

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

1	Introduction	1
1.1	Gene regulation	2
1.1.1	General aspects	2
1.1.2	Modular organization of biological systems	4
1.1.3	Combinatorial regulation	5
1.2	Large-scale experimental approaches	6
1.2.1	Protein-DNA interactions	6
1.2.2	Transcript expression profiling	8
1.2.3	Functional annotations	11
1.3	Computational approaches to data integration	12
1.4	Contributions of the thesis	13
2	Functional analysis of transcriptional modules	15
2.1	Background	15
2.2	Transcriptional modules from binding data	16
2.3	Characterization of functional modules	17
2.4	Condition-invariance and condition-specificity	29
2.5	Summary	31
3	Prioritization of gene regulatory interactions	32

3.1	Background	32
3.2	Overview of our approach	33
3.3	Coherent modules	34
3.3.1	Putative transcriptional modules from binding data	36
3.3.2	Defining coherent modules	38
3.4	Prioritization of gene regulatory links	39
3.5	Evaluation of the method	41
3.5.1	Reference datasets	43
3.5.2	Validation	44
3.5.3	Comparison with other methods	45
3.5.4	Difficulty of comparisons	48
3.6	Biological examples	49
3.6.1	Functionally interacting proteins	50
3.6.2	Conserved binding sites for three regulators of CIS3	51
3.7	Summary	53
4	Condition-specific combinatorial regulation	55
4.1	Background	55
4.2	Identification of condition-altered TFs by a hypergeometric test	57
4.3	Systematic study of condition-specific co-factors	58
4.4	Condition-specific combinatorial regulation is statistically significant	64
4.5	Support for condition-specific combinatorial regulation	66
4.5.1	Expression analysis	66
4.5.2	Conserved motif	71
4.5.3	Protein-protein interaction	71
4.6	Summary	72
5	Conclusions	73

CONTENTS

Bibliography	79
Zusammenfassung	91

List of Tables

2.1	Summary of module analysis results	18
2.2	Significant modules identified	22
2.3	Significant functions identified	23
3.1	Positive predictive values and parameters	41
3.2	Comparison of performance measures	46
4.1	ChIP-chip conditions and the numbers of TFs assayed	57
4.2	Predicted combinatorial TF pairs and supports	62
4.3	Expression data sources	67

List of Figures

1.1	ChIP-chip data matrix	9
1.2	Gene expression data matrix	11
2.1	Expression coherence test	20
2.2	Statistically significant modules	21
3.1	Overview of our method	35
3.2	Summary of our final predictions	42
3.3	Conservation of TF binding sites	52
4.1	Condition-altered TFs and condition-specific co-factors	59
4.2	An alternative way of identifying condition-specific co-factors	61
4.3	Significance test	65
4.4	Synergistic expression analysis	70