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.