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.