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

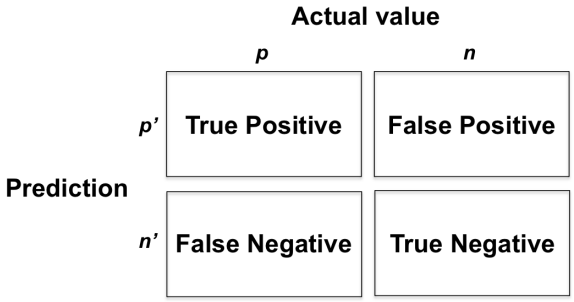


Figure 1: Definitions of true positive (TP), false positive (FP), false negative (FN) and true negative (TN) in binary classification. Positive (p) and negative (n) are the two classes, and p' and n' are the prediction outcomes. A true positive occurs when a positive instance is predicted as positive; 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.

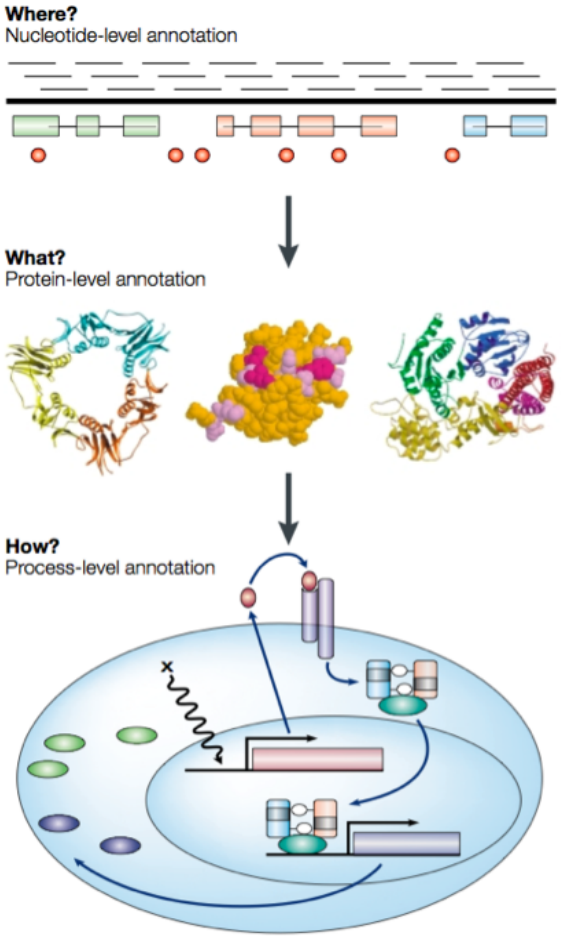


Figure 2: Three layers of genome annotation. Nucleotide-level annotation aims for identifying the physical map of the functional units. Protein-level annotation aims for identifying 3D protein configurations and protein-protein interactions. Process-level annotation aims for identifying the biological 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

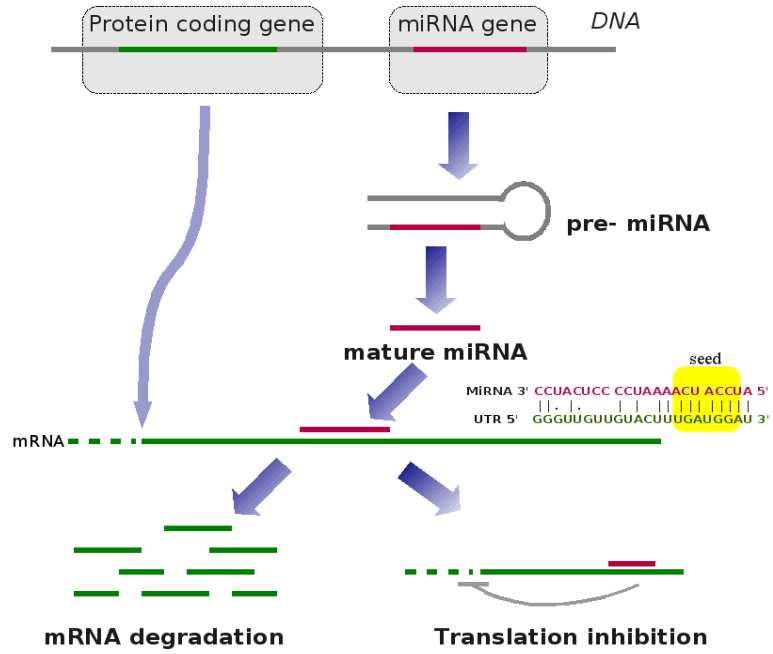


Figure 3: Simplified illustration of miRNA biogenesis and function. miRNA genes are first transcribed to pre-miRNA, and then processed to mature miRNAs. Upon binding to these miRNAs through sequence complementarity, the messenger RNAs (mRNAs), which are called the targets of miRNAs, will be either degraded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

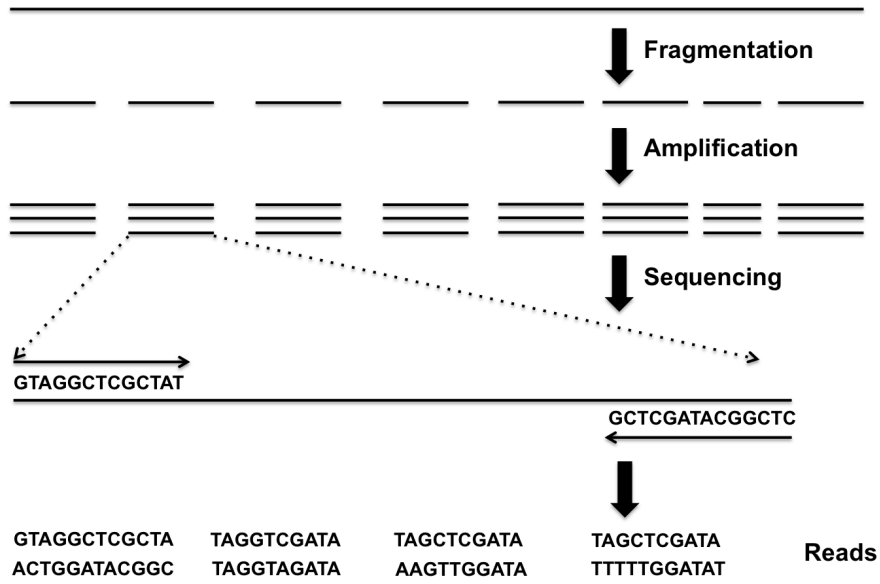
The miRNAs are also found to be involved in the pathogenesis of infectious diseases and cancer. It was discovered that miR-107 is associated with Alzheimer's disease [33]; miR-133b is related to Parkinson's disease; miR-1 plays a role in the development of cardiovascular diseases [9]. These findings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

4.2 Genome project

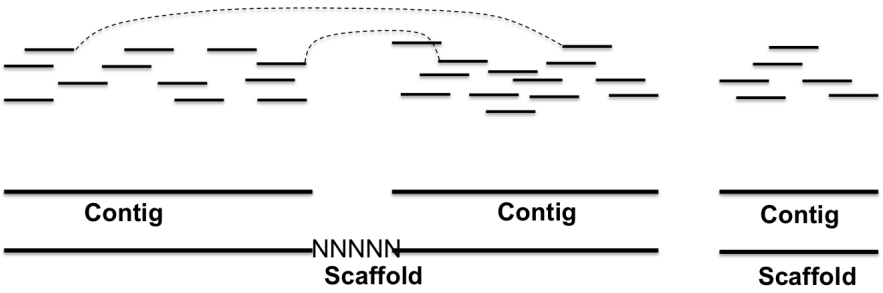
The human genome project, initialized in 1993, released a draft and a complete genome assembly in 2000 and 2003 respectively. These groundbreaking results showed that scientists are capable of decoding the full set of DNA that make a human. Since then, many genome projects of different species, such 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.