Chapter 5

Conclusions

We have presented computational analyses of transcriptional regulation using genome-scale heterogenous data sources. We started with analysis of putative modules derived from ChIP-chip data and integrated gene expression and functional annotation data for detection of functional signals. Different expression data enabled us to identify condition-specific and condition-invariant properties of such functional modules. By utilizing those modules, we further investigated regulatory relationships between transcription factors (TFs) and target genes to prioritize them for more detailed experimental and mechanistic studies. Multiple TFs appearing in modules imply combinatorial regulation and our novel approach revealed condition-specific combinatorial TF pairs and showed that it is statistically significant.

Given the fact that we explored genome-wide data, it may come as a surprise that our predictions are very limited. In principle, current data integration methods aim to utilize diverse data sources to reduce noise in those data. No experimental measurements are complete and precise in the first place. Cells are dynamic and constantly changing. It is difficult to collect data which provide a coherent and consistent picture out of them. All analyses are bound to be *ad hoc* in this sense. Thus, we tried to show that simple computational methods can serve as *practical* tools to generate meaningful biological hypotheses or predictions. Those outcomes are considered as robust signals which can be detected from noisy and inconsistent multiple data sources under a particular condition. Under these circumstances, those data sources give rise to small overlaps among them as results of our analyses imply. Although we also attempted to take care of experimental consistency in data integration such as careful selection of gene expression data with respect to ChIP-chip data, computational efforts will remain immature without a better understanding of experimental parameters and dynamic details associated with all types of data which will be produced with higher accuracy and coverage.

Under the assumption that computational methods have been applied to high quality data, one salient problem in bioinformatics research remains in general as follows. There have been virtually millions of hypotheses from a large number of computational studies about genes, proteins, regulatory motifs, cellular processes and particularly biological functions, which would not be possible to test all of them experimentally for verification or falsification. The list of hypotheses is expected to be ever increasing as experimental technologies generate more and more high-throughput data (let alone arbitrary statistical thresholding). Hence, the critical scientific issue is how to *disprove* such hypotheses generated from various methods and even sophisticated systematic tools in different scientific disciplines which deal with the same biological problems.

In conclusion, our simple but general computational approaches can be applied to large-scale biological data for effective and automated analysis as a first step. Our analyses successfully revealed meaningful biological findings and generated concrete hypotheses from heterogeneous genome-wide data. Therefore, this work is expected to contribute to both helping experimentalists and dissecting underlying biological mechanisms.