Computational Genomic Analysis of Transcriptional Regulation

Ho-Joon Lee

Dissertation zur Erlangung des Grades
eines Doktors der Naturwissenschaften (Dr. rer. nat.)
am Fachbereich Mathematik und Informatik
der Freien Universität Berlin

Februar 2008

1. Referent: Prof. Dr. Martin Vingron

2. Referent: Prof. Dr. Hanspeter Herzel

Tag der Promotion: 29. April 2008

Acknowledgements

As is true for everyone, I have also arrived at this point of achieving a goal in my life through various interactions with and help from other people. However, written words are often elusive and harbour diverse interpretations even in one's mother language. Therefore, I would not like to make efforts to find best words to express my thankfulness other than simply listing those people who have contributed to this thesis itself in an essential way. This work was carried out in the Department of Computational Molecular Biology at Max Planck Institute for Molecular Genetics in Berlin.

First of all, I would like to thank my advisor and office mate, Dr. Thomas Manke, without whom this work would have not been possible. Also I would like to thank Prof. Martin Vingron for his overall support; Dr. Ricardo Bringas for his excellent collaboration during his visits to Berlin; Prof. Hanspeter Herzel for his stimulating communications; members of the department for their helpful feedbacks and interactions. There are numerous other people too who have shown me their constant support and friendship in various ways, directly or indirectly related to my academic life. I will remember them in my heart and hope to find a more appropriate place to acknowledge them in the future.

Finally, my special and unique thanks must go to my parents and brother for their unconditional love and support for every part of my life.

Abstract

Modern technological advances have been producing a huge amount of high-throughput genome-/proteome-wide data which are to be analyzed for inferring biological knowledge. Computational and statistical analyses are an appropriate and efficient way for such large-scale data analysis. In this thesis we investigate genome-wide transcriptional systems by data integration, which is also a prerequisite for systems biology. Computational and statistical methodologies are developed and applied to heterogeneous genome-wide data sources in a model organism, *Saccharomyces cerevisiae*. We aim to discover strong functional signals and related mechanisms from noise-prone genome-scale transcriptional data.

First, our analysis starts with groups of genes bound by common transcription factors, called transcriptional modules. They are derived from protein-DNA interaction data and coupled to gene expression and functional annotation data in order to identify functional signals. Standard methods applied to various large-scale gene expression data show that those identified functional modules can be condition-invariant or condition-specific. Second, we extend our module analysis to prioritization of gene regulatory interactions in functional modules identified on a large scale. Our simple integrative approach to such prioritization yields a statistically significant increase of prediction accuracy for two types of reference datasets compared with an original analysis of genome-wide protein-DNA interactions data alone. In addition, our predictions include those regulatory interactions that were not predicted by other algorithms with as good prediction accuracy. Finally, in view of ubiquitous combinatorial regulation by multiple transcription factors, we turn our attention to different sets of target genes in different conditions regulated by pairs of regulators. We develop a method to identify condition-specific co-factors of those regulators that significantly change their target genes in different conditions. We apply the method to genome-wide

protein-DNA interactions data generated in diverse cellular conditions. Our predictions include novel cooperative regulator pairs as well as known ones with evidences from gene expression, protein-protein interactions, and conserved motifs data. Further analysis shows that such condition-specific combinatorial regulation occurs more abundantly than expected by chance.

In conclusion, our analyses successfully reveal meaningful biological findings and generate concrete hypotheses from heterogeneous genome-wide yeast data. Therefore, this work is expected to contribute as a first step to guiding experimentalists and studying more detailed biological mechanisms.

Contents

List of Tables

List of Figures