 Gbp sequences for this strain.

We also sequenced the total mature messenger RNAs of common carp at different developing stages and conditions. Four mRNAs samples, wild-type carp and carp infected with the *Mycobacterium marinum* pathogen both at embryonic and adult stages, were extracted and then sequenced using the Illumina Genome Analyzer IIx. For each sample, an RNA-Seq sequencing library was constructed from which single 51 bp reads were sequenced. Details of all the data sets are listed in Table 1.

3.1 Genome assembly strategy

In the absence of a carp reference genome, the first task is to generate a carp genome assembly. Considering the fact that the homozygous carp genome sequencing data is approximately 4 times abundance than that of the heterozygous carp genome and is about 13x genome coverage, we chose to assemble the homozygous carp genome and considered to use 5 Kb mate-pair library from the heterozygous carp only for the purpose of scaffolding.

Dataset 1: gDNA, heterozygous carp

	Library size (bp)	Lanes	Read length (bp)	Size (GB)
Single-end	-	2	51	0.6
Paired-end	200	10	51	5.1
Mate-pair	5 K	7	36	3.7

Dataset 2: gDNA, homozygous carp

	Library size	Lanes	Read length (bp)	Size (GB)
Paired-end	200 bp	6	76	40

Dataset 3: RNA-Seq

		Lanes	Read length (bp)	Size (GB)
Single-end	Embryo, wt	1	51	2.6
Single-end	Embryo, infected	1	51	2.6
Single-end	Adult, wt	1	51	2.7
Single-end	Adult, infected	1	51	3.1

Table 1: The genomic reads and RNA reads generated from homozygous carp using Illumina sequencing.

The strategy of the carp genome assembly is illustrated in Fig. 3. First all the raw reads are filtered by quality control criteria and further pre-assembled in the pre-processing stage. Having the high quality reads, three *de novo* assemblers are applied and compared in order to achieve the best assembly.

The raw reads generated from the Illumina sequencer included base-calling errors and adapter contamination. It has been found that Illumina sequencing quality decreases specially at the end of the read [17]. Adapter contamination is mainly caused by insert sizes smaller than the read length. Therefore the reads were filtered at a threshold if either Read 1 or Read 2 contained an adapter sequence longer than 6 bp and the low quality reads were eliminated.

We then merged the remaining paired-end reads into a longer single-end read if they had 7 overlapping nucleotides. Pairs were not collapsed into a longer read if repetitive sequences within them tended to create ambiguous connections. This preassembly procedure not only produces long reads which will potentially improve the efficiency and quality of the assembly, but also provides confirmation for the quality of the 3' end of the reads. After the pre-processing, 3.5% of nucleotides are discarded and 69.9% of pairs are merged.

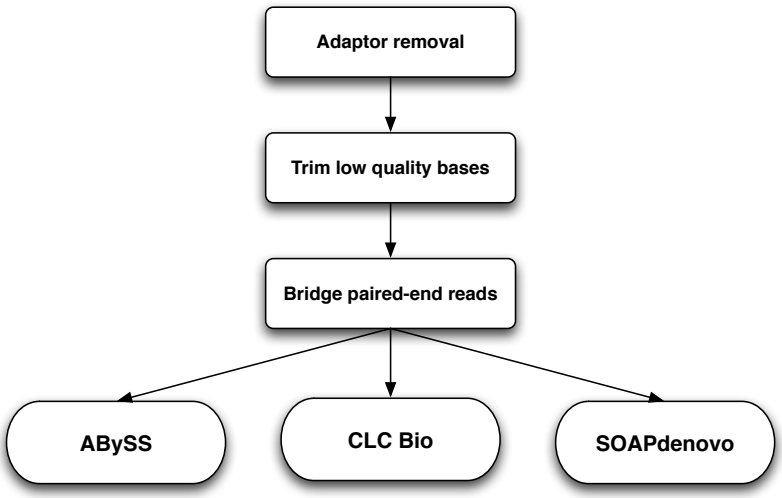


Figure 3: Genome assembly pipeline

Next, we assembled the high quality and merged genomic DNA reads using three *de novo* assemblers: ABySS, CLCBio and SOAPdenovo. All of them are de Bruijn graph-based assemblers. A de Bruijn graph is a directed graph representing overlaps between sequences of symbols [14]. In the graph, basically, short reads are broken into smaller sequences of DNA, called *k*-mers. Each node represents a *k*-mer nucleotides; an edge exists only if the adjacent nodes are overlapped by *k-1* nucleotides. Extracted contiguous sequences are represented by unambiguous paths through the nodes.

Both ABySS and SOAPdenovo are free software; whereas CLCBio is a commercial product. Each of the three assemblers has its own innovations. ABySS improves the memory efficiency by using a distributed de Bruijn graph and has been successfully applied in assembling of an African male human genome [20]. SOAPdenovo performs error correction before the de Bruijn graph is built and specifies different libraries for assembly and scaffolding. It was successfully applied for giant panda genome assembly [10]. CLCBio has strength in handling repeats using the information from paired-end reads. Unlike ABySS and SOAPdenovo, which require manually optimization of the parameter *k*, CLCBio tunes and optimizes it automatically [2].

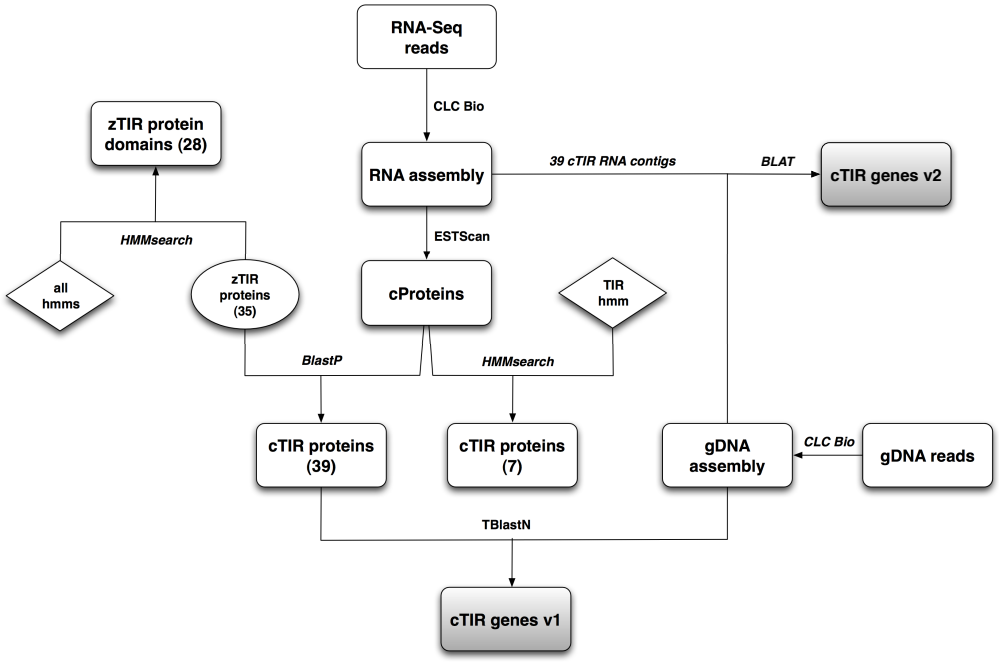


Figure 4: TIR containing gene annotation pipeline. The output TIR gene sequences (cTIR genes v1 and cTIR genes v2 depicted in shade) derived from different methods are compared and further integrated to construct the final TIR genes

3.2 Annotation strategy

Animal genome assemblies based on NGS data only are generally highly fragmented. We therefore developed an integrative strategy to maximize the probability of identification of carp genes; specifically TIR containing genes and transcripts.

Firstly all the RNA-Seq reads from all the samples were pooled and assembled using the CLCBio *de novo* assembler. The resulting RNA contigs were used as potential gene product fragments. These sequences were then translated to protein sequences using the ESTScan algorithm [8]. After that, the protein sequences obtained were searched for the TIR domain found in Interpro [7] using the HMMsearch algorithm [5].

The zebrafish is evolutionarily close to the common carp (both are cyprinids) and the zebrafish genome is relatively well covered and annotated in the Ensembl database [6]. Therefore we used this genome to facilitate the annotation of the carp TIR containing genes. In this manner, we performed a comparative genomic analysis using zebrafish resources: the zebrafish TIR containing proteins found in Ensembl were BLASTed against the carp peptides obtained from the RNA-Seq contigs resulting in the putative carp TIR

Assembly	k	n	n:N50	median	mean	N50	max	sum
Without preprocessing	25	1637271	250045	384	735	1409	17597	1.20E+09
Without preprocessing	27	1847118	-	-	680	1389	18684	1.26E+09
After preprocessing	**25**	**1086163**	**159656**	**587**	**1135**	**2260**	**26293**	**1.23E+09**

Table 2: Performance of CLCBio

proteins allowing us to identify potentially new carp TIR transcripts. Considering the fact that the obtained assembly is fragmented, the transcript and protein sequences are used to bridge fragmented DNA contigs. To achieve this, the candidate TIR transcript and protein sequences were further mapped to the carp genomic contigs by using TBlastN [12] and BLAT [9]. The DNA contig hits are connected using a number of 'N' as gap sequences and finally result in an alternative set of the carp TIR genes for comparison. The entire pipeline is shown in the Fig. 4.

4 Results

4.1 Carp genome assembly

We ran the CLCBio *de novo* assembler on both the raw genomic reads and the reads after preprocessing. Parameter k was optimized to 25, as we manually changed it to 27, the N50 value slightly dropped as showed in Table 2. N50 is defined as the length N for which 50% of all nucleotides in the contigs are in a length of at least N nucleotides long. It is a useful heuristic for measuring the quality of an assembly. A higher N50 represent a longer average contig length. As illustrated in the table, the N50 increased from 1409 to 2260 after the preprocessing step, a 60% improvement compared to the assembly derived from the raw data. This result shows that the preprocessing is a crucial step that makes a huge difference in the final assembly.

ABySS and SOAPdenovo were applied to the preprocessed data. Parameter k for k-mer is tuned from 17 to 55 in order to achieve the best performance. ABySS produced the longest N50 contig length (N50 = 716 bp) when the k-mer length is 50. As for SOAPdenovo, it restricts the value of k to an odd number. The best assembly with N50 of 729 bp was achieved when k is 45. The assemblies generated by three assemblers are summarized in Table 2, 3, 4.

Compared to ABySS and SOAPdenovo, the CLCBio assembler performs much better.

k	n	n:N50	median	mean	N50	max	sum
23	4.19E+07	1217339	150	181	187	5275	6.97E+08
25	3.23E+07	1124372	166	212	233	10026	8.48E+08
30	2.13E+07	866650	192	275	349	16613	1.04E+09
35	1.56E+07	681428	211	333	479	16601	1.14E+09
40	1.21E+07	566732	227	384	606	16601	1.21E+09
45	9812550	516612	240	421	694	16601	1.26E+09
50	**8151538**	**515595**	**249**	**433**	**716**	**16601**	**1.29E+09**
55	7015100	585470	237	403	653	16601	1.33E+09

Table 3: Performance of ABySS

k	n	n:N50	median	mean	N50	max	sum
17	1.40E+08	41018	110	116	112	338	1.07E+07
21	4.27E+07	1180069	149	179	184	4210	6.62E+08
25	2.81E+07	1007523	177	236	274	7225	9.17E+08
29	2.10E+07	830409	196	284	369	11122	1.05E+09
31	1.85E+07	760064	203	306	418	11124	1.10E+09
35	1.47E+07	647073	215	348	520	10342	1.18E+09
41	1.09E+07	537078	238	409	662	10342	1.24E+09
45	**9054965**	**500758**	**257**	**443**	**729**	**10342**	**1.27E+09**
51	7106660	524543	257	438	724	10141	1.31E+09
55	6201398	570679	280	438	676	8881	1.31E+09

Table 4: Performance of SOAPdenovo

Data	Assembler	n	n:N50	median	mean	N50	max	sum
RNAseq	ABy	1127477	67629	151	195	203	6720	4.66E+07
RNAseq	CLC	315316	76037	160	227	255	7939	7.16E+07
contigs+EST +mRNA	CLC	25157	6621	634	721	896	9919	1.81E+07

Table 5: Transcriptome assembly

The best assembly from CLCBio is the one with N50 of 2260 bp. This number is about two times higher than that from ABySS and SOAPdenovo. According to this, CLCBio is selected as the final assembler.

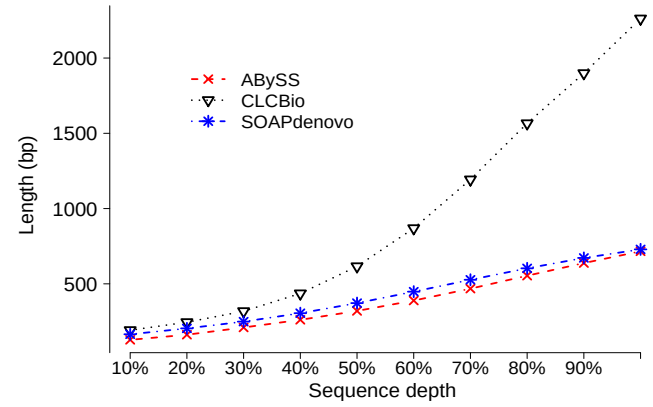
In order to determine whether the current genomic sequencing data is sufficient to generate a reasonably good assembly, we analysed the dependency of the number, length and total size of contigs with sequencing depth. Sequencing depth is represented by different subsets of the read data size. All three assemblers have similar trends for different properties. Fig. 5(a) shows that the average contig length is increasing with sequencing depth in all the assemblers. In Fig. 5(b), it shows how the number of contigs changes along with sequencing depth. The number of contigs first increases as most contigs only consist of one single read. As the number of reads increases, the chance that reads will overlap to form longer contigs increases. After a certain point, the increase of reads number leads to a decrease in the number of contigs. Fig. 5(c) illustrates that the size of the assembly almost reached the saturation status.

From the plots in Fig. 5, we conclude that current data is sufficient to cover the whole carp genome but the assembly is still fragmented. Given more data, CLCBio will generate a better assembly with fewer but longer contigs at higher speed when comparing to the rest.

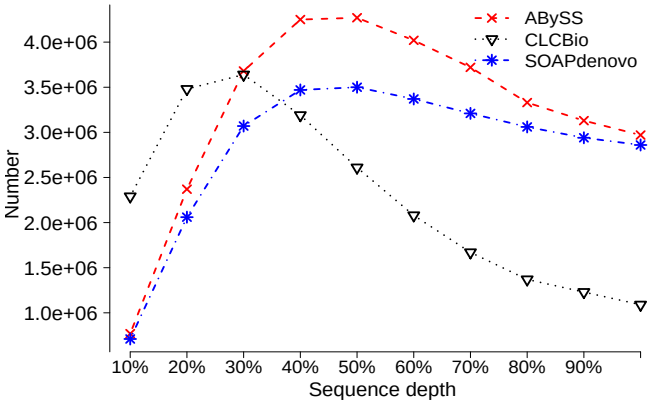
In order to identify expressed sequences, an assembly of the RNA-Seq data was performed. Using the CLCBio assembler, we were able to achieve RNA contigs from RNA-Seq data with a total size of 71.6 Mbp and an N50 of 255 bp. Integrating RNA contigs with the existing EST and mRNA sequences from GenBank, the RNA assembly reaches to an N50 contig length of 896 bp and 18.1 Mbp in total size. The details are showed in Table 5.

4.2 4.2 TIR containing genes and products

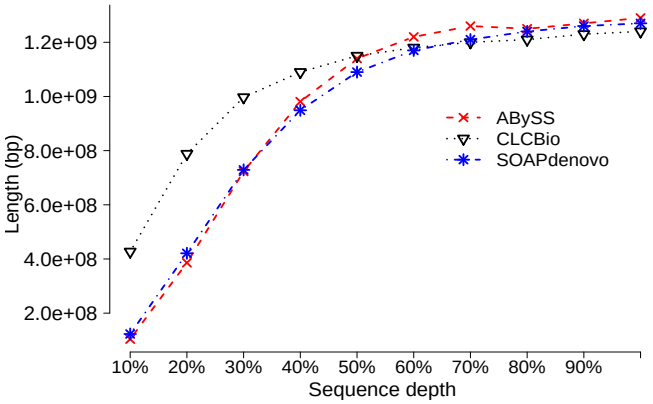
The search for zebrafish TIR genes allowed us to identify characteristics of the gene structures and protein domain structures which will help us to annotate the corresponding genes in the carp genome. We used HMMsearch for mining of the protein domains in 35 zebrafish TIR proteins obtained from Ensembl, and we found that this class of proteins contains 28 domains in total. We also performed a search for all TIR domain-containing genes in zebrafish and found that out of 33 zebrafish TIR genes 16 of them are single-exon genes, 15 genes contain multiple exons and two gene structures are missing from



(a) N50



(b) Number of contigs



(c) Assembly size

Figure 5: Comparison of ABySS and CLCBio and SOAPdenovo.

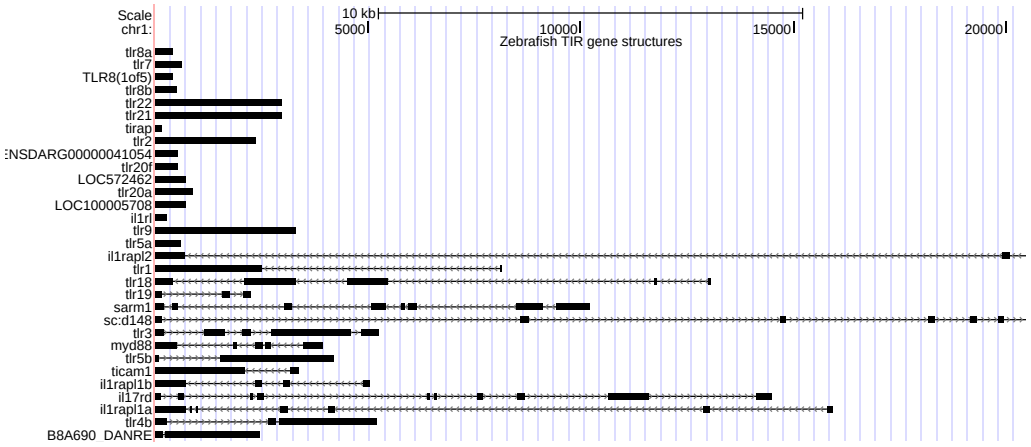


Figure 6: Gene structures of 31 zebrafish TIR domain-containing genes.

Ensembl. The structures of zebrafish TIR domain-containing genes are displayed in Fig. 6.

We were thus able to compare the carp RNA contigs to the zebrafish TIR containing proteins, as well as using their translated sequences to BLAST the carp genome assembly in order to identify potentially fragmented DNA contigs. By setting the cut-off E=1e-05, we discovered 39 carp TIR protein contigs similar to 34 zebrafish TIR RNA sequences as shown in the Supplementary Table 1. Both the protein and derived RNA sequences are further used to discover TIR genomic sequences.

Using protein or RNA contigs as references, we can create longer artificial DNA scaffolds with unknown gap sizes, which can hopefully contribute in obtaining a complete gene model. RNA sequences can be more diverse than DNA due to the alternative splicing, which will increase the chance of connecting the DNA contigs in the wrong order. In this case, we focus on detecting TIR domain-containing genes. In zebrafish, half of these genes contain only one exon which can not lead to alternative splicing. Moreover, there are 35 TIR containing proteins derived from 33 TIR genes. The ratio of gene to transcript is almost 1. Finally, if the alternative splicing events do exist, as long as they do not happen between the joint contigs, the right order of contigs can still be obtained. Therefore, we believe that using RNA contig or peptide sequences to scaffold TIR DNA contigs in carp is a good practical solution.

Using 39 TIR proteins and RNAs, we finally generated two versions of TIR gene sequences, i.e. cTIR genes v1 and cTIR genes v2 depicted in Fig. 4. In Fig. 7, an example is shown of the comparison of the protein sequence est_contig_6303 7(a) and RNA contig

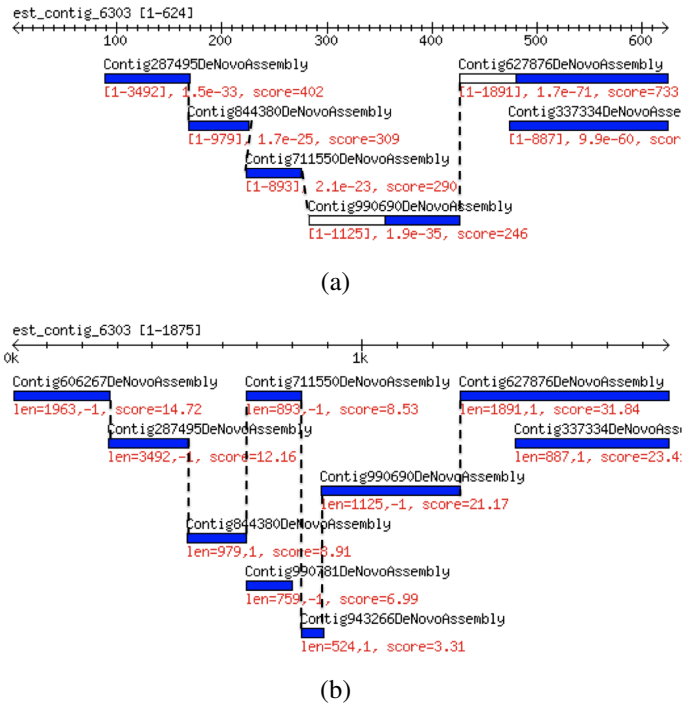


Figure 7: Using RNA and protein contigs to scaffold genomic contigs. Genomic contigs shown in blue are aligned to protein contig 6303 (a) and RNA contig 6303 (b) depicted by the first line in each panel. Hits in the same region are ordered by the mapped scores, with the best matches at the top. Dotted lines indicate that the contigs can be joined with unknown gap size.

est_contig_6303 7(b) mapping to the carp genome. In Fig. 7(a), DNA contigs (No.287495, 843380, 711550, 990690 and 627876) can be joined together as a scaffold; in Fig. 7(b), it shows not only the previous 5 DNA contigs but also DNA contig No.606267 and 943266 can be bridged. We also found that the connected DNA contigs are largely identical between the two versions of TIRs. Scaffolding genomic contigs by BLAT RNA contigs (version 2) against the genome performs better than using protein sequences (version 1), since more DNA contigs can be joined and the RNA sequences can be constructed from the DNA contigs without any missing sequences. In total, 162 genomic DNA contigs are scaffolded using 39 TIR transcript sequences.

5 Conclusions

We have generated a draft assembly for common carp with N50 of 2260 bp and genome size of 1.23 Gbp. Due to the fact that the assembly still contains many fragments, we

could not directly apply *ab initio* gene prediction method for gene discovery. Therefore, we developed an annotation pipeline which integrates whole genome sequencing, RNA-Seq data and available zebrafish data to detect the TIR containing genes in carp. We identified 39 TIR domain-containing transcripts. Using these transcripts as references, 162 DNA contigs are stitched to 39 DNA scaffolds. Potentially, the extended genome contigs will stand a high probability of containing the entire gene. Considering the facts that the ratio of known TIR genes and proteins in zebrafish is 33:35, and that half of these genes contains only a single exon (ruling out alternative splicing), a 1:1 ratio between TIR transcripts and genes in carp is a reasonable approximation. Therefore we established that there are around 39 TIR containing genes and transcripts in common carp.

Based on the performance of ABySS, SOAPdenovo and CLCBio, we decided to apply CLCBio as the final genomic sequence assembler since it produced relatively long contigs covering a similar amount of genome as the other two approaches. Without performing scaffolding, the initial carp genome assembly has reached an N50 of 2260 bp. The N50 contig length is longer than the initial assembly of giant panda genome of which the N50 contig length is 1483 bp [10]. In the giant panda project, the assembly was further improved to contig length N50 of 39886 bp by iterating scaffolding using extra 500 bp, 2 Kbp, 5 Kbp, and 10 Kbp libraries. Compared to this case, the limitation of our experimental design is that we currently did not have different sequence libraries for the haploid fish strain without which scaffolding or the finishing step is difficult to carry out. In the future, with more and larger size libraries available, the assembly will be further improved.

We found that when the data is limited, gene identification analysis is not as straightforward as the standard analysis which usually consists of gene assembly (if necessary) and *ab initio* gene prediction analysis or mapping RNA reads to the well-assembled genome to discover expressed sequences. In our study, we noticed that although the sequencing depth of genomic and transcriptomic data was not sufficient to produce the complete genome and transcriptome assemblies independently, these data are related and can be used as a complementary resource to support each other. Therefore, we developed a complicated gene identification analysis that integrated different data sources and types to maximize the probability of detecting the target genes. We first assemble the carp genome. Lacking long library for scaffolding, RNA-Seq data is not only used for measuring gene expression level but also for scaffolding the DNA contigs. Finally, the TIR domain-containing gene sequences in carp are captured by a comparative genomics analysis using zebrafish resources, since TIR domain is highly conserved and zebrafish genome is well annotated.

In the end, we scaffolded 162 DNA contigs to 39 TIR domain-containing gene sequences. However, only knowing the sequences is not enough. In the future, an ab initio gene prediction algorithm, e.g. AUGUSTUS [21], will be applied on these TIR sequences to further define the gene structures such as the precise start and stop position of a gene and its exons. After having a more complete genome assembly and the gene structures, the TIR containing gene expression in different samples can be measured by mapping the RNA reads to the carp genome using the tools such as TopHat [25] and/or Cufflinks [26].

Acknowledgements

We thank M. Forlenza and G. Wiegertjes from the Cell Biology and Immunology Group, Wageningen University for providing the zebrafish immune gene sequences. This project is partially supported by European Science Foundation FFG Exchange grant No.2952 and BioRange program of the Netherlands Bioinformatics Centre (NBIC, BSIK grant).

References

[1] Serafim Batzoglou, David B. Jaffe, Ken Stanley, Jonathan Butler, Sante Gnerre, Evan Mauceli, Bonnie Berger, Jill P. Mesirov, and Eric S. Lander. ARACHNE: a whole-genome shotgun assembler. *Genome research*, 12(1):177–189, January 2002.

[2] CLC Bio. http://www.clcbio.com/.

[3] A. B. J. Bongers, M. Sukkel, G. Gort, J. Komen, and C. J. J. Richter. Development and use of genetically uniform strains of common carp in experimental animal research. *Lab Anim*, 32(4):349–363, 1998.

[4] Jonathan Butler, Iain MacCallum, Michael Kleber, Ilya A. Shlyakhter, Matthew K. Belmonte, Eric S. Lander, Chad Nusbaum, and David B. Jaffe. ALLPATHS: de novo assembly of whole-genome shotgun microreads. *Genome research*, 18(5):810–820, May 2008.

[5] Richard Durbin, Sean R. Eddy, Anders Krogh, and Graeme Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, July 1999.

[6] Paul Flicek, M. Ridwan Amode, Daniel Barrell, Kathryn Beal, Simon Brent, Yuan Chen, Peter Clapham, Guy Coates, Susan Fairley, and et. al. Ensembl 2011. *Nucleic acids research*, 39(Database issue), January 2011.

[7] Sarah Hunter, Rolf Apweiler, Teresa K. Attwood, Amos Bairoch, Alex Bateman, David Binns, Peer Bork, Ujjwal Das, Louise Daugherty, and et. al. InterPro:

the integrative protein signature database. *Nucleic acids research*, 37(Database issue):D211–D215, January 2009.

[8] C. Iseli, C. V. Jongeneel, and P Bucher. Estscan: a program for detecting, evaluating, and reconstructing potential coding regions in est sequences. *Proc Int Conf Intell Syst Mol Biol*, pages 138–148, 1999.

[9] W. James Kent. BLAT–the BLAST-like alignment tool. *Genome Research*, 12(4):656–664, April 2002.

[10] Ruiqiang Li, Wei Fan, Geng Tian, Hongmei Zhu, Lin He, Jing Cai, Quanfei Huang, Qingle Cai, Bo Li, and et. al. The sequence and de novo assembly of the giant panda genome. *Nature*, 463(7279):311–317, January 2010.

[11] Ruiqiang Li, Hongmei Zhu, Jue Ruan, Wubin Qian, Xiaodong Fang, Zhongbin Shi, Yingrui Li, Shengting Li, Gao Shan, Karsten Kristiansen, Songgang Li, Huanming Yang, Jian Wang, and Jun Wang. De novo assembly of human genomes with massively parallel short read sequencing. *Genome Research*, 20(2):265–272, December 2009.

[12] Scott Mcginnis and Thomas L. Madden. Blast: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res*, 32:20–25, 2004.

[13] A.H. Meijer, S.F. Gabby Krens, I.A. Medina Rodriguez, S. He, W. Bitter, B. Ewa Snaar-Jagalska, and H.P. Spaink. Expression analysis of the toll-like receptor and tir domain adaptor families of zebrafish. *Mol Immunol*, 40(11):773–783, 2004.

[14] Eugene W. Myers. The fragment assembly string graph. *Bioinformatics*, 21(suppl 2):ii79–ii85, September 2005.

[15] Mihai Pop. Genome assembly reborn: recent computational challenges. *Briefings in bioinformatics*, 10(4):354–366, July 2009.

[16] Mihai Pop, Adam Phillippy, Arthur L. Delcher, and Steven L. Salzberg. Comparative genome assembly. *Briefings in Bioinformatics*, 5(3):237–248, September 2004.

[17] Wei Qu, Shin-ichi Hashimoto, and Shinichi Morishita. Efficient frequency-based de novo short-read clustering for error trimming in next-generation sequencing. *Genome Research*, 19(7):1309–1315, July 2009.

[18] F. Sanger, S. Nicklen, and A. R. Coulson. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74(12):5463–5467, December 1977.

[19] Stephan C. Schuster. Next-generation sequencing transforms today's biology. *Nature Methods*, 5(1):16–18, December 2007.

[20] Jared T. Simpson, Kim Wong, Shaun D. Jackman, Jacqueline E. Schein, Steven J. Jones, and Inanç Birol. ABySS: a parallel assembler for short read sequence data. *Genome research*, 19(6):1117–1123, June 2009.

[21] Mario Stanke, Oliver Keller, Irfan Gunduz, Alec Hayes, Stephan Waack, and Burkhard Morgenstern. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucl. Acids Res.*, 34(suppl_2):W435–439, July 2006.

[22] L. Stein. Genome annotation: from sequence to biology. *Nature reviews. Genetics*, 2(7):493–503, July 2001.

[23] O. W. Stockhammer, A. Zakrzewska, Z. Hegedus, H. P. Spaink, and A. H Meijer. Transcriptome profiling and functional analyses of the zebrafish embryonic innate immune response to salmonella infection. *J Immunol*, 182(9):5641–5653, 2009.

[24] Granger G. Sutton, Owen White, Mark D. Admas, and Anthony R. Kerlavage. TIGR assembler: A new tool for assembling large shotgun sequencing projects. *Genome Science and Technology*, 1(1), 1995.

[25] Cole Trapnell, Lior Pachter, and Steven L. Salzberg. TopHat: discovering splice junctions with RNA-seq. *Bioinformatics (Oxford, England)*, 25(9):1105–1111, May 2009.

[26] Cole Trapnell, Brian A. Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J. van Baren, Steven L. Salzberg, Barbara J. Wold, and Lior Pachter. Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology*, 28(5):511–515, May 2010.

[27] Daniel Zerbino and Ewan Birney. Velvet: Algorithms for de novo short read assembly using de bruijn graphs. *Genome Research*, 18(5):821–829, January 2008.

Supplementary results

Gene	Transcript	Name
ENSDARG00000010610	ENSDART00000004200	sarm1
ENSDARG00000010169	ENSDART00000011143	myd88
ENSDARG00000016065	ENSDART00000013021	tlr3
ENSDARG00000040249	ENSDART00000014310	-
ENSDARG00000022048	ENSDART00000034852	tlr4bb
ENSDARG00000026663	ENSDART00000036422	tlr19
ENSDARG00000069592	ENSDART00000044482	CR392351.1
ENSDARG00000037553	ENSDART00000054687	il1rapl2
ENSDARG00000037758	ENSDART00000055006	tlr2
ENSDARG00000058045	ENSDART00000060142	tlr21
ENSDARG00000041054	ENSDART00000060155	-
ENSDARG00000041164	ENSDART00000060337	si:dkey-193n17.7
ENSDARG00000042714	ENSDART00000062685	ticam1
ENSDARG00000043032	ENSDART00000063176	tlr1
ENSDARG00000044415	ENSDART00000065229	TLR5 (1 of 2)
ENSDARG00000044490	ENSDART00000065340	-
ENSDARG00000052322	ENSDART00000074153	tlr5b
ENSDARG00000062045	ENSDART00000089206	il1rapl1a
ENSDARG00000062204	ENSDART00000089680	sigirr
ENSDARG00000038843	ENSDART00000098676	-
ENSDARG00000038843	ENSDART00000098677	il17rd
ENSDARG00000068812	ENSDART00000099649	-
ENSDARG00000062045	ENSDART00000101171	il1rapl1a
ENSDARG00000031859	ENSDART00000101407	-
ENSDARG00000069593	ENSDART00000101409	si:dkey-100n23.2
ENSDARG00000070392	ENSDART00000103242	tlr19
ENSDARG00000075479	ENSDART00000108837	-
ENSDARG00000074371	ENSDART00000109673	tirap
ENSDARG00000078496	ENSDART00000110194	tlr8a
ENSDARG00000079621	ENSDART00000112641	-
ENSDARG00000078740	ENSDART00000112871	il1rapl1b
ENSDARG00000079737	ENSDART00000113028	-
ENSDARG00000075671	ENSDART00000113952	tlr4al
ENSDARG00000076245	ENSDART00000114883	-
ENSDARG00000068609	ENSDART00000099295	il1rl (no hits in carp)

Supplementary Table 1: 35 TIR domain containing peptides in zebrafish. It is found that 34 out of 35 have homologous sequences in carp. We did not find the homologous for i1lr1 gene which is highlighted and showed in the last of the table.

Chapter 6

Conclusions

1 Structure

Annotation is not only about identification of biological molecules but also fully understanding the biological function of these elements. Like the other aspects of molecular biology, the whole procedure of annotation consists of hypothesis creation, validation and refinement. With an abundance of data and tools available, integration is a trend for efficient hypothesis generation. In this thesis, we demonstrated how to improve the target prediction specificity for miRNAs, the post-translational gene regulators, and how to maximize the chance of finding the TIR domain-containing genes in carp using next-generation sequencing data. Integration is the main applied strategy and the complicated procedures are presented using workflows.

In this final chapter, we are going to summarize the main findings of miRNA target prediction and gene discovery. Moreover, we will provide a summary of the lessons learned from the perspective of annotation and integration. Finally, the importance of biological validation will be discussed followed by a vision of the future research.

2 Summary of miRNA target prediction

The mechanism of miRNA regulation in animals is sophisticated. Previous studies have identified several characteristics of miRNA target recognition such as sequence or seed complementarity, stable free energy and target site conservation. However these features cannot fully explain miRNA function mechanism leaving many targets unidentified and many false positive targets. In our study, we found several interesting features.

In Chapter 2, we discovered that there is a correlation between the genomic location of predicted target genes and miRNAs by showing that many targeted genes are physically located close to their miRNAs. Knowing the genomic distance is a related feature, in Chapter 3, we further found that many functionally similar miRNAs are also located in clusters. From these findings, we conclude that genomic distance plays a role in miRNA-target interaction. If two miRNAs or one miRNA and its targets are genomically close, the chance of co-transcription is high. The co-occurrence implies that they might have similar functions or interact with each other. By studying the features of the validated miRNA-target relationships in human, in Chapter 4 we found that some miRNAs tend

to bind their targets at the end of 3' UTR sequences. Integrating our new findings with previous features, we predicted targets for the miRNAs with unknown functions.

With the advances of high-throughput computational approaches for miRNA target prediction, many target candidates are reported. However, the low prediction consistency among the computational tools makes it difficult to screen targets for a final biological validation. In Chapter 4, we tested currently frequently used tools in different datasets as benchmarks for a systematic evaluation. We concluded that TargetScan performs better than miRanda and RNAhybrid with respect to both sensitivity and specificity. Focusing on the overlaps among different tools is not efficient to discover miRNA targets, since it will discard the strength of each method. The proper way is to construct a model to integrate these approaches.

3 Summary of gene discovery

Completion of a genome project including sequencing, assembly and annotation stages is a time, money and labor consuming task. Next-generation sequencing technology is currently still expensive; *de novo* assembly is extremely expensive with respects to computational resources such as CPUs and memory; annotation, however, is the most time consuming. For the model species and human whose genomes have been completely sequenced and well assembled, the current mission is mainly adding the biological context to the sequences. For the non-model species without sequence assembly available, the genome sequences need to be established, before fully annotating the genome.

In Chapter 5, we demonstrated how we annotated a non-model species, i.e. the common carp. We generated 40 GB genomic reads of one paired-end library. The assembly derived from this library is capable of covering almost the whole genome but with fragmented contigs. Lacking long libraries for scaffolding, we used RNA-Seq data to join the fragments together in order to maximize the chance of having the complete gene structure. TIR domain-containing genes are further identified using zebrafish sequences and comparative genomics methods since the TIR domain is highly conserved.

From this project, we learned that data preprocessing is important. Although it will reduce the amount of data, the remaining high quality reads can achieve a better DNA assembly. When having limited genomics data which result in a segmented genome assembly, the

downstream routine analysis such as mapping RNA-Seq data to the assembly using Tophat and Cufflinks in order to measure transcriptome profiling is not practical. In this case, we need to first achieve gene structure. RNA-Seq data can not only measure the expression level but also reveal the exon regions that make up a gene. When lacking long libraries for scaffolding, RNA-Seq libraries could be used for this purpose. Comparative genome analysis such as using BLAST to find highly conserved sequences will speed up the gene finding process.

4 Summary at integration level

In Chapter 2 and Chapter 3, different sources of data, such as genomic distance, sequence similarity, free energy and GO terms are integrated to make the final decision as to whether the target is true or false. Integration is performed by a panel of data mining techniques such as decision trees, relative subgroup discovery and a linear and quadric classifier. In Chapter 4, the intermediated features generated by the three prediction tools are recorded and then further integrated using a Bayesian Network classifier. As discussed in Chapter 4, data integration at a low level, which integrates the raw data, can help to reduce or avoid an error cascade as is seen in the high level integration, that is focused on the integration of the results from other studies. In Chapter 5, data such as genomic DNA reads, RNA-Seq reads and motifs are integrated sequentially. At each step in the workflow, one extra type of data serves as a filter to screen the TIR domain contained candidate sequences.

5 Summary at error reduction level

With the development of current miRNA target prediction tools such as miRanda, TargetScan and RNAhybrid predict, hundreds of targets for each miRNAs are predicted. Among them, only a very small portion is validated as real targets. The high amount of unvalidated predictions not only indicates a high false positive rate but also renders validation of biological experiments rather unpractical. In Chapter 2, by integrating genomic distance between miRNAs and their targets and other enrichment information, targets for dre-miR-10 and dre-miR-196 have been reduced to less than 10 for each. In Chapter 3, using functional similar miRNA for functional unknown miRNA target prediction, 6 new

targets have been predicted as the target candidates for 5 miRNAs. Using heterogeneous data, we greatly reduced the number of candidates to a scale in which the wet experiments can be easily preformed to validate the results.

There is usually a trade-off between sensitivity and specificity. In these two chapters, our aim was to improve the specificity, and the cost of our integration strategy was the reduction of sensitivity. High specificity tools will speed up the process of finding the real targets. In Chapter 4, we integrated three target prediction methods using three integration strategies with the aim to achieve the best performance. The performance is defined with a criterion considering both sensitivity and specificity. In the end, we confirmed the idea that proper integration can improve performance more than any other single method.

6 Limitations

Nevertheless, there are some limitations in our research. First, some cutoffs were set based on our experience and observations. How to decide the cut-off in order to select the candidates is a difficult problem. In our studies, some cut-offs can be optimized according to the error rate using cross validation. Some are set according to a rule of thumb, e.g. p-value ≤ 0.05 is significant. Some cut-off settings are based on the experience of users or references. For example in Chapter 5, when BLASTing sequences within the carp species, we selected the hits if the E-value $\leq$ 1e-20; while BLASTing sequences between zebrafish and carp, the E-value cut-off was set to less than 1e-5. These are based on the fact that the sequence similarity should be higher within the same species than between different species. However, different users or research groups may have difference experiences, therefore the outcome can be different.

Secondly, for data mining, the size of training data is relatively small. In Chapter 3, we used 127 true and false functionally similar miRNA pairs as the training set. This number was, at the time of our experiments, the maximum available number in the human, which has the most validated targets available in the database. In the rule generation of functionally similar miRNAs, we did not mix species. Since most of the miRNAs are conserved, maybe the rules found for humans are transferable to other species.

Thirdly, wet lab validation is currently missing. In Chapter 2 and 3, we predicted high confident targets for several unannotated miRNAs. The number of targets for each miRNA

has been scaled down to less than 10. This can be easily validated by performing the experiments. Since the results of our studies are based on the hypothesis and computational predictions, biological validation is urgently needed.

7 Microarray projects

Microarrays was the main techonolgy measuring transcriptome composition a few years ago. We have also participated in two microarray analysis projects which have not been described in this thesis. The first project concerned the interpretation of microarray time series data of the *Streptomyces coelicolor* ssgC mutant. In this project, the transcriptome of the wildtype and ssgC mutant was measured at 9 different time points over their life span and the main goal was to look for genes which have a different expression in the mutant compared to the wildtype. The second project was zebrafish embryogenesis microarray interpretation using functional and anatomical annotation. The aim was to study the temporal-spatial patterns of developmentally regulated genes during zebrafish embryogenesis. In both research projects, the bottleneck we experienced was the data normalization which is a sophisticated process to remove the bias and noise within and between arrays. Choosing different statistics models or methods for normalization led to different candidates, which had great impact on the downstream analysis. The lesson we learned from these two microarray projects is that bioinfomaticans and biostaticicians should be involved beyond the data analysis. They should be involved in the stage of experiment design as well, since the experiment design directly decides how the data should be analyzed later on. A weak and messy experiment design will not lead to very significant results.

Currently, a new technology for trancriptome analysis is RNA-Seq. It has been applied in carp genome project described in Chapter 5. Compared to microarrays, RNA-Seq can measure the dynamic transcriptome without prior knowledge of genome sequence and has a much higher range of detection, base-level resolution and the ability to detect the previously unknown transcripts. Besides these, the advantages for the data analysis are that it is digital data and does not require sophisticated normalization.

The price of RNA-Seq has dropped dramatically recently. Although currently it is still more expensive than microarrays, in a few years, it will be possible to have 1000 dollar

genome and transcriptome. Although data analysis for RNA-Seq and microarrays differ significantly during the data preprocessing steps, eventually they both measure the gene expression. Therefore many annotation methods for microarrays can theoretically be transferred to RNA-Seq.

8 Conclusion

Annotation is a broad topic and will be one of the main research themes for biology in the future. In this thesis, we demonstrated how we use bioinformatics, and integration in particular, to annotate miRNAs and a novel genome. Although this is just a small part of annotation, we have shown that bioinformatics can guide wet experiments by providing the candidates for validation. By incorporating integration in the workflow, the efficiency and accuracy of bioinformatics predictions can be further improved. Currently, in life science studies high-throughput experiments, multiple platforms and different species as model system are very commonly used. Therefore, heterogeneous data integration is no doubt a trend for the analysis of biological data.

Summary

Bioinformatics is an interdisciplinary field of science which uses methodologies from computer science, mathematics and statistics with the specific aim to create deeper insights from the large amounts of experimental biological data. In molecular genomics the use of bioinformatics is indispensable as the large volumes of data only obtain added value through thorough computational analysis.

Genes are the building blocks for the machinery of cells. Genes are regions of DNA that can be transcribed to messenger RNA and subsequently translated to proteins. The proteins are the chief actors within the cell. Some of the RNA molecules are controlled by small RNA-like structures called microRNAs. These, recently discovered, microRNAs are very short messenger RNAs that are also transcribed from DNA sequences. However, instead of being further translated to protein, these short RNAs bind to messenger RNAs, and, in this manner, inhibit expression of their target.

The complete set of hereditary material of an organism is referred to as the genome. In order to understand the genome we need to be able to label all functional parts. This labeling is referred to as annotation and this is typically the domain of bioinformatics. In this thesis annotation is achieved, in particular, through the use of heterogeneous data integration. The analysis focuses on annotation of genes and molecular structures that control the expression of genes, the microRNAs. The heterogeneous aspect refers to the integration of multiple resources within the analysis so that one can reason efficiently about the data.

The main goal of this thesis is efficient and accurate annotation of microRNAs (miRNAs) and functionally unknown DNA sequences. Gene annotation is the process detecting the structure and biological function of the raw DNA sequences. It is the most time-consuming analysis in a genome project. As for miRNA annotation, the major task currently is to identify miRNAs targets since miRNAs modify gene expression by binding to their target genes. To achieve these goals, we developed several complex workflows which integrate the current most relevant data sources and tools.

In Chapter 2, we explained an integrative method which investigates several aspects of the relationships between miRNAs and their targets with the nal purpose of extracting high

confident targets from the target pool. The applied techniques include statistical tests, clustering and association rules. The research comprised a case study for two miRNAs, i.e. dre-miR-10 and dre-miR-196, in which seven high confidence target candidates were predicted, all of which belong to *hox* gene family and have similar characteristics as already validated target genes.

In Chapter 3, we presented an approach for analyzing miRNA-miRNA relationships and subsequently utilizing these relations for target predictions in human. In support of this a machine learning pipeline was developed in order to reveal the feature patterns between known miRNAs. Subsequently, the observed patterns were applied to miRNAs of which the targets are not yet known so as to see if new targets could be predicted. Our method contributes to the improvement of target identification by predicting targets with high specificity and without constraints on evolutionary conservation.

In Chapter 4, we evaluated the performance of different target prediction algorithms and used integration methods to improve prediction accuracy. To this end, high-level integration approaches, i.e. algorithm combinations and ranking aggregation, as well as low-level integration approaches, e.g. a Bayesian Network classification, were performed. All of the methods were tested on miRNA-target interactions that were experimentally validated and on several compiled negative control data sets. The results showed how each individual prediction algorithm has its own advantages. Moreover, among different integration strategies, the application of the Bayesian Network classifier on the features calculated from multiple prediction methods significantly improved target prediction accuracy.

In Chapter 5, we focused on the assembly and functional annotation of the carp genome. The common carp is a candidate model system that can be used for high throughput screens of pharmaceutical compound libraries. In this chapter, we develop a genome assembly and an annotation pipeline with the final aim of identifying innate immune response genes, especially Toll/Interleukin-1 receptor (TIR) domain-containing genes, using next generation sequencing data. The genome assembly pipeline consists of data cleaning, pre-assembly and assembly using CLCBio, ABySS and SOAP-denovo. A basic gene annotation pipeline is developed by using a simple gene prediction that is based on protein-based gene model prediction as well as comparative annotation. The latter is focused on prediction of orthologues with respect to the zebrafish genome.

As indicated, the central theme throughout this thesis is heterogeneous data integration. In

Chapter 2 and Chapter 3, different features, such as genomic distance, sequence similarity, free energy and Gene Ontology terms are carefully combined to make the final decision whether the target is true or false. Integration is performed by a panel of data mining techniques such as decision trees, relative subgroup discovery and a linear and quadric classifier. In Chapter 4, the intermediated features generated by the three prediction tools are recorded and then further integrated using a Bayesian Network classifier. In Chapter 5, the genome annotation section, different data such as genomic DNA reads, RNA-Seq reads and motifs are integrated in a sequential fashion. Each step in the workflow, adds one extra type of data to serve as a filter to screen the TIR domain containing candidate sequences.

The purpose of using integration is to improve sensitivity and/or specificity of the system. These two measurements characterize the system performance. Sensitivity is defined as the ratio of actual positives which are correctly identified. Specificity measures the probability that the negatives are correctly identified. For each algorithm, it is desirable to achieve both high sensitivity and specificity. There is, however, a trade-off between the measures; high sensitivity will sacrifice specificity by increasing its false positive rate and vice versa. In Chapter 2, by including a feature for genomic distance between miRNAs and their targets and other enrichment information, the number of targets for dre-miR-10 and dre-miR-196 has been reduced to less than 10 for each. In Chapter 3, using functionally similar miRNAs for functionally unknown miRNA target prediction, 6 new targets have been predicted as target candidates for 5 of the miRNAs. Using heterogeneous data, we greatly reduced the number of candidates to a scale in which biologist can easily validate the results. In these two chapters, our aim was to improve the specificity, and the cost of our integration strategy was a slight reduction of sensitivity. Tools with a high specificity will speed up the process of finding the real targets. In Chapter 4, we integrated three target prediction methods using three integration strategies with the aim to achieve the best performance. Performance is defined with a criterion considering both sensitivity and specificity. In the end, we substantiated a concept that proper integration can improve the performance than any other single method. In Chapter 5, by considering both genomic and RNA sequencing data, our purpose was to maximize the probability of finding TIR containing genes in the common carp, therefore sensitivity has been the main focus in this chapter.

The application of the aforementioned methods promotes our understanding of miRNA regulation as well as the structures and function of the novel genes. New biological insights were gained during these studies.

Currently, the mechanism of miRNA regulation in animals is acknowledged as being sophisticated but not yet fully understood; as such many targets are left unidentified and many false positive targets remain. In our study, we found several interesting new features. In Chapter 2, we discovered that there is a correlation between the genomic location of predicted target genes and miRNAs by showing that many targeted genes are physically located close to their miRNAs. Knowing the genomic distance is a related feature, in Chapter 3, we further found that many functionally similar miRNAs are also located in clusters. From these findings, we conclude that genomic distance plays a role in miRNA-target interaction. If two miRNAs or one miRNA and its targets are genomically close, the probability of co-transcription is high. The co-occurrence implies that they might have similar functions or interact with each other. By studying the features of the validated miRNA-target relationships in human, in Chapter 4 we found that some miRNAs tend to bind their targets at either the beginning or the end of 3' UTR sequences.

Gene annotation is a time and labor intensive task. For the non-model species without a sequence assembly available, the genome sequences need to be established, before being able to fully annotate the genome. In Chapter 5, we demonstrated how we annotated a non-model species, i.e. the common carp. We generated huge amount of genomic reads together with RNA sequencing data. In the end, the preliminary carp genome assembly was achieved with an N50 contig length of 2260 bp and it is estimated that the carp genome is about 1.23 Gbp. Compared to zebrafish innate immune genes, we estimated that there are 39 TIR domain-containing genes and transcripts in the common carp.

To sum up, annotation is a broad topic and will be one of the main research themes for biology, and thus bioinformatics, in the future. In this thesis, we demonstrated how we use bioinformatics, and integration in particular, to annotate miRNAs and a novel genome. Although this is just a small part of annotation, we have shown that bioinformatics can guide wet experiments by providing the candidates for validation. By incorporating integration in the workflow, the efficiency and accuracy of bioinformatics predictions can be further improved. Currently, in life sciences high-throughput studies are being incorporated in the experimental workflow, multiple platforms and different model species are

very commonly used. In line with this trend heterogeneous data integration is no doubt an important strategy for the analysis of biological data.

Samenvatting

Bioinformatica is een interdisciplinair onderzoeksveld waarbij methoden uit de computer wetenschappen, wiskunde en statistiek worden gebruikt met het specifieke doel betekenis te geven aan grote hoeveelheden biologisch experimentele gegevens. In de moleculaire genomica is bioinformatica onontbeerlijk; de grote hoeveelheden data die worden gegenereerd verdiepen pas ons inzicht juist door grondige computationele analyse.

Genen zijn de bouwstenen voor de machinekamer van de cel. In feite zijn genen regios in het DNA die kunnen worden overgeschreven naar boodschapper RNAs (mRNA) die vervolgens kunnen worden vertaald naar eiwitten. De eiwitten zijn de belangrijkste actoren in de cel. Sommigen RNA moleculen worden gecontrolleerd door kleine RNA-achtige strucuturen die microRNA worden genoemd. Deze, recentelijk ondekte, microRNAs (miRNA) zijn in feite hele kleine mRNAs en worden net als mRNA ook uit het DNA overgeschreven. Echter, in plaats van de normale vertaling naar eiwit binden deze korte RNA fragmenten aan mRNA en op deze manier kunnen ze het aflezen van het eiwit (het doel) verhinderen.

Het complete erfelijke materiaal van een organisme wordt ook wel het genoom genoemd. Teneinde het genoom te begrijpen moeten we alle functionele dele labelen. Dit proces van labelen wordt annotatie genoemd en de annotatie komt tot stand door bioinformatica. In dit proefschrift wordt annotatie gerealiseerd door middel van het gebruiken en integreren van verschillende, i.e. heterogene, bronnen. De nadruk ligt op het verkrijgen van annotaties voor genen en moleculaire structuren waarmee genen worden gecontrolleerd, de micro RNAs. Het gebruik van heterogene bronnen vergroot daarbij de mogelijkheden voor het redeneren over de data.

Het hoofddoel van dit proefschrift is efficiente en nauwkeurige annotatie van miRNAs en DNA sequenties waarvan tot nu toe geen functie bekend is. Gen-annotatie is het proces waarmee de structuur en functie uit ”ruwe” DNA sequenties wordt verkregen. In een genoom-project is dit het meest tijdrovende deel van de analyse. Wat betreft miRNA annotatie is, in het huidige onderzoek, de belangrijkste taak de doel-RNAs (target) te kunnen vaststellen van een miRNA. Dit omdat miRNA de expressie van een gen controlleert door het binden aan een specifiek doel-mRNA. Teneinde deze verschillende annotatie taken te kunnen realiseren zijn een verscheidene complexe werkschemas (workflows) opgesteld

waarin de op dit moment relevante bron data als ook de analyse technieken worden geintegreerd.

In hoofdstuk 2 wordt een integratie method behandeld waarmee een aantal aspecten worden onderzocht van de relaties tussen miRNAs en het doel-mRNA met de bedoeling om uit de poel van mogelijke doel-mRNAs juist die te selecteren die met een hoge mate van waarschijnlijkheid juist zijn. De technieken die hierbij zijn toegepast omvatten onder andere statistische testen, clustering en zogenaamde associatie regels. Het onderzoek dat in dit hoofdstuk wordt beschreven omvat ook een "case" studie voor twee specifieke miRNAs, te weten, dre-miR-10 en dre-miR-196. Voor deze twee miRNAs werden zeven candidaat mRNAs voorspeld met een hoge waarschijnlijkheids score; alle voorspelde kandidaten behoren tot de hox-gen familie. De voorspelde candidaten delen karakteristieken met genen die reeds gevalideerd zijn.

In hoofdstuk 3 wordt een strategie gepresenteerd voor het analyseren van relaties tussen miRNA's en daarbij wordt vervolgens aangegeven hoe deze relaties gebruikt kunnen worden in de voorspelling van doel-genen zoals die in de mens gevonden kunnen worden. Om dit te ondersteunen is een proces-koppeling ontwikkeld teneinde patronen van kenmerken die tussen bekende miRNAs bestaan, te onthullen. De patronen die gevonden zijn, zijn vervolgens toegepast op miRNAs waarvan de doel-genen nog niet bekend zijn om op die manier mogelijke voorspellingen te kunnen doen over doel-genen van deze miRNAs. Deze methode draagt bij aan de verbeteringen die noodzakelijk zijn voor het identificeren van de doel- genen waarbij de specificiteit vergroot is en de beperking die wordt opgelegd vanwege het principe van conservering van evolutie, niet behoefd te worden toegepast.

In hoofdstuk 4 hebben zijn de verschillende algoritmes die worden toegepast om doelgenen te voorspellen aan een evaluatie onderworpen; daarbij hebben we gebruik gemaakt van methodes van integratie om de nauwkeurigheid van de voorspelling te kunnen vergroten. Daarbij zijn zowel integratie methodes toegepast aan de bovenkant van het spectrum, i.e. combinaties van algoritmen en ranking aggregatie technieken, als ook methodes aan de onderkant van het spectrum, i.e. Bayesiaanse Netwerk classificaties. Al deze methodes zijn getest op miRNA-doel interacties die experimenteel gevalideerd zijn als ook op datasets die samengesteld zijn als negatieve controle data. Uit de resultaten komt naar voren hoe ieder van de voorspellings algoritmen zijn eigen voordelen heeft. Bovendien blijkt dat het gebruik van de Bayesiaanse Netwerk classificatie zoals toegepast op de ken-

merken die berekend zijn uit de verschillende voorspellings algoritmen een significante verbetering geven op de nauwkeurigheid van de voorspelling van het doel-gen.

In hoofdstuk 5 ligt de nadruk op het maken van een assemblage van het genoom van de karper en het annoteren van functies binnen dat genoom. De gewone karper is een nieuw model systeem dat uitermate geschikt is voor zogenaamde ”high-throughput” screenings van verzamelingen van farmaceutisch actieve stoffen. In dit hoofdstuk beschrijven we hoe de genoom assemblage is gerealiseerd en hoe we een proces-koppeling voor annotatie van het genoom maken waarbij we vooral gericht zijn op het het identificeren van de genen verantwoordelijk voor de aangeboren immuunrespons; dit zijn met name de genen die domeinen bevatten voor de Toll/Interleukin-1 receptor (TIR). De analyse is gebaseerd op zogenaamde volgende generatie sequentie gegevens. De proces-koppeling voor genoom assemblage bestaat uit het opschonen van de data, pre-assemblage en assemblage gebruik makend van CLCBio, ABySS en SOAP-denovo software. Een recht toe recht aan gen voorspelling gebaseerd op een proteine gebaseerd gen model voorspellingsmodel tesamen met een vergelijkignsannotatie gebaseerde annotatie vormen een werkbare proces-koppeling voor het annoteren van genen in dit nieuwe genoom. In de vergelijkingsannotatie wordt gebruik gemaakt van de ortologe genen in het genoom van de zebravis.

Zoals eerder vermeld, heterogene data integratie is het centrale thema in dit proefschrift. In de hoofdstukken 2 en 3 worden verschillende kenmerken zoals afstand op het genoom, sequentie similariteit, vrije energie en concepten uit de Gene Ontology zorgvuldig gecombineerd om tot een eindoordeel te komen of een voorspelling omtrent een doel-RNA goed of fout is. Integratie wordt gerealiseerd door een combinatie van data mining technieken zoals, beslisbomen, relatieve subgroup discovery, een lineaire classifier en een kwadratische classifier. In hoofdstuk 4 worden de intermediaire kenmerken, gegenereerd uit de drie geselecteerde voorspellingsmethoden, vastgelegd en geintegreerd met een Bayesiaanse Netwerk Classifier. In hoofdstuk 5, in de sectie die handelt over genoom annotatie, worden verschillende typen van data zoals DNA fragmenten, RNA-Seq fragmenten en RNA motieven geintegreerd op een sequentieele wijze. Elke stap in het werkschema voegt een nieuw type data toe dat werkt als een filter om te zoeken naar de TIR-domein bevattende kandidaat sequenties in het karper genoom.

Het doel van de integratie is het verbeteren van de sensitiviteit en/of de specificiteit van het

systeem. Deze twee maten karakteriseren de prestatie van het systeem. Sensitiviteit wordt gedefinieerd als de ratio van het aantal positieven en het aantal correct geidentificeerde positieven. De specificiteit drukt de waarschijnlijkheid uit dat de negativen correct geindentificeerd worden. Context van predictie is het wenselijk dat een algoritme zowel een hoge sensitiviteit als een hoge specificiteit kan realiseren. Er is echter een afweging tussen deze maten, een hoge sensitiviteit zal de specifiteit benadelen omdat de maat van foute positieven toeneemt en vice versa. In hoofdstuk 2 wordt een kenmerk voor genomische afstand tussen miRNA en hun doel genen toegevoegd en de analyse wordt verder verrijkt met aanvullende informatie. Hierdoor kan het aantal potentieele doel genen voor de miRNAs die werden onderzocht, dre-miR-10 and dre-miR-196, worden teruggebracht tot minder dan 10 per miRNA. In hoofdstuk 3 worden functioneel vergelijkbare miRNAs gebruikt voor de predictie van doel genen van miRNA waarvan de functie nog niet bekend is. Op deze wijze zijn 6 nieuwe doel-genen geindentificeerd voor 5 van de miRNAs uit het experiment. Op basis van heterogene data bronnen zijn we in staat geweest het aantal potentieele candidaat doel-genen terug te brengen tot een omvang waarbinnen de bioloog een validatie experiment kan opzetten. In de hoofdstukken 3 en 4 was het doel te onderzoeken hoe de specifiteit kan worden verbeterd. Door het toepassen van een strategie van data integratie zijn we in staat geweest dit te realiseren met slechts een kleine vermindering van de sensitiviteit. Een hoge specificiteit versnelt het proces van het vinden van de doel-genen aanzienlijk. In hoofdstuk 4 zijn 3 doel voorspellings methoden geintegreerd waarbij 3 verschillende integratie strategieen zijn gebruikt, hierbij is specifiek gelet op het behalen van de best mogelijke prestatie. De prestatie is uitgedrukt in een maat waarin zowel specificiteit als sensitiviteit worden meegenomen. Op basis van de resultaten hebben we een concept kunnen uitwerken hoe een correcte integratie de prestatie aanzienlijk kan verbeteren in vergelijking met ieder van de methoden afzonderlijk. In hoofdstuk 5 hebben we gestuurd naar het maximaliseren van de waarschijnlijkheid om de TIR-domein bevattende genen te vinden in het genoom en in RNA fragmenten van de karper. Hierdoor heeft de focus vooral gelegen op de sensitiviteit.

Met het toepassen van de hier genoemde methoden wordt ons inzicht in de miRNA regulatie als ook de structuur en functie van nieuwe genen bevorderd. In deze studies zijn daarmee ook nieuwe inzichten in de biologie verkregen.

Het mechanisme van miRNA regulatie in dieren wordt gezien als zeer verfijnd, echter

tot op heden is het nog niet geheel doorgrond; getuige het feit dat veel doel-genen nog niet zijn geindentificeerd en daarbij nog veel fout positieven overblijven. In onze studies hebben we een aantal interessante nieuwe kenmerken kunnen vaststellen. In hoofdstuk 3 hebben we het verband opgehelderd tussen doel-genen en miRNA door te laten zien dat veel van deze doel-genen fysiek dicht gelokaliseerd zijn bij het miRNA dat ze reguleert. Uitgaande van dit feit hebben we in hoofdstuk 3 kunnen laten zien dat functioneel vergelijkbare miRNAs gelokaliseerd zijn in clusters. Uit deze bevindingen kunnen we conlcuderen dat afstand op het genoom een belangrijke rol speelt in miRNA-doel interactie. Als twee miRNAs of een miRNA en zijn doel-gen dicht bij elkaar op het genoom zijn gelokaliseerd, dan is de waarschijnlijkheid dat ze tegelijk worden afgelezen groot. Dit gezamelijk voorkomen impliceert dat ook de functies vergelijkbaar zijn of dat ze in een interactie voorkomen. Door de kenmerken van reeds gevalideerde humane miRNA-doel gen relaties te bestuderen hebben we kunnen vaststellen dat sommige miRNA's binden op hun doel juist aan het begin of aan het eind van de 3' UTR sequentie.

Het annoteren van genen is een tijdrovende en arbeidsintensieve taak. Voor niet-model organismen waarvoor geen genoom assemblage beschikbaar is, moet eerst dit bewerkstelligd worden voordat men in staat kan zijn het genoom te annoteren. In hoofdstuk 5 wordt gedemonstreerd hoe een niet-model organisme, i.e. de gewone karper, kan worden geannoteerd. Een grote hoeveelheid data wordt gegenereerd voor de genoom en RNA sequentie analyse. Uit deze data is een voorlopig karper genoom samengesteld met een N50 contig lengte van 2260 base paren. De inschatting is dat het karper genoom een lengte heeft van ongeveer 1.23 giga base-paren. In een vergelijking met de aangeboren immuno genen van de zebravis kunnen we schatten dat er 39 TIR-domein bevattende genen en afschriften zijn in de gewone karper.

Samenvattend, annotatie is een veelomvattend onderwerp en een belangrijk onderzoeksthema voor de biologie, en daarmee voor de bioinformatica, voor nu en voor de toekomst. In dit proefschrift laten we zien hoe bioinformatica en integratie kan worden gebruikt, in het bijzonder voor de annotatie van microRNA's en genomen. Dit is slechts een klein deelgebied van annotatie, desalniettemin hebben we laten zien hoe bioinformatica een leidraad kan zijn voor laboratorium experimenten door te voorzien in kandidaten die gevalideerd kunnen worden.

List of publications

1. Yanju Zhang, Elia Stupka, Christiaan V. Henkel, Hans J. Jansen, Herman P. Spaink and Fons J. Verbeek. Identification of Common Carp Innate Immune Genes with Whole-Genome Sequencing and RNA-Seq Data. *Journal of Integrative Bioinformatics*, 8(2):169, 2011. Online Journal: http://journal.imbio.de/index.php?paper_id=169

2. Yanju Zhang and Fons J. Verbeek. Comparison and Integration of Target Prediction Algorithms for microRNA Studies. *Journal of Integrative Bioinformatics*, 7(3):127, 2010. Online Journal: http://journal.imbio.de/index.php?paper_id=127

3. Yanju Zhang, Jeroen S. de Bruin and Fons J. Verbeek. Specificity Enhancement in microRNA Target Prediction through Knowledge Discovery. *Chapter 20 In: Machine Learning*, Eds. Yagang Zhang, ISBN: 978-953-307-033-9, INTECH, February 2010

4. Yanju Zhang, Jeroen S. de Bruin, Fons J. Verbeek. miRNA target prediction through mining of miRNA relationships. *Proceedings of the 8th IEEE International Conference on Bioinformatics and Bioengineering*. ISBN: 978-1-4244-2844-1, 1-6, 2008

5. Yanju Zhang, Joost M. Woltering, Fons J. Verbeek. Screen of microRNA targets in zebrafish using heterogeneous data sources: a case study for dre-miR-10 and dre-miR-196. *International Journal of Mathematical, Physical and Engineering Sciences*, Vol. 2 (1), 10-17, 2007

6. Gihan Dawelbait, Christof Winter, Yanju Zhang, Chrisitian Pilarky, Robert Grtzmann, Jrg-Chrisitan Heinrich and Michael Schroeder. Structural templates predict novel protein interactions and targets from pancreas tumour gene expression data. *Bioinformatics*, 23:i115–24, 2007

7. Gihan Dawelbait, Christian Pilarsky, Yanju Zhang, Robert Grtzmann and Michael Schroeder. Structural protein interactions predict kinase-inhibitor interactions in upregulated pancreas tumour genes expression data. In Kay Diederichs, Robert Glen, and Oliver Kohlbacher, editors, *Proceedings of the 1st International Symposium on Computational Life Science*. Springer LNBI, 2005

Curriculum Vitae

Yanju Zhang was born on 11 June 1980 in Guilin, Guangxi, China. She studied Computer Science at Guilin University of Electronic Technology, China and received the bachelor degree in 2002. After that she worked as a teaching assistant at this university for one year. From 2003 to 2005, she studied Molecular Bioengineering at Technological University Dresden, Germany. She obtained her master degree with a thesis on bioinformatics, i.e. *using protein-protein interactions and metabolic pathways to interpret tumor gene expression data*. Since 2006, she started her PhD at the Leiden Institute of Advanced Computer Science, Leiden University in the section Imaging and BioInformatics under the supervision of Dr. Ir. Fons J. Verbeek. This project was part of the BioRange project, a large national BioInformatics programme under the supervision of the Netherlands BioInformatics Centre. The focus of her research has been heterogeneous data integration and methodology development for the analysis of biological data. Her research interests include applying machine learning algorithms and visualization in bioinformatics, pathway and interaction network modeling, sequencing analyses and comparative genomics. In March 2010, she was selected for a Frontiers of Functional Genomics grant award and did her internship at the group of Dr. Elia Stupka in the Ccancer Institute of London, UK. Here she elaborated on a carp genome sequencing, assembly and annotation project that was initiated at the Molecular Cell Biology of the Leiden Institute of Biology. From 2010 until now, she has been working in the lab of Prof. dr. Eline P. Slagboom in department of Molecular Epidemiology at Leiden University Medical Center. Her research work concerns methodology and algorithm development for RNA-Seq data and methylation arrays.

Acknowledgements

Wrapping up all my research into this small thesis has not been an easy job. This dissertation would not have been possible without the support, inspiration and help from all the kind people around me.

I enjoyed talking to Dr. Erno Vreugdenhil who patiently explained the background and biological function of microRNA to me. Thanks, also, to Prof. Herman P. Spaink for getting us involved in the carp genome project. This next-generation sequencing project has had a great impact on my subsequent research.

I am very grateful to Dr. Christiaan Henkel who showed me lots of techniques and approaches for analyzing sequencing data. Chris, your suggestions and inspiration played an important role in developing my approach to the carp project.

I would like to thank Elia Stupka for hosting my internship at UCL, London. Elia, I learned a lot from you and I had a great time.

Many thanks to my colleagues at the LIACS. It has been a pleasure working with you. In particular, I am indebted to Amalia, Laura and Joris for helpful suggestions in English writing. As friends, you guys always comforted and encouraged me at times I was struggling which I have appreciated very much. Kuan, I would specifically like to thank you for criticizing my work and thereby encouraging me to consider every analysis critically. And I am grateful to Julia, Peng, Dome and Zi for being kind to me and giving me support in both science and life. Thanks, Yun, for me you are like a sister.

I would like to thank Fabrice who shared the office with me during the first three years of my PhD. Your quick completion of your PhD study set an example for us and your positive attitude towards research affected me a lot. And thanks, Juan, for inviting me to the Dutch Conversation Night and showing me many weird and funny things in Holland.

I owe a debt of gratitude to Eline and Kai for giving me an opportunity to work in the MolEpi group at the LUMC after my PhD studies. Thanks Steffan and Eric-Wubbo for offering me all kinds of help to finish up this thesis.

Thanks Xiaohong and Lu for being my paranymphs. Xiaohong, we have known each other for almost five years. You are such a good friend and secret keeper with whom I can

share both happiness and sadness. Lu, such a nice girl living upstairs of my apartment, I always enjoy having tea or dinner with you at your place when I feel tired and lonely downstairs.

I also would like to thank my badminton friends, especially Hying, Richard and Kian. Playing badminton has been my major way to release stress. I am also glad that we are friends beyond the badminton courts. It was fun having dinner and boat trips with you. Thanks Elise, Xiaodong and Gary for your company.

I owe my deepest gratitude to my dear brother Fulin. Thanks for taking good care of our parents. You are also a good brother whom I can always count on. I thank my sister-in-law Aiqiong and my niece Wenjia for bringing so much happiness to our family.

Finally, millions of thanks to my parents. Dad and Mom, you are the best in the world. Your kindness, persistence and other good traits set me a good example. Thank you for supporting me in studying abroad, for always being there for me whatever happened, for sending all kinds of good food to me from many miles away. I will be very grateful to you for all my life.

Yanju Zhang, October 2011