PROKARYOTIC NETWORKS RECONSTRUCTION

Peyman Zarrineh

Jury:
Prof. dr. ir. Yves Willems (chairman)
Prof. dr. ir. Bart De Moor (promotor)
Prof. dr. ir. Kathleen Marchal (co-promotor)
Prof. dr. ir. Yves Moreau
Prof. dr. ir. Jos Vanderleyden
Dr. ir. Katrijn Van Deun
Prof. dr. ir. Victor M. Eguíluz (Institute for Cross-Disciplinary Physics and Complex Systems, Spain)
Dr. ir. Tom Michoel (Freiburg Institute for Advanced Studies School of Life Sciences, Germany)

Dissertation presented in partial fulfilment of the requirements for the degree of Doctor in Electrical Engineering

November 2011

© Katholieke Universiteit Leuven – Faculty of Engineering
Address Kasteelpark Arenberg 10, B-3001 Leuven (Belgium)

Alle rechten voorbehouden. Niets uit deze uitgave mag worden vermenigvuldigd en/of openbaar gemaakt worden door middel van druk, fotokopie, microfilm, elektronisch of op welke andere wijze ook zonder voorafgaandelijke schriftelijke toestemming van de uitgever.

All rights reserved. No part of the publication may be reproduced in any form by print, photoprint, microfilm, electronic or any other means without written permission from the publisher.

D/2011/7515/147
ISBN 978-94-6018-446-8

Acknowledgments

When I was accepted in University of Tehran as a bachelor student of Computer Science, I could not even think one day I will end up here. After six years of education and research in bioinformatics, I used to answer the first question of ordinary people “what is bioinformatics?”. I never forget the time that I saw a PhD job proposition in a system biology website, and I contacted Prof. Kathleen Marchal. Fortunately, two months later I was invited to Leuven for an interview, and I was accepted. Therefore, I had opportunity to expand my knowledge in various branches of sciences such as Genetics, Statistics, and even algorithm and computing. In addition, my knowledge about history, culture, and languages was extremely increased due to the unique location and characteristics of Levuen as a city and Belgium as a country in the heart of Europe. Here I want to express my gratitude towards people and institutes who supported me during my academic education.

First of all, I want to thank my promoters prof. Bart De Moor and Prof. Kathleen Marchal for their scientific, financial, and spiritual supports. I would also like to thank three special colleagues, Dr. Carolina Fierro, Dr. Alejandra Herrada, and Aminael Sanchez Rodriguez, with whom I directly contributed, and we tackled some interesting research problems in my field of research. I would also like to express my gratitude to Prof. Victor M. Eguiluz, and Dr. Jose Javier Ramasco from institute for Cross-Disciplinary Physics and Complex Systems (IFISC) in Palma de Mallorca for their scientific supports which empower me to analysis complex biological networks.

In addition, I want to express my special thanks to all the professors, colleagues, and secretaries in ESAT, CMPG, and IFISC who supported me and with whom I shared great memories in the office and also outside of the office, specially the former and current members of Prof. Marchal’s research group, Dr. Kristof Engelen, Dr. Pieter Monsieurs, Dr. Inge Thijs, Dr. Karen Lemmens, Dr. Abeer Fadda, Dr. Tim Van den Bulcke, Dr. Hui Zhao, Dr. Riet De Smet, Dr. Valerie Storms, Dr. Hong Sun, Marleen Claeys, Thomas Dhollander, Qiang Fu, Ivan Ischukov, Pieter Meysman, Lore Cloots, Lyn Venken, Yan Wu, and Dries De Maeyer.

Furthermore, I would also like to thank the chair prof. Yves Willems and members of the jury and also my assessors: Prof. Jos Vanderleyden, Prof. Yves Moreau, Prof. Iven Van Mechelen, Dr. Tom Michoel, and Dr. Katrijn Van Deun for providing valuable comments and suggestions to improve this PhD dissertation.

I am highly grateful to my previous professors in University of Tehran in Iran, Chalmers University of Technology in Sweden, Katholieke Universiteit Leuven, and also my parents who have encouraged and supported me during my long time education. I would like to mention some Iranian friends in Leuven including CMPG doctoral student Hassan, with whom we organized several tea breaks with hard social-political discussions, and also Sia, Kamran, and Pooya who helped me to introduce new faces of Iranian culture in Leuven. Thanks to Anita, Elsy, Ida, Ilse, Mimi, Veronique, and Lodewijk who helped me to overcome the complicated bureaucracy in administrative related matters.

SUMMARY

Availability of various genome-wide datasets provides the opportunity to study the whole genome behavior of the organisms as well as predicting new functions for unknown genes. With the advent of the *omics* data, molecular biology has evolved from a rather data-poor to an extremely data-rich era. Several smaller scale studies already have shown how the integration of different omics data can result in a better mechanistic understanding of the cellular organism. In addition to omics data, co-expression cross-species comparison is also useful to expand the available information from better studied organism to other organisms or strains where the available data is limited.

In the first part of the study, we described a new co-expression cross-species comparison method to analyze microarray datasets comparatively across species and to identify the co-expressed modules of genes. For this aim, we developed a method referred to as COMODO (COnserved MODules across Organisms) that uses an objective selection criterium to identify conserved expression modules between two species. The method uses as input microarray data and a gene homology map and provides as output pairs of conserved modules and searches for the pair of modules for which the number of sharing homologs is statistically most significant relative to the size of the linked modules. We demonstrated the performance of COMODO using distantly related two model bacterial systems, *Escherichia coli* and *Bacillus subtilis*. As a notable result, we identified larger size of conserved co-expressed modules than previously predicted to exist. In addition, we identified co-expressed modules of similar elementary processes with totally different regulatory mechanisms. Later, we discussed the statistics to assess co-expression conservation between two or three organisms, and we expanded COMODO to detect the co-expression conservation across three organisms. We applied COMODO to study the expressional conservation and divergence across *E. coli*, *Salmonella enterica*, and *B. subtilis*. We observed several modules just conserved in *E. coli* and *S. enterica* including many modules related to response to various stimuli and signal transductions, even though some aspects of their life style are remarkably different (pathogenicity of *S. enterica*). Moreover, based on the conserved co-expressed modules, we could predict some conservation in the regulatory interaction of *E. coli*

and *S. enterica* although the regulatory network is not available for *S. enterica*. Furthermore, we also investigated the co-expression conservation of genes involved in two special functions, *quorum sensing* and pathogenicity across *E. coli* and *S. enterica*, and we could observe fair conservation for genes involved in quorum sensing, but almost no conservation for genes involved in pathogenicity. In fact, *S. enterica* contains a much larger number of genes related to pathogenicity that are considered the main causes of difference in life style of the two phylogentically close species *E. coli* and *S. enterica.*

In the second part of this study, first we explored the mutual relation between the regulatory network and microarray expression compendium in *E. coli*. For this aim, we tried to detect modules in the regulatory network which may resemble combinatorial regulators by using Fisher exact test and Monte Carlo sampling. As both of the methods, Fisher exact test and Monte Carlo sampling, failed to find modules in the regulatory network, at the next attempt we tried to define a similarity measure for each pair of genes based on their common regulators; we called it co-regulatory similarity. PageRank value was used as a measure to assess the importance of a regulator in the regulatory network. Based on this measure, the more important regulators were happened to be the more global regulators. This facilitated to define the co-regulatory similarity measure between each pair of genes based on the PageRank value of their common regulators. In our definition, regulators with lower PageRank values (more local regulators) contribute more in the co-regulatory similarity of their targets. We showed this co-regulatory similarity measure exhibits high correlation with the observed co-expression on the microarray expression compendium. Based on this study we could conclude that the observed co-expressed modules are the effect of the structure of the whole regulatory network rather than a set of combinatorial regulators.

We also studied the mutual relation between the regulatory network as the controlling network and the other interaction networks with non-controlling roles in the cell. To process non-controlling interaction networks, we detected biological modules. These biological modules included modules detected in protein-protein interaction network and EcoCyc cellular pathways. The average co-regulatory similarity values of all gene pairs in each biological module were much higher than what is expected for random genes. We also performed the analysis in the

other direction, we detected modules with high co-regulatory similarity values derived from the regulatory network. We found high similarity between these expected modules based on the regulatory network and actual biological modules. In addition, we also compared the hierarchy of biological modules, built by using regulatory networks, with the one, built by using functional GO terms. The regulatory similarity between each two modules could easily be calculated by averaging our defined co-regulatory similarity value between each pair of genes across two modules. For the functional similarity, we introduced new species-specific functional similarity measure for a pair of genes, and we calculated the average value of this similarity measure between each pair of genes across two modules. We could observe rather high correlation between functional similarity value and co-regulatory similarity of two modules, implying the hierarchies built by these two measures are highly related. Based on our observation, we could explain that despite the rapid evolution of the regulatory network, the rewiring in this network would be in the direction to keep the biological modules conserved and also in higher level preserve the functional hierarchy.

SAMENVATTING

De beschikbaarheid van genoomwijde datasets biedt de mogelijkheid om organismen in hun globaliteit te bestuderen en de functie van nog ongekende genen te voorspellen. Moleculaire biolgie is geëvolueerd naar een datarijk onderzoeksdomein. Verschillende studies hebben reeds aangetoond dat integreren van omics data vaak resulteert in een beter globaal inzicht in het cellulaire gedrag. Bovendien, laat het vergelijken van omics informatie over de species heen toe om informatie van gekende organismen te extrapoleren naar minder bestudeerde organismen.

In het eerste deel van dit werk beschrijven we een nieuwe cross-species coclustering strategie, COMODO (COnserved MODules across Organisms) die toelaat om coexpressie informatie te vergelijken tussen species. De methode gebruikt als input microarray data en homologie relaties en geeft als output paren van geconserveerde coexpressie modules waarvoor het aantal gedeelde homologen statistisch significant is t.o.v. het aantal genen in de modules. We hebben de performantie van COMODO aangetoond door expressie-informatie te vergelijken tussen twee evolutionair ver verwijderde bacteriële modelsystemen *Escherichia coli* en *Bacillus subtilis*. In een later hoofdstuk hebben we COMODO uitgebreid voor de vergelijking van coexpressie-informatie tussen drie organismen waarbij we COMODO hebben gebruikt om coexpressie modules te zoeken die geconserveerd zijn in *E. coli*, *Salmonella enterica*, and *B. subtilis*.

In het tweede deel van de thesis, hebben we de relatie bestudeerd tussen het regulatorisch network en microarray expressie data in *E. coli*. Hiervoor hebben we een nieuwe netwerk gebaseerde similariteitsmaat voor coregulatie gedefinieerd op basis van de PageRank. In onze definitie komen regulators met een lagere PageRank overseen met meer locale regulators die meer bijdragen tot de totale coregulatorische similariteit tussen de targets. Genen met een hoge regulatorische similariteit op basis van de pagerank waren ook sterk coexpressed. Dit liet ons toe te besluiten dat het geobserveerd coexpressie gedrag (modulariteit in coexpressienetwerk) kan verklaard worden door een globaal network effect.

Bijkomend hebben we ook de relatie bestudeerd tussen het regulatorische netwerk en andere cellulaire interactienetwerken die geen regulerende functie hebben, zoals protein-protein

interactie- en metabole netwerken (EcoCyc). De gemiddelde coregulatorische similariteit (PageRank) voor genparen die behoren tot deze functionele netwerken was hoger dan verwacht op basis van een random associatie. Modules geïndentificeerd in deze functionele netwerken vertoonden ook een gemiddeld hogere coregulatorische similariteit dan verwacht op basis van random associatie. Ook werd de hiërarchie van de biologische modules zoals afgeleid op basis van onze netwerk gebaseerde regulatorische similariteit vergeleken met de functionele hierarchie gebruikt door GO. Deze vergelijking toonde aan dat beide hiërarchieën sterk gerelateerd zijn m.a.w. dat de functionele hiërarchie zoals gebruikt door GO, de regulatorische hiërarchie reflecteert. Deze observaties tonen aan dat gedurende evolutie het regulatorisch netwerk wellicht wijzigt om aanpassingen aan nieuwe situaties te accommoderen, maar dat wijzigingen onderheving zijn aan beperkingen opgelegd door de netwerkstructuur (zoals het behoud van functionele hiërarchie).

ABBREVIATIONS

BM	*Bacillus subtilis* module
cDNA	complementary DNA
ChIP	chromatin immunoprecipitation
ChIP-chip	chromatin immunoprecipitation (CHIP) on a microarray (chip)
ChIP-Seq	chromatin immunoprecipitation (CHIP) and sequencing
COG	clusters of orthologous groups of proteins
COLOMBOS	collection of microarrays for bacterial organisms
COMODO	conserved modules across organisms
DAG	directed acyclic graph
DNA	deoxyribonucleic acid
DBTBS	database of transcriptional regulation in *Bacillus subtilis*
EM	*Escherichia coli* module
FDR	false discovery rate
FFL	feed-forward loop
GO	Gene Ontology
ISA	iterative signature algorithm
MCL algorithm	Markov cluster algorithm
MCL	multi component loops
MIM	multi input motif
mRNA	messenger RNA
NCBI	national center for biotechnology information
OSLOM	order statistics local optimization method

SCSC	soft cross-species co-clustering
sRNA	small non-coding RNA
SIM	single input motif
SM	*Salmonella enterica* module
RNA	ribonucleic acid
rRNA	ribosomal RNA
TF	transcription factor
TFBS	transcription factor binding site
tRNA	transfer RNA

TABLE OF CONTENTS

Summary i

Samenvatting ..... **Error! Bookmark not defined.**

Abbreviations viii

Chapter 1 1

Introduction 1

1.1. Context of the thesis 1

1.1.1. Systems Biology: systematic approaches to study life 1

1.1.2. Comparative genomics 2

1.1.3. Gene expression compendia 3

1.1.4. Gene ontology terms 4

1.1.5. Physical interactions and cellular pathways 5

1.2. Objectives of the thesis 10

1.3. Overview of the thesis 12

Chapter 2 15

COMODO: an adaptive co-clustering strategy to identify conserved co-expression modules between organisms 15

2.1. Introduction 15

2.2. Materials and Methods 17

2.2.1. COMODO co-clustering procedure 17

2.2.2. Gene-gene threshold matrix 17

2.2.3. Selection of seed modules 18

2.2.4. Extension of seed modules 22

2.2.5. Chi-square test statistic as optimization criterium 23

2.2.6. Filter procedure 24

2.2.7. Application of the methodology to the E. coli and B. subtilis datasets 25

2.2.8. Condition selection for module visualization 25

2.2.9. Microarray compendia 25

2.2.10. Homology map and sequence similarity 25

2.2.11. Essential genes 26

2.2.12. Enrichment analysis of Gene Ontology terms, metabolic pathways, protein complexes, and regulatory data 26

2.2.13. Operon Information 27

2.3. Results 27

2.3.1. COMODO: a method to identify cross-species expression conservation 27

2.3.2. Identifying evolutionary conserved modules between E. coli and B. subtilis 28

2.3.3. Assessing the conservation of co-expression within homologous operons 30

2.3.4. Optimized co-expression threshold is module-dependent 31

2.3.5. Comparison with SCSC, a probabilistic co-clustering approach 34

2.3.6. Evolutionary conserved processes and essential genes 35

2.3.7. Regulation of evolutionary conserved modules 37

2.3.8. Differentiation in expression by divergence of regulation 40

2.3.9. Expression behavior of linker genes 41

2.3.10. Sensitivity towards the choice of the prespecified maximal co-expression stringency value 44

2.4. Discussion 47

Chapter 3 50

Extending COMODO to three organisms: application on *S. enterica* 50

3.1. Introduction 50

3.2. Materials and Methods 51

3.2.1. Statistics to assess co-expression conservation between two or three organisms 52

3.2.2. Application of the methodology to the E. coli, B. subtilis, and S. enterica datasets 55

3.3. Results 56

3.3.1. Identifying evolutionary conserved and non-conserved co-expressed modules between E. coli, B. subtilis, and S. enterica 56

3.3.2. Regulatory network conservation 57

3.3.3. Expression comparison of genes involved in quorum sensing and pathogenecity 60

3.4. Discussion 60

Chapter 4 63

Inferring co-regulated genes from regulatory network 63

4.1. Introduction 63

4.2. Materials and Methods 67

4.2.1. Regulatory network (Transcriptional and post-transcriptional interactions) 67

4.2.2. Co-expression microarray data 67

4.2.3. Monte Carlo sampling in regulatory network to assess the collaboration of regulators 68

4.2.4. PageRank value of regulators to assess importance of a regulator 69

4.2.5. Co-regulatory similarity measure between pair of genes and pair of modules based on PageRank similarity of common regulators 70

4.2.6. Finding modules in a network using OSLOM 70

4.3. Results 71

4.3.1. Detecting collaborative regulators 71

4.3.2. Co-regulatory similarity a measure to predict co-expression 73

4.4. Discussion 77

Chapter 5 79

The relation between physical interaction networks and functional data sources: application to the *E. coli* genome 79

5.1. Introduction 79

5.2. Materials and Methods 82

5.2.1. Current available physical interaction data sources in E. coli 82

5.2.2. Functional data sources 83

5.2.3. Jaccard similarity coefficient 83

5.2.4. Detecting modules in each physical interaction data source 84

5.2.5. Functional similarity measure between two modules 84

5.3. Results 85

5.3.1. Studying mutual relation between physical interaction data sources and functional data sources 85

5.3.2. Detecting modules of genes involved in the same biological processes 86

5.3.3. Exploring the mutual relation between genes involved in similar biological processes and the regulatory network 91

5.3.4. Comparing regulatory network hierarchy and GO terms hierarchy 95

5.4. Discussion 97

Chapter 6 102

Conclusions and perspectives 102

6.1. Conclusions 102

6.2. Perspectives 104

6.2.1. Data integration 104

6.2.2. Cross-species comparison 109

References 112

Supplementary Tables 120

Appendix A 121

CHAPTER 1
INTRODUCTION

1.1. CONTEXT OF THE THESIS

1.1.1. SYSTEMS BIOLOGY: SYSTEMATIC APPROACHES TO STUDY LIFE

Systems biology is the study of an organism with system point of view. Here, an organism is viewed as an integrated and interacting network of genes, RNAs, proteins, enzymes, and biochemical reactions which can sustain its life by interacting with its environment. Systems biology aims to describe the modular organization of an organism instead of analyzing individual components or aspects of the organism.

Systems Biology can be seen as a revolutionary approach to analyze biological complexity and biological systems function which reshaped the life sciences, and provided a deep understanding of DNA sequences, RNA synthesis, and the generation and interaction of proteins. During the past decades a tremendous evolution in molecular techniques allowed measuring the different biological components and their interactions on a genome-wide scale, giving rise to genome-wide data sets. Systems Biology approach to analyze the results derived from genome-scale experiments has accumulated vast amounts of data.

The complete sequencing of many genomes specially the human genome has ushered in a new era of systems biology referred to as omics. The English language neologism omics informally refers to a field of study in biology ending in *-omics*, such as genomics or proteomics (Wikipedia definition). The availability of omics data for various organisms has provided the opportunity to analyze conserved molecular mechanisms between different model organisms (Stuart, Segal et al. 2003; Lefebvre, Aude et al. 2005; Fierro, Vandenbussche et al. 2008; Chikina and Troyanskaya 2011).

The emergence of high-throughput technologies, such as genome sequencing technologies, microarray technology, Yeast two-hybrid screening, facilitate the vast growth in available omics

data. High-throughput technologies allow researchers to quickly conduct millions of biochemical, genetic or pharmacological tests. To understand the underlying biology of these data, systems biology is relying on an intimate integration of both mathematical and biological methods.

One major issue in systems biology is to develop proper data mining tools to integrate knowledge derived from various omics data. Because first, different omics data (e.g. genome sequence, transcriptome, proteome, interactome, metabolome) unveil distinct aspects of a cell as a biological system and integrating them leads to a more comprehensive insight into the cell life. Second, experimental and biological noise in the individual data measurements can be so prohibitive that each data type alone has a limited utility.

In addition to data integration, comparing the genomic properties across various species has revealed evolutionary and functional relations among different genes. The field comparative genomics was originally initiated to study of functional links and evolutionary relation mainly based on sequence similarity. Recent developments in data integration has made this field richer as coupling sequence similarity with other data sources provides more accurate source of information to study evolutionary and functional relations.

1.1.2. COMPARATIVE GENOMICS

The aim of comparative genomics is to study the relation of genome structure and function across different biological species or strains to shed light on evolutionary and functional conservation and divergence, and also to expand available knowledge from the well-studied organisms to the ones which this knowledge is limited. The study of functional links and evolutionary relation was accomplished mainly based on sequence similarity. Genes or proteins with high sequence similarity are called homologous. Homologous sequences are orthologous if they were separated by a speciation event: when an ancient species diverges into two separate species, the divergent copies of a single gene in the resulting species are said to be orthologous. Although this sequence-homology based prediction has been successful in practice it has certain drawbacks. For example, it may fail to predict the real orthologous gene pairs; Orthologous proteins with rather divergent sequences may be responsible for the same biological function. On

the other hand, two proteins with quite similar sequences may be involved in different biological processes or molecular functions(Lefebvre, Aude et al. 2005). In addition, the existence of the large number of homologous protein families to which the sequence-homology based prediction fails to ascribe a known function for any member is another major limitation (Karimpour-Fard, Detweiler et al. 2007; Chikina and Troyanskaya 2011).

Considering the mentioned problems, coupling other functional data sources is inevitable. Recently, there has been growing interest in utilizing co-expression data derived from different microarray experiments as another data source to predict functionally related genes among different organisms (Bergmann, Ihmels et al. 2004). Previous studies demonstrate that genes with similar functions are often co-expressed (Ihmels, Bergmann et al. 2005). In addition, revealing evolutionary conserved expression patterns has gained a lot of interest recently (Tirosh, Bilu et al. 2007; Chikina and Troyanskaya 2011). The next step in this field may accomplish by integrating physical interaction data to gain higher insight of conservation and divergence across different species or strains in the context of evolution.

1.1.3. GENE EXPRESSION COMPENDIA

Gene expression is the process by which information from a gene, which is typically a DNA strain, is used in the synthesis of a functional gene product. The first products of this process are RNA strains (e.g. mRNA, rRNA, tRNA, sRNA). Later, Messenger RNAs (mRNAs) can give rise to the proteins (see also central dogma of molecular biology in 1.1.5). The set of all RNA molecules in a cell at a certain stage or an environmental condition is referred to as the *transcriptome*. Revealing this transcriptome allows gaining insight into the functions of the individual genes and their interrelationships. Microarray technology has facilitated measuring the whole transcriptome on one chip.

Microarray experiments are made publicly available in specialized databases (Barrett, Troup et al. 2007; Demeter, Beauheim et al. 2007; Parkinson, Kapushesky et al. 2007). To fully exploit the large resource of information offered by the public databases, all the publicly available microarrays in one organism should be combined as large species-specific gene expression *compendia* (Figure 1.1). Compendia can be considered as a matrix containing the organism's

genes (rows) microarray expression values for all conditions (columns) in which microarrays were performed (Figure 1.1).

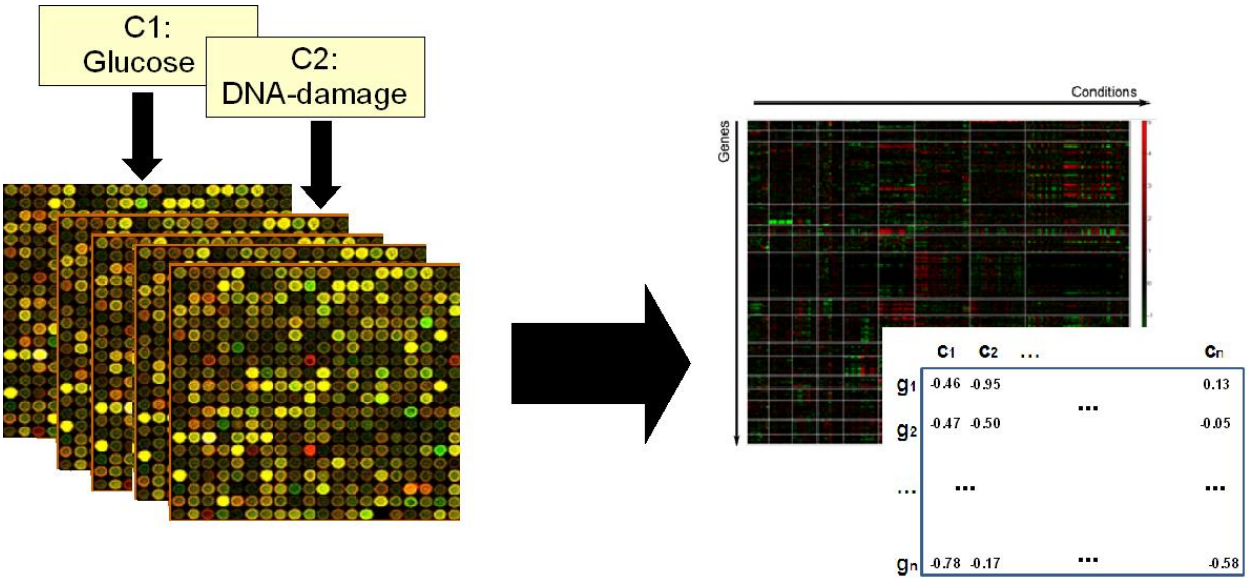


Figure 1.1. For an organism, gene expression compendia contain all publicly available microarray experiments measured at different stages or environmental conditions (left panel). It can be considered as a matrix containing the organism's genes (rows) microarray expression values for all conditions (columns) in which microarrays were performed (right panel).

Genes, which show similar expression pattern for a large set of conditions, are referred as co-expressed genes. As mentioned above co-expression is considered as a functional data and it was widely used to analyze functional relation in one organism (Ihmels, Bergmann et al. 2004) (Bergmann, Ihmels et al. 2003; Fadda, Fierro et al. 2009; Lemmens, De Bie et al. 2009) or across organisms (Ihmels, Bergmann et al. 2004; Ihmels, Bergmann et al. 2005; Zarrineh, Fierro et al. 2011).

1.1.4. GENE ONTOLOGY TERMS

Gene ontology (GO) terms are the most standard functional classes to validate the biological results, usually high-throughput experiments (Hu, Karp et al. 2009). Each GO term includes genes were annotated for a certain function. GO terms are divided to three main domains:

biological process, molecular function, and cellular component. Each domain of gene ontology is a tree-like directed acyclic graph (DAG) in which each node is a GO term and the direction of edges shows the parents GO terms. Thus, if a gene is assigned to a certain GO term, this gene should be assigned to all its parents GO terms. The problem with using GO terms is to detect the informative terms in the mentioned DAG as some GO terms contains hundreds of genes while some may contain only one gene. Here the main problem is how to deduce the functional relation between two genes or two clusters of genes considering the structure of GO terms DAG. One way is to reduce the GO terms just to a set of informative ones. As an example (Hu, Jiang et al. 2010) took just 32 informative GO terms from biological process domain, and they even removed any proteins with NCBI product descriptions as "hypothetical", "predicted" or "putative" to perform their analysis.

1.1.5. PHYSICAL INTERACTIONS AND CELLULAR PATHWAYS

According to central dogma of molecular biology, DNA information can be copied into mRNA, which is called transcription, and proteins can be synthesized using the information in mRNA as a template, which is called translation. The last products of this process, proteins, are known as the building blocks of cellular components and functions by forming protein complexes and enzymes. However, this is not the whole story, and a large network of physical interactions in all levels (e.g. protein-DNA, RNA-DNA, RNA-RNA, protein-protein, protein-compound) exists to control the transcription and translation process. This controlling system enables cells to sustain their life and react to their environmental perturbations. In addition, DNA information can be copied into non-coding RNA's such as transfer RNA (tRNA), ribosomal RNA (rRNA), and bacterial small RNAs (sRNA). These RNA's can carry on catalyst activities and controlling activity inside the cell.

In bacteria, gene expression is controlled by specific proteins (or protein complexes) called transcription factors (TFs), and also sigma and anti-sigma factors in transcription level. Gene expression can also be controlled by sRNA's in post-transcriptional level. Both mentioned protein-DNA transcriptional interactions and RNA-RNA post-transcriptional interactions can be inhibitory or activatory. Finally, some proteins like protein kinases can control other proteins in

post-translational phase through a phosphorylation activity. This kind of protein-protein interactions can also appear as both inhibitory and activatory.

New technologies have facilitated the prediction of massive number of physical interactions in different levels. To portrait the global mechanism of cell as a living system, these interactions have to be processed and integrated in a biological meaningful manner. In this way, network representations are the most natural and successful representation of physical interactions. A biological network is represented as an undirected graph like protein-protein network where nodes are proteins and edges are the interactions between them. A biological network can also be represented as directed graphs like regulatory network where in bacterial case the nodes are protein, sRNA, and genes and edges are regulatory relations between regulatory elements (protein, sRNA) and the targets genes. Phosphorylation network is another directed network where nodes are proteins and edges represent a kinase activity.

In addition to direct physical interactions, cellular pathways are also another available source of data that can be represented as a network. A cellular pathway is a chain of biological reactions to reform some initial compounds to the final compounds, and we call this chain of reactions a metabolic pathway if the final compounds which can be used by the cell, and in case that this chain of reactions convey a cellular signal in to a cell we call it signaling pathway. A cellular pathway is usually represented as a directed graph where the nodes are compounds and the edges are the reactions between them. A reaction can also be represented by the enzymes which catalyzed the reactions. As illustration, Figure 1.2 represents L-arabinose degradation I pathway in *E. coli* derived from EcoCyc (Keseler, Bonavides-Martinez et al. 2009).

To gain a global understanding of the mode of action in a cell (comprehensive mechanistic network), the network becomes an overlay of different individual networks and cellular pathways with nodes representing different molecular entities and edges different physical interactions or pathway directions. Here the problem is how to interpret this large and heterogeneous network. The simple solution is to restrict the nodes in the network to just genes, or both genes and proteins especially in eukaryotic cases where one gene can be translated to several different forms of proteins (alternative splicing) (Huang and Fraenkel 2009; Hyduke and Palsson 2010). The edges can also be restricted to actual physical interactions (Huang and

Fraenkel 2009; Hyduke and Palsson 2010) (e.g. Figure 1.3), or they may reflect the functional relations between genes derived from different physical interaction networks or cellular pathways (Myers, Chiriac et al. 2009; Narayanan, Vetta et al. 2010). In later case, the other functional data like co-expression data can also contribute to build the functional interaction network (Myers, Chiriac et al. 2009; Narayanan, Vetta et al. 2010). Naïve Bayesian approach is the most famous approach to build the functional interaction network (Myers, Chiriac et al. 2009).

Each biological interaction network displays a special topology which is evolutionary favorable due to the biological function of the network. The network, which gained the most attention from topological point of view in recent studies, is the regulatory network which evolves faster than other networks in the cell (Shou, Bhardwaj et al. 2011). The regulatory network consists of transcriptional, post-transcriptional, and post-translational interactions and controls the regulation of every cellular process. The highly repetitive topologies in the regulatory network are called regulatory motifs. In (Yu and Gerstein 2006), different regulatory motifs was highlighted such as single input motif (SIM), multi input motif (MIM), feed-forward loop (FFL), and multi component loops (MCL) (Figure 1.4). It has been shown that feed-forward loop is the most abundant circuit in a regulatory network (Babu, Teichmann et al. 2006; Yu and Gerstein 2006). Later in (Michoel, Joshi et al. 2011), the controlling motif circuits were expanded to protein-protein interactions and also post-translational interactions such as phosphorylation, which are highly abundant in higher evolutionary organisms. They showed that feed-forward loop is not only the favorable pattern in regulatory network consists of translational and post-translation interactions, but also it is highly abundant adding post-transcriptional interactions to the mentioned network in yeast (Michoel, Joshi et al. 2011).

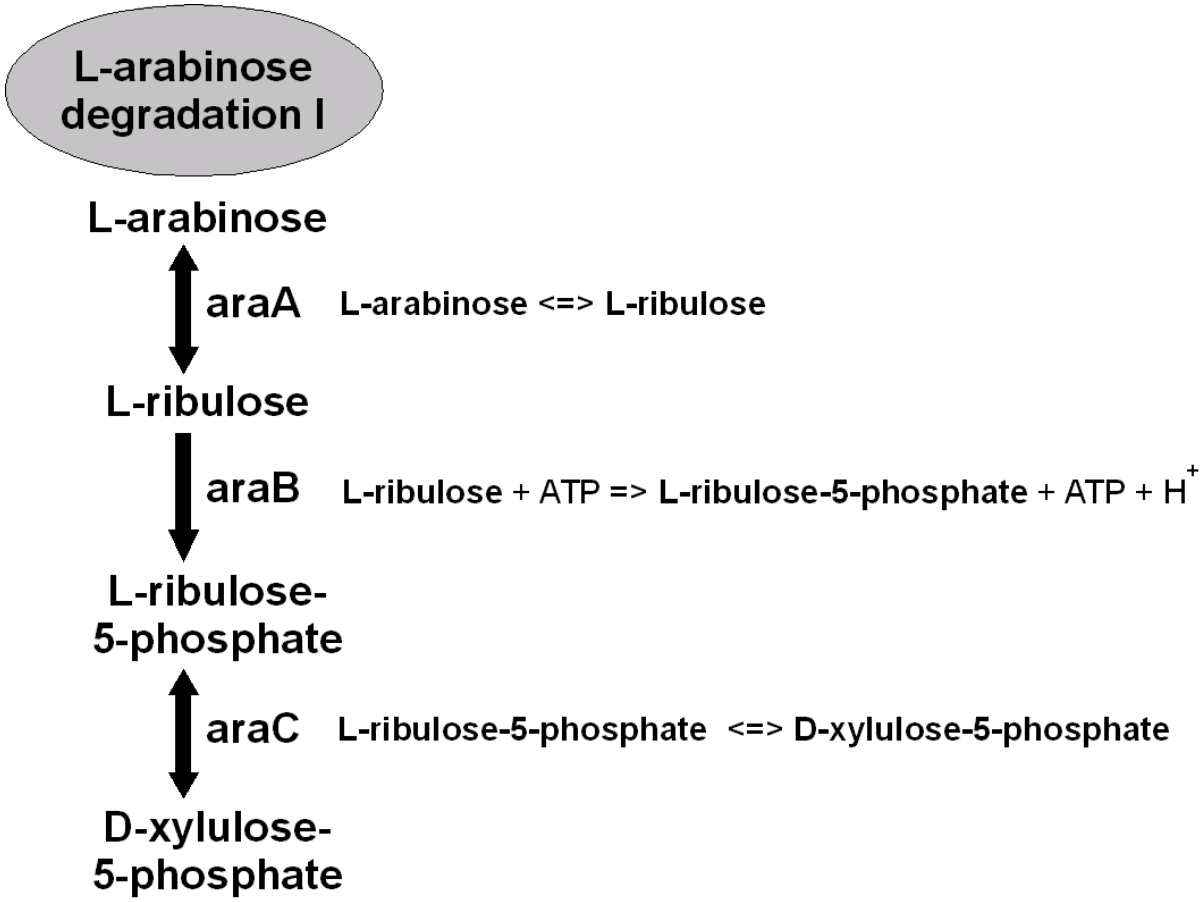


Figure 1.2. Network representation of L-arabinose degradation I pathway in *E. coli*. In this pathway compound L-arabinose is converted to another compound D-xylulose-5phosphate through three biological reactions. Three proteins/enzymes araA, araB, and araC catalyzed these reactions.

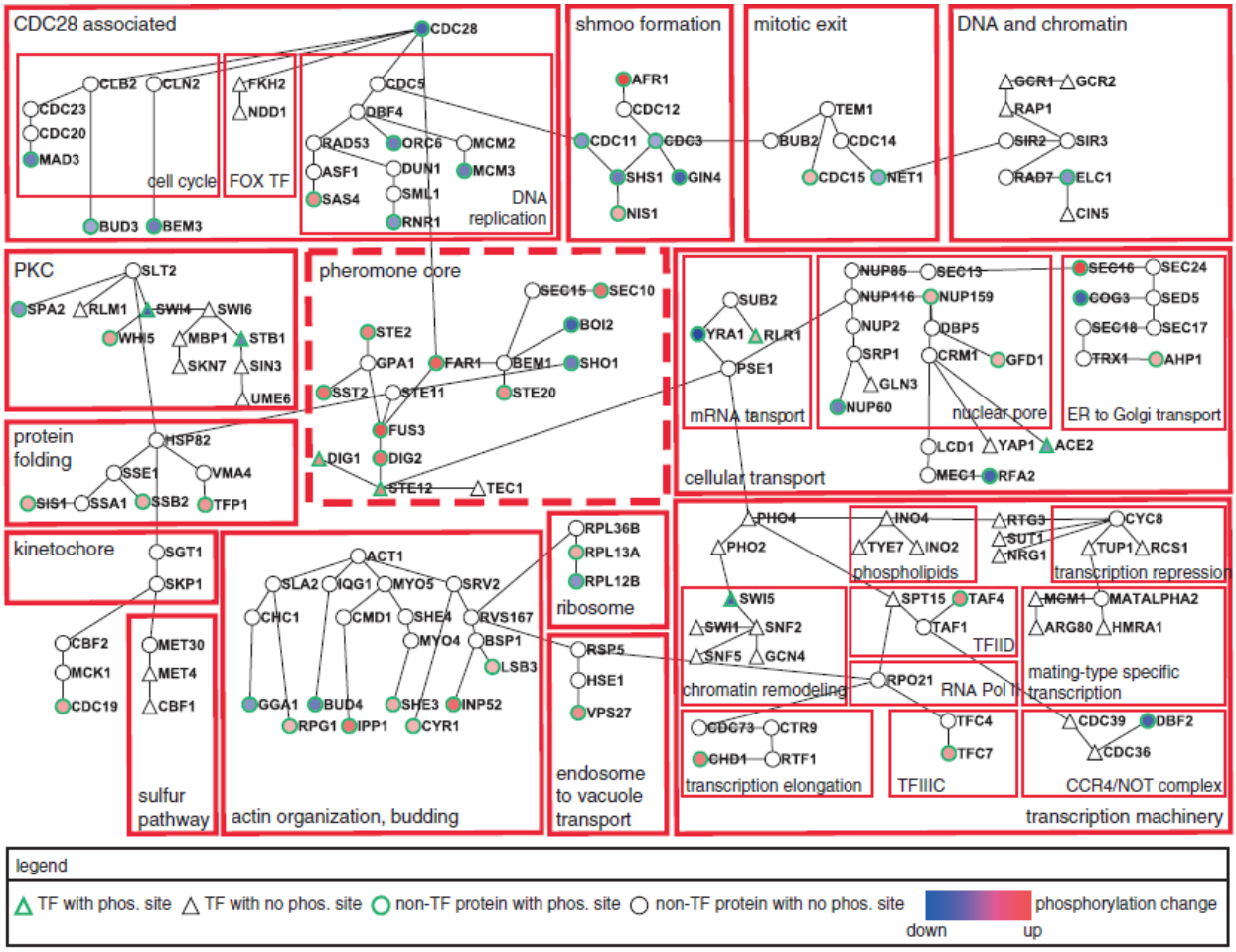


Figure 1.3. An example of comprehensive mechanistic network. pheromone response network in yeast. Here each module consists of different kind of interactions and different types of genes (TF and non-TF). The genes are categorized in different boxes based on their GO terms. Taken from (Huang and Fraenkel 2009).

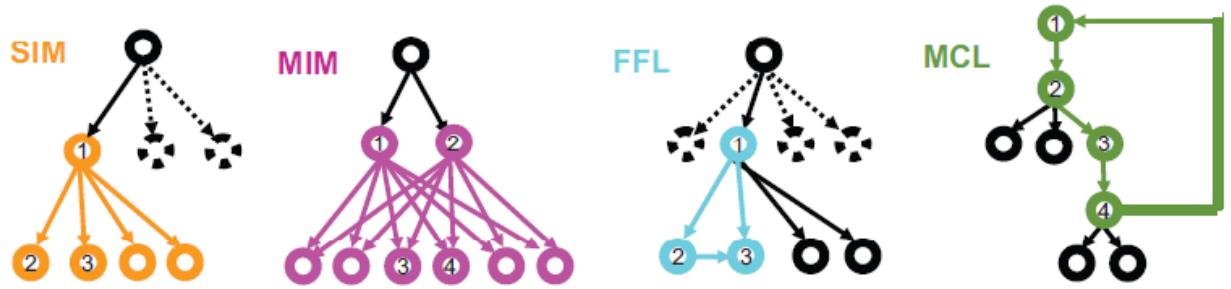


Figure 1.4. Illustration of regulatory network motifs. Four common network motifs in regulatory networks. Different colors represent different motifs. (*I*) Single-input motifs (SIM). For example, node 1 regulates nodes 2 and 3. (*I*) Multiple-input motifs (MIM). For example, nodes 1 and 2 are regulators, and nodes 3 and 4 are there common targets. (*III*) Feed-forward loop (FFL). For example, node 1 is the higher regulator in the hierarchy, and node 2 is its target while this node is a local regulator itself, and node 3 is the shared target of both regulatory. (*IV*) Multi-component loops (MCL). Node 1 is the higher regulator in the hierarchy, node 2 is target of node 1 while it regulates node 3. Node 4 is target of node 3, but it regulates node 1 on top of the hierarchy. Taken from (Yeger-Lotem, Riva et al. 2009).

1.2. OBJECTIVES OF THE THESIS

Availability of various genome-wide datasets provides the opportunity to study the whole genome behavior of the organisms as well as prediction of new functions for unknown genes. Integrating different types of data can lead to a better understanding of the cellular behavior and better functional annotation of genes (Kelley and Ideker 2005; Beyer, Bandyopadhyay et al. 2007; Huang and Fraenkel 2009). However, data generation efforts in bacteria have for a long time been lagging behind related efforts in yeast and other eukaryotic organisms. According to (Hu, Janga et al. 2009) one-third of the 4,225 protein-coding genes of the best studied bacterial strain, *Escherichia coli K-12*, remain functionally un-annotated (orphans). The number of annotated genes decline sharply for other bacteria. In addition, the annotated information regarding the physical interactions and cellular pathways is even more limited. This limitation is even more critical for the regulatory networks, especially for the less studied organisms. For example, 2697 transcriptional interactions and 203 small RNA interactions were annotated for

Escherichia coli K-12 in RegulonDB database (Gama-Castro, Jimenez-Jacinto et al. 2008), and the available data drops to 120 binding factors and 1475 gene regulatory relations for *Bacillus subtilis*, annotated in DBTBS (Sierro, Makita et al. 2008). Finally, there is no regulatory database for the other model bacteria *Salmonella enterica enteric*.

One way to overcome the data source limitation is to expand the information from the well-studied organisms to the ones that the available information was limited. Comparative genomics was the classical approach to expand information across organisms by considering sequence homology. Right now, fairly good functional annotation, operon prediction, and metabolic pathways, derived from genes sequence similarity, are available for different bacterial genome in BioCyc (Karp, Ouzounis et al. 2005). Co-expression similarity is another functional data which can easily be coupled to sequence data to enrich the accuracy of comparative genomics. Therefore, we developed new software, called COMODO (COnserved MODules across Organisms), to systematically integrate the sequence homology relations and co-expression relation derived from microarrays experiments. We demonstrated its performance using two distantly related model bacterial systems, *Escherichia coli* and *Bacillus subtilis*. As the results, we have shown the larger size of conserved co-expressed modules than previously predicted (Chapter 2). Later, we formalized the co-expression conservation for three organisms, and we demonstrate the efficiency of the cross-species co-expression comparison by studying the co-expression conservation as well as divergence of less studied model organism *Salmonella enterica enteric* in comparison to the other gram negative model organism *Escherichia coli* and the gram positive model organism *Bacillus subtilis* (Chapter 3).

Integrating various data sources is another way to overcome the data limitations. Integrating different data sources derived from high-throughput to assign new function to genes with unknown genes have been applied over different species especially *E. coli* (Andres Leon, Ezkurdia et al. 2009; Hu, Janga et al. 2009) and yeast (Zhu, Zhang et al. 2008; Myers, Chiriac et al. 2009; Narayanan, Vetta et al. 2010). Although current data integration methods based on network could predict new function for many genes of different genome successfully, still the mutual relation between physical interaction networks with controlling role inside the cell (the regulatory network) and other physical interaction networks and also other functional data

sources is not completely explored. For the first time, we could formulate the co-regulation of genes based on the regulatory network, and we have shown that our co-regulatory similarity measure is in line with the observed co-expression on the microarray compendia (Chapter 4). Later, we have shown the relation of the internal interactions responsible to assemble structural and functional components and cellular pathways with their regulatory program (Chapter 5). In addition, we could display the relation between functional hierarchy of genes and their regulatory hierarchy.

1.3. OVERVIEW OF THE THESIS

This section provides a chapter-by-chapter overview of the thesis (see Figure 1.5). The main topics of this thesis are cross-species co-expression comparison (Chapter 2 and 3) and the mutual relation between the regulatory network and other data sources (Chapter 4 and 5). In **Chapter 1**, comparative genomics is defined, and cross-species co-expression comparison is described as an improvement to the classical comparative genomics. In addition, gene expression compendium is introduces as a proper data set for cross-species co-express comparison. Furthermore, Different data sources like gene ontology term, as the most standard functional data source, and also physical interaction and cellular pathways are also introduced in this chapter. Finally, some basic properties of the regulatory network as the network with controlling role inside the cell like highly repetitive topologies, motifs, are also introduced in this chapter.

In **Chapter 2**, a new methodology for cross-species co-expression comparison, referred to as COMODO (COnserved MODules across Organisms) that uses an objective selection criterium to identify conserved expression modules between two species, is introduced. The method uses as input microarray data and a gene homology map and provides as output pairs of conserved modules and searches for the pair of modules for which the number of sharing homologs is statistically most significant relative to the size of the linked modules. To demonstrate its principle, we applied COMODO to study co-expression conservation between the two well studied bacteria *Escherichia coli* and *Bacillus subtilis*. The work in this chapter has been accepted for publication (Zarrineh, Fierro et al. 2011):

Zarrineh P., Fierro A. C., Sanchez-Rodriguez A., De Moor B., Engelen K., Marchal K. COMODO: an adaptive coclustering strategy to identify conserved Co-expression modules between organisms (2011). *Nucleic Acids Research*, 39 (7):e41.

In **Chapter 3**, the extended COMODO methodology is discussed. The extended COMODO can capture conservation across three species. The conservation and divergence inferred from extended COMODO methodology applied on three well studied bacteria *Escherichia coli*, *S. enterica*, and *Bacillus subtilis* is described in this chapter. Since regulatory network information does not exist in *S. enterica*, some possible regulatory interactions which can be deduced from co-expression conservation of target genes are also highlighted. The work presented in this chapter is still on-going:

Zarrineh P., Sanchez-Rodriguez A., Marchal K. Extending COMODO to three organisms: application on *S. enterica*. *In preparation*.

In **Chapter 4**, we introduce a new co-regulatory measure based on the regulatory network structure. To demonstrate its capabilities we applied this measure over *E. coli* regulatory network as the regulatory network of *E. coli* is one of the most complete regulatory networks. For the first time, we could show the co-regulatory measure is in agreement with the observed co-expression in microarray expression compendia. Using this co-regulatory measure in **Chapter 5**, we could project the regulatory network over physical interaction data including the protein-protein interaction network and the cellular metabolic and signaling pathways in *E. coli*. We could introduce a new species-specific functional similarity measure using GO terms in *E. coli*, and we could demonstrate the relation between regulatory program and hierarchy of functions in *E. coli*. The work presented in chapter 4 and 5 is an on-going collaboration research with institute for Cross-Disciplinary Physics and Complex Systems, in Palma de Mallorca:

Zarrineh P., Herrada A. C., Ramasco J. J., Eguiluz V. M., De Moor B., Marchal K. The mutual relation between the regulatory interaction network and other data sources: application to the *E.coli* genome. *In preparation*.

Finally, **Chapter 6** summarizes the results and provides a perspective on the future of both cross-species co-expression comparison and the mutual relation study between controlling interactions and other data sources. In this chapter we emphasize the co-regulatory similarity measure and the functional similarity measure derived from GO terms can be useful for data integration methods. As more data sources are becoming available in different organisms, both cross-species comparison and data integration fields can be enriched by new available data. These two fields are not completely independent, and the new progresses in data integration may be beneficial for cross-species comparison in near future.

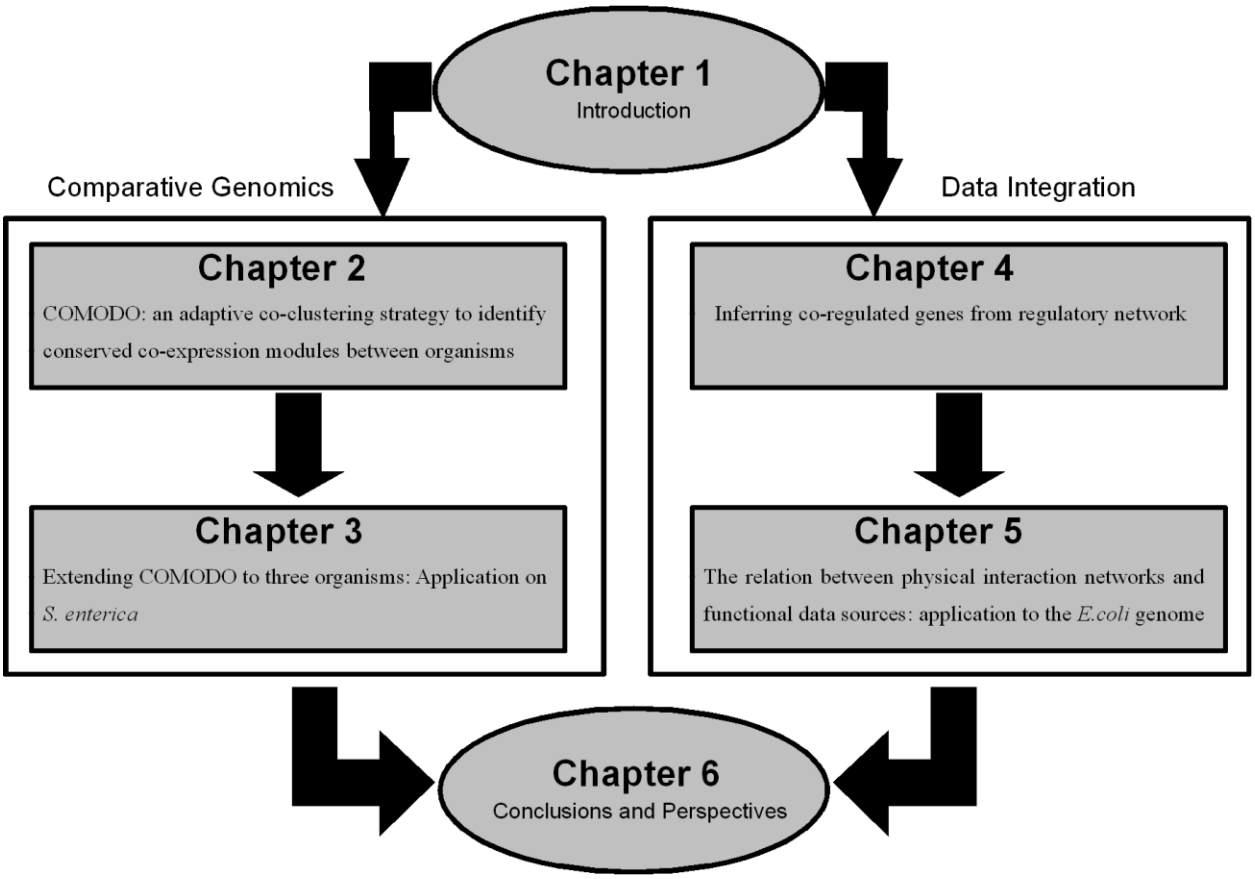


Figure 1.5. Overview structure PhD thesis. The thesis contains an introduction chapter (Chapter 1) and a conclusions and perspectives chapter (Chapter 6). The main part of the thesis consists of two parts, in the first part a new methodology is introduced for cross-species co-expression comparison (Chapter 2 and 3). In the second part, the mutual relation between the regulatory network and the other data sources, proper for data integration, is described in details (Chapter 4 and 5). This mutual relation study can be used for data integration (Chapter 6).

CHAPTER 2
COMODO: AN ADAPTIVE CO-CLUSTERING STRATEGY TO IDENTIFY CONSERVED CO-EXPRESSION MODULES BETWEEN ORGANISMS

2.1. INTRODUCTION

The availability of large scale expression compendia in combination with gene sequence conservation makes it possible to compare expression networks across organisms, in order to study their evolution or to identify functional counterparts in different species as homologs with 'conserved expression behavior' (Tirosh, Bilu et al. 2007; Fierro, Vandenbussche et al. 2008; Lu, Huggins et al. 2009). Besides custom made datasets that measure exactly the same experimental conditions in the different analyzed species (Lelandais, Tanty et al. 2008), also large heterogeneous compendia based on collecting publicly available expression datasets confer a useful resource for cross-species analysis of co-expression (Stuart, Segal et al. 2003; Bergmann, Ihmels et al. 2004). In contrast to the custom made homogeneous datasets, such heterogeneous expression compendia do not allow for a direct comparison of the expression patterns between orthologs in the different data sets, but instead rely on the search for 'conserved expression behavior'. With conserved expression behavior, we refer to the conservation of a mutual relation between genes across species (such as the conservation of the mutual correlation between the expression profiles of a pair of genes across species). This conserved behavior is usually derived by defining co-expression modules (i.e. genes sets that behave similarly in all or a subset of the conditions), inferred by either biclustering (searching for co-expressed gene sets) (Cai, Xie et al. ; Bergmann, Ihmels et al. 2004; Ihmels, Bergmann et al. 2005; Lu, He et al. 2007) or by the analysis of a co-expression network (a network constructed from the data where the nodes refer to the genes and the weighted edges to the degree of co-expression between the connected nodes) (Stuart, Segal et al. 2003; Lefebvre, Aude et al. 2005; Oldham, Horvath et al. 2006). These conserved modules are then compared across the species. Methods differ in the way they perform this module comparison. A first set of approaches starts from a reference species in which an initial set of modules is built (Bergmann, Ihmels et al. 2004; Ihmels, Bergmann et al. 2005; Oldham, Horvath et al. 2006; Lu, He et al. 2007). The corresponding homologous modules

are then identified in the target species by using gene homology. The approaches allow determining if the expression of a group of co-expressed genes in the reference organism is fully, partially, or not at all conserved at the level of co-expression in the target organism. To make an exhaustive comparison of all conserved modules between both species, each species has once to be used as a reference and once as a target. These approaches are most often applied using one-to-one gene homology relations (Ihmels, Bergmann et al. 2005; Lelandais, Tanty et al. 2008). A second set of approaches obviates the need of reference species: in the multi-species co-expression network proposed by Stuart *et al.* (Stuart, Segal et al. 2003), nodes correspond to genes that are conserved across the studied species (one-to-one map) and edges indicate significant pairwise co-expression levels between those genes in the different species. A clustering approach is used to identify conserved modules in this multi-species co-expression network. Alternatively, co-clustering strategies exploit homology and co-expression information to identify in both species simultaneously co-expression modules. Depending on the implementation results focus on modules containing only homologous genes that link up related modules (Lefebvre, Aude et al. 2005) or on finding mixed modules containing both homologous linker genes together with other genes that are co-expressed with those linker genes in a species specific way (Cai, Xie et al. 2010).

The difficulty with most previous methods is that they rely on the choice of a particular co-expression threshold or clustering parameter that determines the final module sizes (e.g. minimal degree of co-expression within a cluster or a minimal correlation coefficient to define subsets of co-expressed genes in a co-expression network, the number of clusters, etc.). However, choosing such parameter is not trivial as the definition of a relevant biological module is not a fixed one: different parameters can result in equally valid modules differing from each other in number of genes and/or conditions. Moreover, the relation between the degree of co-expression and a particular parameter or threshold usually is dataset-dependent (noise level, number of arrays tested, etc.) (Van den Bulcke, Lemmens et al. 2006). As it is hard to decide in advance on the most optimal co-expression threshold or parameter to define modules in each of the species-specific compendia and to decide upon the threshold or parameter combination that would allow for a proper cross-species comparison of modules, we developed a cross-species co-clustering approach referred to as COMODO (COnserved MODules across Organisms) that exploits

homology relations to determine the most optimal 'conserved co-expression modules' between two species (Zarrineh, Fierro et al. 2011). COMODO can take as input both one-to-one and many-to-many homology relations. The way we exploit the homology relations makes COMODO mainly suitable to search for processes with conserved co-expression behavior. Modules in a conserved pair are composed of homologous genes that share a mutual co-expression in each of the species, together with additional genes for which the co-expression with the homologous linker genes was found to be species-specific. We applied COMODO to search for conserved modules in two evolutionary distant prokaryotic model organisms: *Escherichia coli* and *Bacillus subtilis*. For those prokaryotic organisms we found conserved co-expression modules with a considerably larger fraction of genes than the number of conserved transcriptional units previously reported based on comparative genome analysis (Snel, van Noort et al. 2004; Okuda, Kawashima et al. 2007) and that cover a wider range of biological processes with conserved co-expression behavior than previously detected (Vazquez, Freyre-Gonzalez et al. 2009). Our results also showed how distantly related bacteria support the co-expression behavior of similar elementary processes with a completely different regulatory program. In chapter 3 we will formulize co-expression in more general way to extend COMODO to three organisms.

2.2. MATERIALS AND METHODS

2.2.1. COMODO CO-CLUSTERING PROCEDURE

An overview of COMODO is given in Figure 2.1 while in Figure 2.2 the detailed steps of the co-clustering procedure are displayed.

2.2.2. GENE-GENE THRESHOLD MATRIX

Conceptually all theoretically potential modules in each of the species can be represented as nested chains of partially overlapping modules that were obtained by gradually decreasing the threshold of the distance measure used by the clustering or distance approach (Figure 2.1). Biologically each chain of nested modules corresponds to the hierarchical organization of a

certain cellular processes (e.g. ranging from the production of an essential specific amino acid to a general response on a diauxic shift) (Bergmann, Ihmels et al. 2003). Different chains can share genes as the same genes can be involved in more processes. We used a symmetric gene-gene threshold matrix to concisely represent such chains of nested modules (Figure 2.1). Each axis of this matrix corresponds to the genes of one organism. The order of the genes in the X- and Y-axis of the matrix is determined by their assignment to modules under the most stringent tested threshold i.e. genes that are co-expressed at the most stringent tested threshold will be grouped. The values in the ith row and j^{th} column of the gene-gene threshold matrix represent the most stringent threshold at which respectively genes i and j appear together in at least in one of the detected modules. For the results shown in the main text the pairwise similarity between the genes was based on the Pearson correlation over all conditions in the compendium. The gene-gene threshold matrix in this case contains for each cell a discretized pairwise correlation value and the gene order on the X- and Y-axis of the gene-gene threshold matrix equals the order of the genes at the leaves of a hierarchical clustering applied on the non-discretized gene-gene correlation matrix. The number of bins used for the discretization depends on the parameter step size (see also below). We also built a gene-gene threshold matrix by using the gene thresholds defined by the iterative signature algorithm (ISA) to assign its genes to modules (Bergmann, Ihmels et al. 2003; Zarrineh, Fierro et al. 2011). In the latter case, the gene-gene threshold matrix consists of a compact representation of the overlapping clusters (module tree) that can be obtained using ISA with different threshold combinations. In our paper, we demonstrated the generality of the COMODO by analyzing results derived from ISA as a measure to build gene-gene threshold matrix (Zarrineh, Fierro et al. 2011), but in this chapter we will just focus on the results derived by using Pearson correlation across all conditions as co-expression measure since the quality of the results was much higher.

2.2.3. SELECTION OF SEED MODULES

To select the seed modules, we used the values on the first subdiagonal of the gene-gene threshold matrix (the first subdiagonal contains the values directly under those of the main diagonal of the gene-gene threshold matrix). To identify seeds we selected on this first subdiagonal groups of genes that were locally found to be more co-expressed with each other

than with their neighboring genes on the first subdiagonal (Figure 2.2A). For those genes the value on the first subdiagonal corresponds to the most stringent co-expression threshold at which they can be found together. To prevent that we would obtain many very small seed modules,

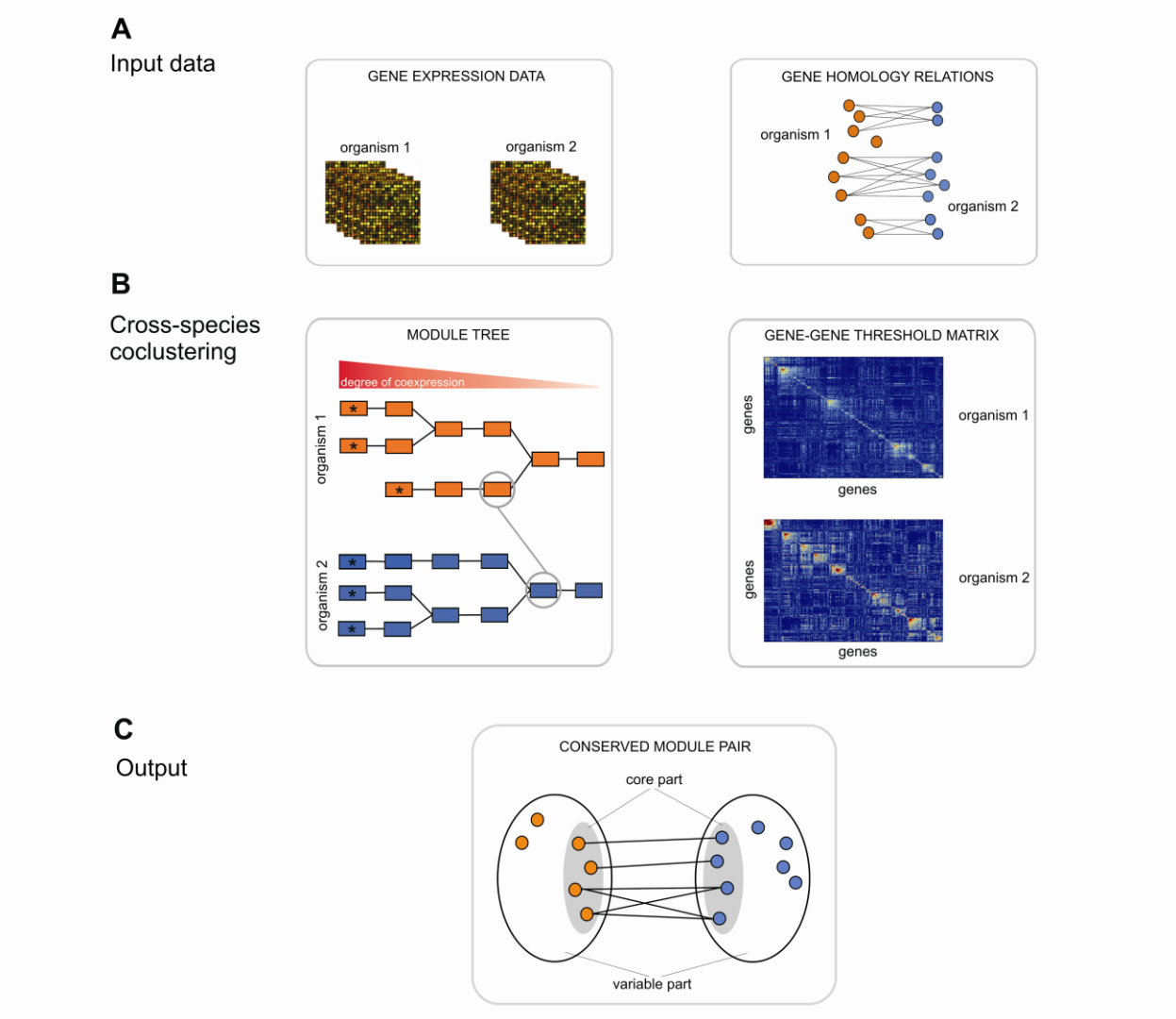


Figure 2.1. Detection of evolutionary conserved expression modules. A: Input data constitute of expression compendia of two distinct organisms (here *E. coli* and *B. subtilis*) (left panel) as well as a homology map between genes of the respective species (here derived from COG) (right panel). In the right panel, nodes correspond to genes and edges indicate the homology relations. **B:** The left panel schematically illustrates the concept of module trees. Conceptually all potential modules (indicated by rectangles) in each of the species can be represented as nested chains of partially overlapping modules that can theoretically be obtained by gradually decreasing the threshold that

determines the degree of co-expression within a module. Consecutive branches of the module trees give a view of all possible module sizes that originate from seed modules (modules indicated by a star correspond to modules obtained with the most stringent threshold). The chains of nested modules are captured by the symmetric gene-gene threshold matrices in each of the species (right panel). Our cross-species co-clustering procedure starts from tightly co-expressed seed modules (indicated by stars) and uses a bottom up approach to traverse these chains of nested modules in both species simultaneously to identify from all possible matching pairs the best matching one (here indicated by the modules connected by a gray line, best is defined based on the Chi-square test statistic). **C:** resulting matching module pairs are referred to as evolutionary conserved module pairs and consist of a core and a variable part.

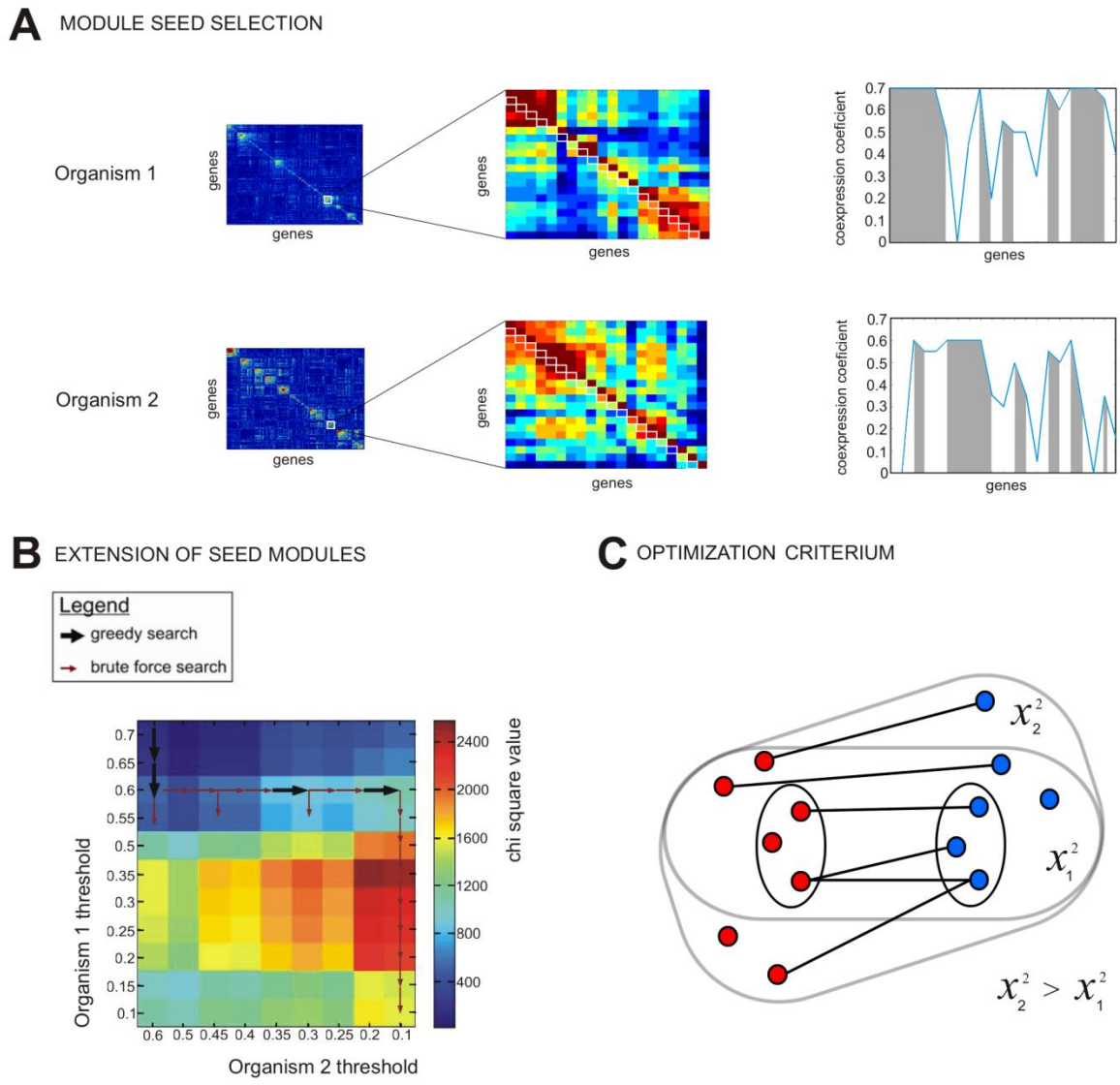


Figure 2.2. Cross-species co-clustering procedure. Displays the overall strategy of the co-clustering approach: first 'module seeds' are selected from the gene-gene threshold matrices in the respective organisms. Module seeds

linked by a sufficient number of homologous gene pairs are then gradually extended by traversing the space of possible cluster threshold combinations represented on the gene-gene threshold matrices of the respective species until optimality is reached. **A: Module seed selection step:** The left panel represents a zoom in on the gene-gene threshold matrices of respectively the first and second organisms. Values on the first subdiagonal of the gene-gene threshold matrix (indicated with white rectangles) are used to select the seed modules. The right panel displays the co-expression values corresponding to this first subdiagonal of the gene-gene threshold submatrices of respectively organisms 1 and organism 2. Groups of genes that are mutually more co-expressed than with any other genes on the first subdiagonal are selected as seeds (gray areas in the plot). To prevent that we would obtain many very small seed modules we set in the gene-gene threshold matrix all values larger than a prespecified maximal co-expression stringency value equal to this value. **B: Extension of seed modules step:** module seeds linked by a sufficient number of homologous gene pairs are gradually extended by traversing the space of possible cluster threshold combinations represented on the gene-gene threshold matrices in the respective organisms until optimality is reached. As it is computationally heavy to compare all possible threshold pairs, a combination of a greedy and brute force search was used to find the optimal module pair. This combination of a greedy and brute force search is represented as a dimensional grid of different threshold pairs, each with their corresponding chi-square values. The arrows indicate how the search space was traversed to find an optimal threshold pair. The search starts from the most stringent threshold pair (seed modules (top left)). Greedy (larger black arrows) and brute force (smaller red arrows) searches are called consecutively to evaluate different thresholds pairs in an efficient way. Plot of consecutive Chi-square values obtained along the search (i.e. for the different evaluated threshold pairs). **C: Optimization criterium:** a Pearson's chi-square test was used to assess the statistical significance of a module pairs i.e. to assess to what extent the number of linking and non-linking gene pairs between two modules differ from what is expected by chance.

containing two genes only, in the gene-gene threshold matrix all values larger than a prespecified maximal co-expression stringency value were set equal to this value. This guarantees a minimal number of genes to be present in the seed modules. We could show that within a certain range our co-clustering procedure is quite robust against the choice of this prespecified maximal co-expression stringency value (see 2.3.10).

2.2.4. EXTENSION OF SEED MODULES

COMODO uses a bottom up approach to build its conserved module pairs. It starts from the seed modules in each of the species of interest. Module seeds linked by a sufficient number of homologous gene pairs are gradually extended by traversing the space of possible cluster threshold combinations as represented on the gene-gene threshold matrices in the respective species until optimality is reached (see below for the chi-square optimization criterium). As it is computationally heavy to pairwisely compare all cluster threshold combinations between the two organisms we developed a dedicated search methodology. The search space of all possible combinations of thresholds can be represented in a two dimensional grid as shown in Figure 2.2B. Moving down the grid corresponds to gradually lowering the thresholds pairs. At each move the optimization criterium is evaluated. The parameter "Step" indicates the size by which the threshold is lowered at each move (in our experiments this was set to 0.05). To move along the grid we applied a combination of a greedy and brute force search. The methodology starts with the thresholds that define the seeds module pairs. By applying a greedy search gradually one or both of the thresholds in a combination are lowered until a local optimum is reached, i.e. further lowering the thresholds does not further improve the optimization criteria. To prevent the methodology from getting trapped in a local optimum, it searches further down in the grid in brute force manner until the stop criteria is reached (see below) to make sure no other threshold pair exists that is more optimal. If a better threshold pair than the current local optimum is found, the whole greedy search procedure is restarted from this more optimal threshold pair.

Two stop criteria are used: first, both thresholds should be larger than a preset value (in our example based on the Pearson correlation coefficient, both thresholds should at least be 0.1). Secondly, the minimal fraction of homologous versus non-homologous genes in the gene sets obtained by a given threshold pair should be higher than a preset number (in our study it was set to 0.1).

To tune the methodology for bacterial applications we introduced the following refinement procedure: genes that belong to the same operon tend to show a higher degree of co-expression with each other than with other genes. To prevent our methodology of getting biased towards finding module pairs that are composed of evolutionary conserved operons (these might always

get the highest chi-square value), we allowed for all module pairs of which one of the composing modules contains less than five genes the following additional threshold relaxations: the threshold of the group that contains less than five genes was relaxed until more genes were included. In such case, both the initially detected module pair and the module pair obtained after threshold relaxation were retained for further analysis.

The method can be applied on any chains of nested modules for which the relation between the modules is hierarchical, meaning that the module(s) obtained with the more stringent thresholds should be subsets of the ones obtained with a more relaxed threshold. Modules obtained with a more stringent threshold can never contain genes that were not detected at a more relaxed threshold.

2.2.5. CHI-SQUARE TEST STATISTIC AS OPTIMIZATION CRITERIUM

The definition of the best matching module pair is bound by the number of homologs that is shared by the selected modules in each of the species and corresponds to the pair for which the number of sharing homologs is statistically most significant relative to the size of the linked modules (Figure 2.2C). We used a Pearson's chi-square test to assess the statistical significance of a module pairs i.e. to assess to what extent the number of linking and non-linking gene pairs between two modules differ from what is expected by chance. To formulate the Pearson's chi-square test, consider *N1* genes in the genome of the first organism and *N2* genes in the genome of the second organism, and *M* linking homologous gene pairs derived from the COG database. If we pick two genes randomly, one from each organism, the probability that a homologous gene pair has been chosen is equal to $\frac{M}{N_1 \times N_2}$. Therefore, the probability that these genes are not homologous is $1-\left(\frac{M}{N_1 \times N_2}\right)$.

Given a pair of modules (one for each organism) containing respectively *g1* genes from the first organism and *g2* genes from the second one (where *g1* and *g2* << *N1* and *N2* respectively), the expected number of homologous gene pairs that would appear assuming that the two modules are randomly selected modules can be estimated by:

$$E_{homologous} = g_1 \times g_2 \times \left(\frac{M}{N_1 \times N_2} \right)$$

The expected number of non-homologous gene pairs appearing between them can be estimated by:

$$E_{non_homologous} = g_1 \times g_2 \times \left(1 - \left(\frac{M}{N_1 \times N_2} \right) \right)$$

We use the Pearson's chi-square test to assess whether the number of homologous and non-homologous gene pairs in an observed module pair is significantly different from the expected one. A chi-square test with one degree of freedom is as follow:

$$\chi^2 = \frac{O_{homologous} - E_{homologous}^{\ 2}}{E_{homologous}} + \frac{O_{non_homologous} - E_{non_homologous}^{\ 2}}{E_{non_homologous}}$$

Where *O* and *E* stand for observed and expected values respectively. Note that as the *p*-value might get very close to zero, we use an optimization criterium that maximizes the actual chi-square values instead of minimizing the corresponding *p*-values.

2.2.6. FILTER PROCEDURE

We selected from the raw output the most interesting module pairs for further analysis: we only retained the most significant module pairs (using a minimal threshold on the chi-square value). To remove redundancy we kept in case of overlapping module pairs (different module pairs that share 75% of homologous linker genes) the one with higher chi-square value.

We included the following additional criteria for our specific application: modules of size smaller that six should be linked up to their counterpart modules in the other organism with at least two homologous linker genes, this to avoid small spuriously linked modules. In addition, we required that the number of linker genes comprises at least 20% of the total number of genes in each of the modules to prevent unbalanced growth of one module compared to its counterpart module as the latter modules were very often found not to be biologically meaningful.

2.2.7. APPLICATION OF THE METHODOLOGY TO THE E. COLI AND B. SUBTILIS DATASETS

Using a the Pearson correlation over all conditions as a distance measure and a prespecified maximal co-expression stringency value of 0.7 for seed identification, we obtained conserved module pairs covering 1687 *E. coli* genes and 2129 *B. subtilis* genes. After filtering (using a chi-square threshold of 470) and removing overlapping module pairs (see above), we retained 445 *E. coli* genes and 481 *B. subtilis* genes being found in 82 non-redundant module pairs. The final 82 conserved module pairs were ordered according to their overlap in gene number in each of the organisms. Modules that shared more than 30% of their genes were assigned to the same biological process as they were enriched in the same GO categories and pathways. To assess the False Discovery Rate (FDR) of our results we randomized the expression values in the original compendia and searched for conserved module pairs using the same procedure as described above (process was repeated 50 times, expression data was randomized by reassigning the gene labels to the expression profiles).

2.2.8. CONDITION SELECTION FOR MODULE VISUALIZATION

For visualization purposes heat maps only display the conditions for which the co-expression behavior was most obvious. Relevant conditions were selected by dividing per condition the mean value of the expression levels in the module by the variance (coefficient of variation). If this coefficient of variation exceeds a predefined threshold (1 in our case), the corresponding condition is visualized.

2.2.9. MICROARRAY COMPENDIA

The microarray compendium of *E. coli* was obtained from Lemmens *et al.* (Lemmens, De Bie et al. 2009) and the one of *B. subtilis* from (Yu and Gerstein 2006) *et al*. (Fadda, Fierro et al. 2009). They contained respectively 870 conditions for *E. coli* and 231 for *B. subtilis*.

2.2.10. HOMOLOGY MAP AND SEQUENCE SIMILARITY

A total of 5459 homologous gene pairs between *E. coli* and *B. subtilis* were annotated based on the COG database (Tatusov, Koonin et al. 1997). This many-to-many COG map was used

throughout the thesis unless specified otherwise. Orthologous gene pairs between *E. coli* and *B. subtilis* were when needed identified by the Reciprocal Smallest Distance approach (Wall and Deluca 2007).

2.2.11. ESSENTIAL GENES

Essential genes in *B. subtilis* and *E. coli* were downloaded from DEG, a database of essential genes (Zhang, Ou et al. 2004; Zhang and Lin 2009). This database contains 271 essential genes in *B. subtilis*, resulting from a single gene deletion experiment (Kobayashi, Ehrlich et al. 2003). For *E. coli* 620 genes were originally determined to be essential based on genetic footprinting (Gerdes, Scholle et al. 2003) and 303 genes were later identified by single gene deletions (Baba, Ara et al. 2006). As 205 genes were found in common between those two *E. coli* lists, we obtained in total 712 essential genes for *E. coli*. Based on the homology relation derived from the COG database, we found 209 homologous pairs of essential genes comprising 191 *B. subtilis* and 195 *E. coli* essential genes. From these 195 *E. coli* genes with homologs in *B. subtilis* 164 were originally identified by the single gene deletion experiment mentioned above (Baba, Ara et al. 2006).

2.2.12. ENRICHMENT ANALYSIS OF GENE ONTOLOGY TERMS, METABOLIC PATHWAYS, PROTEIN COMPLEXES, AND REGULATORY DATA

Gene Ontology (GO) terms, metabolic pathways, and protein complexes of *E. coli* were downloaded from EcoCyc (Keseler, Bonavides-Martinez et al. 2009). GO terms for *B. subtilis* were downloaded from the Comprehensive Microbial Resource (CMR) (Peterson, Umayam et al. 2001). Metabolic pathways and protein complexes of *B. subtilis* were obtained from BioCyc (Karp, Ouzounis et al. 2005). Transcriptional interactions were downloaded from RegulonDB (Gama-Castro, Jimenez-Jacinto et al. 2008) and DBTBS (Sierro, Makita et al. 2008) for *E. coli* and *B. subtilis* respectively. Enrichment analysis was done based on the hypergeometric distribution corrected for multiple testing by the False Discovery Rate (FDR) (Benjamini and Hochberg 1995).

2.2.13. OPERON INFORMATION

Operon structure was derived from RegulonDB (Gama-Castro, Jimenez-Jacinto et al. 2008) and DBTBS (Sierro, Makita et al. 2008) for respectively *E. coli* and *B. subtilis*. As DBTBS only describes experimentally validated *B. subtilis* operons, we used for gene sets not covered by DBTBS the following databases with operon predictions*:* OpeRons (DOOR, http://csbl1.bmb.uga.edu/OperonDB/) (Mao, Dam et al. 2009) and http://www.microbesonline.org/operons/OperonList.html (Price, Huang et al. 2005). Operon predictions were retained: 1) if databases agree with each other in predicting the same operon structure (this was the most frequent situation), 2) if they were only predicted by one database, 3) in the few cases where two databases predicted a different operon structure, we used the structure that was more compatible with our expression results or with the structure of the counterpart operon in *E. coli*. To identify conserved operons (or homologous operons) between *E. coli* and *B. subtilis* we used the following definition: we started from the list of *E. coli* operons as this was the best annotated. We identified as an operon conserved between *E. coli* and *B. subtilis* any annotated operon in *B. subtilis* for which at least two genes showed homology, based on COG database information. This analysis was also repeated using only strict homology links obtained by the Reciprocal Smallest Distance approach (Wall and Deluca 2007) to approximate a definition of 'orthologous' operons.

2.3. RESULTS

2.3.1. COMODO: A METHOD TO IDENTIFY CROSS-SPECIES EXPRESSION CONSERVATION

As we focused on searching processes across species with evolutionary conserved co-expression behavior, we defined the optimal size of the modules in each of the species as the one that maximizes the fraction of homologous genes that links up both modules in an evolutionary conserved module pair. An overview of the analysis flow is given in Figure 2.1. To avoid (bi)clustering the datasets using a fixed parameter setting that determines the cluster size in each of the species separately, we relied on a bottom up co-clustering approach to build the modules. COMODO is initialized with co-expressed seeds or seed modules obtained in each of the

species. These seeds are gradually expanded in each of the species until a pair of modules is obtained for which the number of shared homologs is statistically optimal relative to the size of the linked modules. The optimization criterium is based on a chi-square statistic (see 2.2.5). Our co-clustering procedure that extends the seed modules until optimality is reached is based on greedy and brute force procedure described in 2.2.4.

Eventually pairs of evolutionary conserved modules are obtained, each containing a core and a variable part (Figure 2.1C). The core part consists of the homologous genes that link up both co-expression modules and for which the mutual co-expression behavior is conserved. The variable part contains the additional genes that belong to the composing modules of a given pair in either one of the organisms. These are the genes that either do not have a homologous counterpart in the other organism or that acquired a co-expression behavior similar to that of the core part in only one of two species (Perez and Groisman 2009). Because a module in one species can be linked to more counterparts in the other species (Figure 2.3), COMODO can be used to study both conservation, but also divergence in expression which makes it optimally suited to be used with a many-to-many homology map.

2.3.2. IDENTIFYING EVOLUTIONARY CONSERVED MODULES BETWEEN E. COLI AND B. SUBTILIS

We applied our methodology to study the degree to which co-expression modules have been conserved between two bacterial model organisms *E. coli* and *B. subtilis*. For both species we used cross-platform microarray compendia covering a wide range of experimental conditions (see 2.2.9). Many-to-many homology relations amongst the genes of the two species were defined based on COG (Tatusov, Koonin et al. 1997). Applying our method resulted in the identification of 82 conserved module pairs in *E. coli* and *B. subtilis* that were linked through a statistically significant set of homologous genes. These linked groups are called matching module pairs and they represent processes for which the co-expression is at least partially conserved over the wide evolutionary distance that separates *E. coli* from *B. subtili*s. Figure 2.3 gives an overview of these matching, evolutionary conserved module pairs.

To estimate the potential number of false positives amongst our detected conserved module pairs, we applied COMODO to a randomized dataset from which we did not expect to find any meaningful results (see 2.2.7). The False Discovery Rate (FDR) estimated as the mean number of significantly detected matching module pairs in random expression compendia was 2.24. In general the chi-square statistic values obtained in the randomized datasets were well below the ones observed for the true dataset (t-test, $P < 0.05$), implying that the size of the core to the variable part is much larger in modules obtained from the true dataset than in those obtained from the random dataset.

In those 82 conserved module pairs, on average 60% of the genes constitute the core part and 40% the variable part. Of those genes in the variable part, 33% did not have a homologous counterpart in the organism of comparison. The other 67% found in the variable part with a homologous counterpart in the other organism could correspond to species-specific members of the regulon represented by the core part. A gene assigned to the core part of one module can also be found in the variable part of another module as the same gene can belong to different regulons that do not completely coincide between species. For instance, EM40-BM40 contains in its core part the orthologous operon *nrdEFIH* known to be regulated both in *E. coli* and *B. subtilis* by NrdR (Hartig, Hartmann et al. 2006; Torrents, Grinberg et al. 2007). In EM59-BM59, containing a Fur-dependent conserved core the same *nrdEFIH* is in the variable part of the *E. coli* module. This confirms previous knowledge on *nrdEFIH* being Fur-dependent in *E. coli*, but not in *B. subtilis* (at least not yet observed) (Hartig, Hartmann et al. 2006).

By using a stringent filtering procedure and only maintaining matching module pairs for which the core part was relatively larger compared to the variable part, we focused on the processes for which co-expression behavior was conserved between *E. coli* and *B. subtilis*. The number of genes in the evolutionary conserved modules varies largely and ranges between 2 to 100, with a large overrepresentation of small modules (e.g. 28 module pairs containing 2 to 5 genes only in both matched modules). Smaller modules usually correspond to single operons, subunits of a protein complex or constitute parts of larger biosynthetic pathways. As the size of the conserved modules increases, the modules cover larger pathways. A complete description of the modules can be found in the Table S2.1.

In total 30 of our 82 conserved modules are linked by a single homologous operon: 14 of those by 'an orthologous operon' (according to the definition of operon orthology described in 2.2.13) and 16 by a homologous, but functionally related operon. As by definition genes within an operon (as an estimate of a transcription unit) will be co-expressed, these matching modules, although correctly identified by our method do not contribute more information on the functional relation between the matching operons than the one derived from sequence analysis. Therefore, extrapolation of the operon function between *E. coli* and *B. subtilis* should be treated with care for those modules.

2.3.3. ASSESSING THE CONSERVATION OF CO-EXPRESSION WITHIN HOMOLOGOUS OPERONS

As the operon structure is an important mechanism to guarantee the conservation of co-expression behavior between genes within a species (Snel, van Noort et al. 2004; Okuda, Kawashima et al. 2005; Okuda, Kawashima et al. 2007), we wanted to assess as a validation of our methodology to what extent homologous operons will be found in the core parts of our conserved modules. When using the COG based definition of homologous operons, we could retrieve 289 pairs of *E. coli* and *B. subtilis* operons that share at least 2 homologous genes (see 2.2.13). Based on sequence homology several *E. coli* operons were mapped to at least two different operons in *B. subtilis* that mutually do not share any gene (this was also observed when the comparison was performed the opposite way around i.e. when *B. subtilis* operons were mapped to those of *E. coli*). Of these 289 *E. coli* operons with a homologous counterpart in *B. subtilis*, 91 were found as linkers between conserved modules (i.e. 31% recovery rate), resulting in a total of 135 links between conserved modules as some operons can occur in more modules. Of those 135 linking operons, in 61 cases all the genes of the linking operons were found in the core part, in 33 cases one of the operon genes of the linking operons was missing and in 41 cases at least two genes of the linking operons were missing from the core part. Although in some cases lacking some of the operon genes in the core part might point towards differentiation in regulation, for instance, by means of intra-operonic promoters, it seemed that in many cases it was the last operon gene that was no longer found to be co-expressed with the rest of the operon genes in the core parts of the linking modules. This observation can be explained by the

increased degradation of the mRNA at the 3' end of the transcript (Grunberg-Manago 1999). When using a more stringent definition of homologous operons (see 2.2.13), the fraction of *E. coli* operons with a counterpart in *B. subtilis* that was found in the core part (meaning that their genes were found as linker genes in conserved modules) was much higher (50 of the 100 linking operons, i.e. 50% recovery rate). This higher recovery rate might be partially due to the fact that this more strict mapping as an estimate of orthologous operons results in linking operons that mutually share more genes than with the previously used COG-based mapping (as an estimate of homologous operons). When more genes are shared between the linking operons, the chance to find a module pair that meets our selection criteria (sharing at least two co-expressed linker genes) can be met more easily. On the other hand, it definitely also reflects that many of the operons that can be linked through a COG mapping are not each other's functional counterparts.

We also found that a considerable part of the orthologous operons could not be retrieved in conserved co-expression modules as their composing genes were not found to be co-expressed in *E. coli* or *B. subtilis*, probably due to the still incomplete sampling of conditions in the used expression compendia (this was the case for 50 operon pairs defined using the more stringent definition and for 198 of the conserved operon pairs defined with the less stringent definition). So the 50% recovery rate of orthologous operons (as well as the 31% recovery rate for the homologous operons) in our module cores stems from the incompleteness of the used expression compendia rather than from a bias in the methodology.

2.3.4. Optimized co-expression threshold is module-dependent

Maximizing the statistical significance of the number of linking homologs in the core of a conserved module pair relative to the module sizes in each of the respective species allows us to select in each of the species the modules that best match the conserved processes reflected by the core. Depending on the type of biological process that is conserved in the core the optimal correlation thresholds for the modules in each of the individual species can differ considerably. This is illustrated in Figure 2.4 where the selected correlation coefficient differs largely between the modules of the different conserved pairs.

Globally the correlation thresholds for the *E. coli* modules were lower than those of the corresponding *B. subtilis* modules, most probably because the *E. coli* compendium is larger and contains more conditions than the one of *B. subtilis*. When investigating per organism the relation between the used correlation threshold and the number of genes in a selected module, we observed that it is not only the number of genes within a module that determines the selected correlation threshold, but that there is also a clear influence of the type of process the module reflects (Figure 2.4). Housekeeping processes such as ribosomal metabolism and translation (EM34_35_36-BM34_35_36) were found with very strict thresholds despite containing a relatively high number of genes, while for more specialized processes such as e.g. iron acquisition (EM59-BM52_53_59) and motility and flagella synthesis (EM32-BM32) the opposite was observed (Figure 2.4). This can be related to the number of compendium conditions in which genes are expected to be co-expressed. When using a distance measure that by default considers all conditions (such as Pearson correlation), genes that tend to be active under all conditions (e.g. housekeeping genes) will be found co-expressed with a more stringent correlation threshold than genes that are only co-expressed under a subset of the sampled conditions (e.g. those that belong to the more specialized modules). This observation underlines the need for a module- and dataset-dependent determination of the co-expression threshold or clustering parameter that determines the final module sizes during the co-clustering of heterogeneous expression compendia.

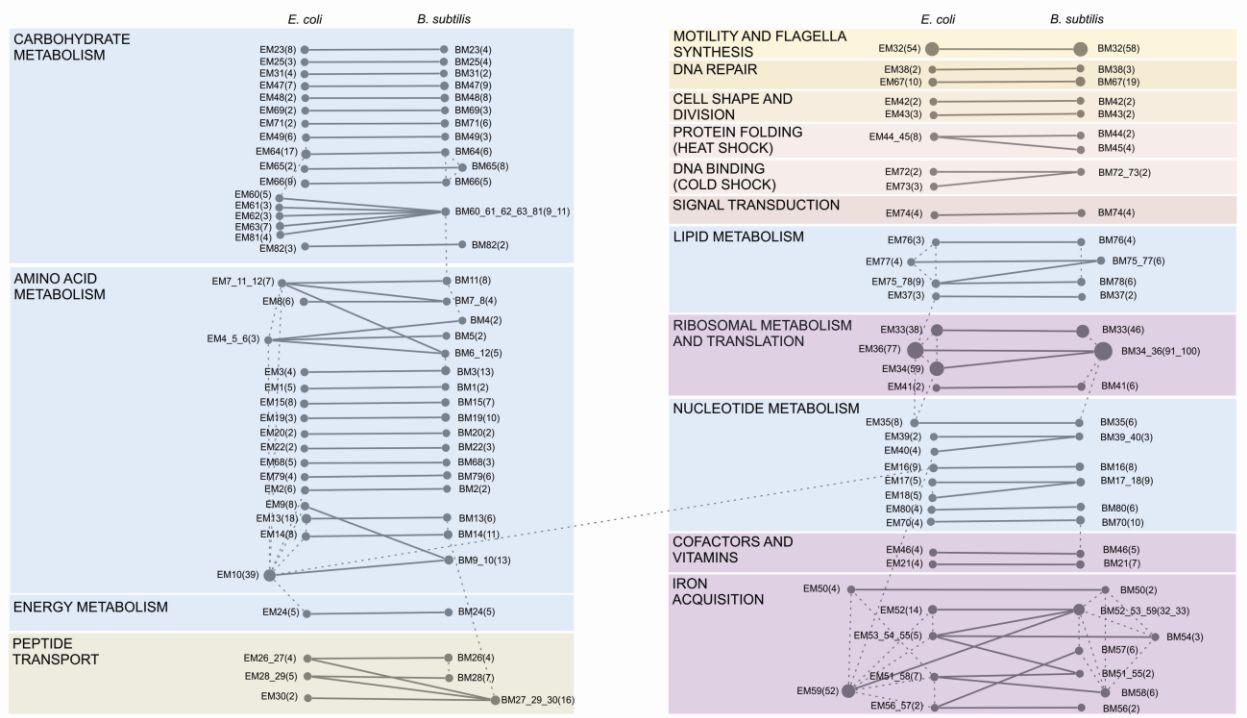


Figure 2.3. Overview of evolutionary conserved modules between *E. coli* and *B. subtilis*. A total of 82 evolutionary conserved module pairs of which the matching modules (connected by solid lines) were linked through a statistically significant set of homologs between *E. coli* and *B. subtilis* are shown. Node sizes are proportional to the number of co-expressed genes in the modules (indicated in parenthesis) and module ids correspond to those used in Table S2.1. Modules showing an overlap of 30% to 75% of their genes within each species were connected by dashed lines. Modules that show an overlap of at least 75% in their gene content were merged. Modules to which a similar functional category was assigned were grouped (as indicated by the different panels. Panels with the same color are involved in a similar general process e.g. metabolism).

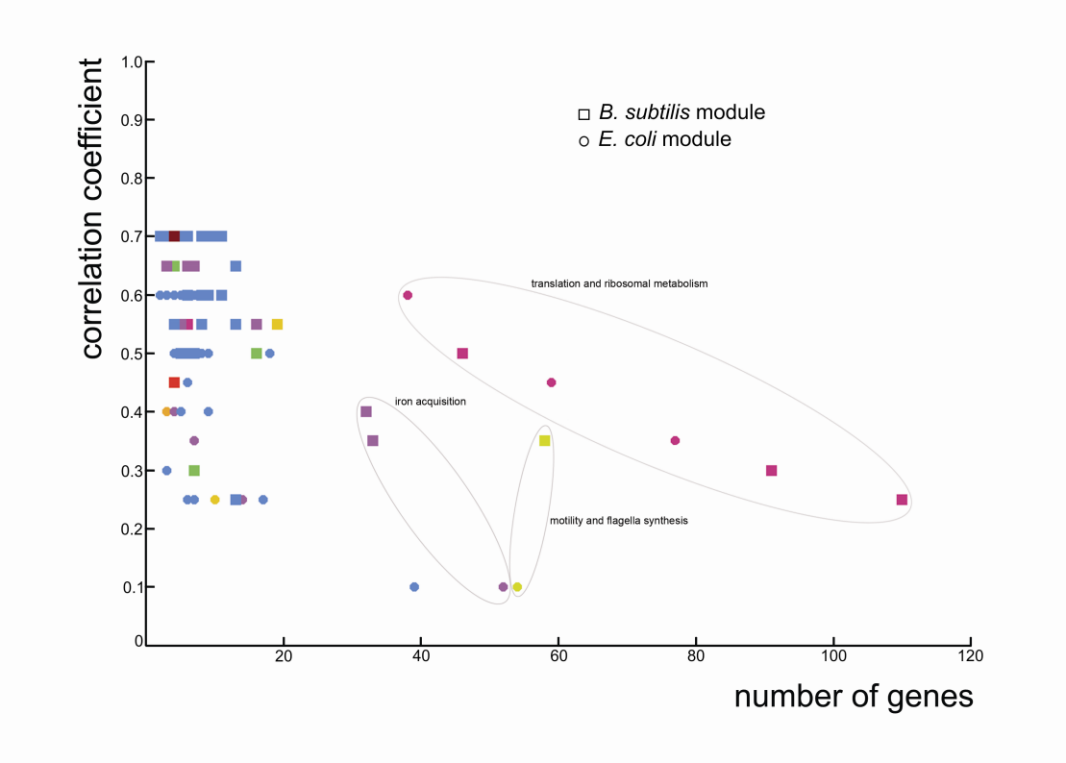


Figure 2.4. Degree of correlation within a co-expression module versus the number of genes it contains. Number of genes: refers to the total number of genes in the module (adding up genes in core and variable parts). A total of 82 evolutionary conserved modules between *E. coli* (circles) and *B. subtilis* (squares) are plotted. In each case the color used to represent a module corresponds to the color scheme in Figure 2.3 to denote the functional class (or group of related functional classes) a module was assigned to.

2.3.5. COMPARISON WITH SCSC, A PROBABILISTIC CO-CLUSTERING APPROACH

We compared the performance of COMODO with the recently developed co-clustering approach SCSC of which the implementation is publicly available (Cai, Xie et al. 2010). We observed that the intrinsically different way in which the co-clustering is performed by respectively SCSC and COMODO affects the characteristics of the detected matching module pairs. SCSC partitions the data in each species in a predefined number of modules (Zarrineh, Fierro et al.). This results in sets of loosely connected modules of which the sizes and co-expression level largely depend on the used dataset pre-filtering and the predefined cluster number. In addition, there is no guarantee that the homologous genes that were added to the modules are amongst the most tightly co-expressed genes in a module. This, in combination with the fact that the identified modules

should only be loosely connected by homologs to be identified as a matching pair complicates distinguishing true matching module pairs from spurious matching ones when using SCSC in combination with a many-to-many homology map.

COMODO in contrast chooses the number of genes in the modules of either species to maximize the enrichment of linking homologs relative to the number of variable genes. This criterium results in adapting the size of the modules to the specificities of the conserved processes: as a result COMODO can cover a wide range of module sizes without compromising the quality of the modules (reflected by a good co-expression level). In addition homologous linker genes are by definition as tightly co-expressed as the rest of the genes in a module. This, together with a selection of the most significantly matching module pairs based on the chi-square statistic facilitates prioritizing the most significant matching module pairs. However, because of this bottom up strategy COMODO might unlike SCSC underestimate in the individual species the true sizes of the pathways represented by the cores.

2.3.6. EVOLUTIONARY CONSERVED PROCESSES AND ESSENTIAL GENES

Figure 2.3 gives an overview of the evolutionary conserved modules ordered within a species according to their overlap in genes. Partially overlapping modules (indicated by dashed lines) were assigned to the same functional category. Biological processes involved in carbohydrate metabolism, amino-acid metabolism, energy metabolism, nucleotide metabolism, lipid metabolism, translation and ribosomal metabolism, motility and flagella synthesis, DNA repair, cell shape and division, protein folding (heat shock), DNA-binding (cold shock), signal transduction, cofactors and vitamins, and iron acquisition all contain genes for which the mutual co-expression behavior was found to be conserved between *E. coli* and *B. subtilis*. As most of these processes with a conserved co-expression behavior are primary processes, we wondered to what extent they contained essential genes, defined as the minimal gene sets required to sustain a living cell. Essential genes are believed to be widespread and highly conserved during evolution (Gerdes, Scholle et al. 2003; Kobayashi, Ehrlich et al. 2003). Previous studies identified a total of 712 essential genes in *E. coli* and 271 genes in *B. subtilis* (Zhang, Ou et al. 2004; Zhang and Lin 2009). Of those, 209 were found to have a counterpart in both species (as homologous gene pairs). 48 (23%) of these essential homologous gene pairs were found as core genes in our

conserved modules. The majority of them (37 pairs) belonged to the large conserved module pair involved in translation and ribosomal metabolism. Another 17% of the homologous essential gene pairs appeared in conserved modules that were linked by a smaller number of conserved core genes than the minimum that was required in our selection (i.e. in the module pairs that were linked more weakly by homologous genes pairs and did not pass our stringent selection criteria). For the remainder of the essential genes that were not found in any module, we found that they exhibited a lower degree of co-expression with other genes in the genome than was observed on average (indicating that most likely they are not co-expressed with any other gene in our dataset).

In addition, some auxiliary processes not generally considered as essential (Kobayashi, Ehrlich et al. 2003) exhibit a highly conserved co-expression behavior between *E. coli* and *B. subtilis*. Remarkably is the group involved in flagella synthesis and motility (EM32-BM32) which recapitulated 68% of the previously characterized motility genes of *E. coli* and 78% of the genes known to be related to motility in *B. subtilis* (Rajagopala, Titz et al. 2007). The majority of the genes known to be involved in flagella synthesis with a homologous counterpart in both *E. coli* and *B. subtilis* were found in the core part (50 homologous links including 34 linked genes in the core part out of 54 total module genes in *E. coli* (63%) and 36 linked genes in the core part out of 48 total module genes in *B. subtilis* (75%)). The variable part then mainly consisted of genes occurring in one of the two species only (14 out of 20 in *E. coli* (70%) and 15 out of 22 in *B. subtilis* (68%)).

Another large group is the one involved in iron acquisition (EM59-BM52_53_59) which contains 70% of the *E. coli* and 65% of the *B. subtilis* Fur targets identified by Ollinger *et al.* (Ollinger, Song et al. 2006). Unlike motility and flagella synthesis case, here most of the known Fur targets of *E. coli* and *B. subtilis* were not found in the core part. The core part only consists of 26 homologous links (13 out of 52 total module genes in *E. coli* (25%) and 18 out of 32 total module genes in *B. subtilis* (56%)) which is a relatively small fraction compared to the large variable parts. The variable part of the *E. coli* module contained in this case 28 out of the 39 genes (72%) without homologous counterpart in *B. subtilis* and the variable part of *B. subtilis*

had 7 out of the 14 genes (50%) without counterpart in *E. coli*. This indicates that the Fur regulon largely changed during evolution to adapt to the specific needs of each organism.

2.3.7. REGULATION OF EVOLUTIONARY CONSERVED MODULES

For all conserved module pairs depicted in Figure 2.3, the co-expression behavior of their genes has largely been conserved during evolution. This does, however, not necessarily mean that also the regulatory mechanism that is responsible for this co-expression behavior is conserved. To study their regulatory mechanisms, we listed all modules with conserved co-expression behavior and assigned to each module the corresponding transcription and sigma factors by calculating the modules' enrichment in genes for a given transcription or sigma factor, according to RegulonDB or DBTBS (see Table S2.1). We used the Reciprocal Smallest Distance approach (RSD) (Wall, Fraser et al. 2003) to identify the best matching transcription and sigma factors pairs between *E. coli* and *B. subtilis*. We then determined whether modules with a conserved co-expression behavior were regulated by matching transcription and sigma factors in both organisms. By doing so, we were able to divide the evolutionary conserved module pairs into three main groups according to the sequence similarity of the transcription and/or sigma factors that were assigned to each of them.

The first group comprises conserved module pairs regulated by reciprocally best matching transcription or sigma factor pairs. To this group belonged 14 of the 82 conserved modules pairs regulated by the pairs NrdR/NrdR (EM39_40-BM39_40), Fur/Fur (EM51_52_53_55_57_58_59-BM51_52_53_55_57_58_59), LexA/LexA (EM67-BM67), BirA/BirA (EM21-BM21) and ArgR/AhrC (EM9_10_79-BM9_10_79) (where the notation corresponds to the *E. coli* gene/*B. subtilis* gene). Each of these best matching transcription factors pairs have previously been identified as functionally conserved counterparts between *E. coli* and *B. subtilis* (with Fur/Fur, LexA/LexA, and ArgR/AhrC being direct orthologs and BirA/BirA being a best matching xenolog pair as pinpointed by Price *et al.* (Price, Dehal et al. 2007)). Moreover, the best matching transcription factors pairs identified by Price *et al.* (Price, Dehal et al. 2007) as non-functional counterparts were never found to regulate our conserved modules, further confirming the power of using co-expression in inferring functionality. Also in this group we found the

conserved modules regulated by two orthologous sigma factor pairs: FliA/SigD (EM32-BM32) and RpoN/SigL (EM79-BM79).

A second group of conserved module pairs appeared to be regulated by transcription or sigma factors showing a homologous link, as predicted by COG, but not being best reciprocal matches. In this group we found four transcription factor pairs: ArcA/ResD (EM24-BM24), FruR/CcpA (EM25-BM25), GalR/CcpA (EM61-BM61), and Gals/CcpA (EM61-BM61) that could be assigned to three conserved module pairs. For the couple ArcA and ResD it is indeed known that they both are sensing aerobic versus anaerobic conditions (Vazquez, Freyre-Gonzalez et al. 2009). They both belong to large gene families for which the evolutionary history is hard to resolve and thus inferring functionality from merely sequence homology can be misleading (Sun, Sharkova et al. 1996). Just like FruR, GalR and GalS in *E. coli*, CcpA in *B .subtilis* is still involved in the regulation of carbon sources, but evolved towards a more global function than its homologous counterparts in *E. coli*. Indeed CcpA is known to be the non-homologous functional counterpart of Crp (Babu, Teichmann et al. 2006; Vazquez, Freyre-Gonzalez et al. 2009). Regarding the sigma factors regulating the modules in this group we observed the pairs: RpoD/SigA (EM1_18_32_39_43_81_82-BM1_18_32_39_43_81_82), RpoH/SigA (EM44_45-BM44_45), and RpoS/SigA (EM62_63-BM62_63). According to the COG homology definition, the house keeping sigma factor SigA of *B. subtilis* (Paget and Helmann 2003) has three homologs in *E. coli*, namely RpoD, RpoH, and RpoS. These multiple sigma factor copies have resulted in a subfunctionalization in *E. coli* of the global role executed by the sigma factor SigA in *B. subtilis* (Paget and Helmann 2003; Wade, Roa et al. 2006). This is clearly visible from our results where we found different combinations of respectively RpoD, RpoH and RpoS being responsible for the regulation of at least 12 *E. coli* modules that were paired with an equal number of *B. subtilis* modules regulated by SigA.

In the third group of conserved module pairs we found those cases where the assigned transcription regulators do not show any significant sequence similarity with each other, but they appear to regulate genes with similar function in both organisms. For 65 of the 82 conserved module pairs, at least one of the assigned transcription factors was different between *E. coli* and *B. subtilis* (summarized in the Table S2.1). For example, the master regulators FlhC and FlhD

responsible for regulation of motility and flagella synthesis in *E. coli* do not have a homologous counterpart in *B. subtilis*, while the co-expression behavior of their cognate modules is conserved (EM32-BM32). We can thus assume that a non-homologous functional counterpart, such as the recently proposed SwrA takes over the mechanism of regulation in *B. subtilis* (Calvio, Celandroni et al. 2005; Calvio, Osera et al. 2008; Smith and Hoover 2009). Indeed SwrA is known to regulate SigD in *B. subtilis* as FlhC and FlhD do in *E. coli* (Hamze, Julkowska et al. 2009).

Additional striking examples are the pairs of conserved modules in *E. coli* and *B. subtilis* regulated respectively by PurR/PurR (EM16_17_18-BM16_17_18), TreR/TreR (EM48-BM48), CysB/YwfK (EM2-BM2), MalT/AbrB (EM47-BM47). A complete list of such non-homologous transcription factors that regulate paired co-expression modules in *E. coli* and *B. subtilis* can be found in Table S2.1. PurR is known in both *E. coli* (Rolfes and Zalkin 1988) and *B. subtilis* (Weng, Nagy et al. 1995) to respond to purine excess by repressing genes of the inositol monophosphate (IMP) to adenine monophasphate (AMP) conversion pathway. TreR on the other hand controls the expression of the trehalose utilization operon in both species and its activity is known to be dependent on the cAMP gene activation protein (CAP) in both *E. coli* and *B. subtilis* (Horlacher and Boos 1997). Both pairs of similarly named transcription factors PurR/PurR and TreR/TreR constitute well documented cases of parallel evolution: despite being each other's functional counterparts in both *E. coli* and *B. subtilis* and being responsible for the regulation of an almost conserved regulon, the proteins in each pair do not exhibit any significant sequence homology, nor any similarity in their molecular mode of action (Schock and Dahl 1996; Horlacher and Boos 1997; Fukami-Kobayashi, Tateno et al. 2003).

In contrast to these well documented cases no studies exists that focus on the direct functional comparison of the pairs CysB/YwfK and MalT/AbrB in respectively *E. coli* and *B. subtilis*. The functional relation between CysB/YwfK was supported by the fact that both regulators belong to the same LysR-type of activators and they do show a low level of sequence homology (28% of sequence homology) (Guillouard, Auger et al. 2002)). Also the regulator pair was assigned to conserved modules involved in cysteine biosynthesis, a role which is well documented for CysB and YwfK. Both regulators are also related to sulfate transport (Sekowska, Kung et al. 2000;

Guillouard, Auger et al. 2002). Phenotypes of *E. coli cysB* mutants were found to be very similar to those of *B. subtilis ywfK* mutants (Guillouard, Auger et al. 2002). The conserved modules regulated by the MalT/AbrB pair were found to be in involved in maltose metabolism. In *E. coli* MalT is known to regulate seven operons of the maltose regulon (Danot, Vidal-Ingigliardi et al. 1996) that are subjected to catabolite repression (Eppler, Postma et al. 2002). For AbrB the direct role on maltose regulation is not reported. Instead AbrB is known to be a dual regulator that regulates a plethora of genes during starvation-induced processes such as those involved in sporulation, production of antibiotics and degradative enzymes (Robertson, Gocht et al. 1989). The fact that AbrB has been found to modulate the cAMP-CAP system by competing with catabolite repressor proteins during growth on carbon sources that induce partial catabolite repression, (Fisher, Strauch et al. 1994) points towards a possible functional link between MalT and AbrB.

2.3.8. DIFFERENTIATION IN EXPRESSION BY DIVERGENCE OF REGULATION

We can also find modules containing sets of genes, co-expressed in one species that got split up in different co-expression modules in the second species (Figure 2.5). We identified such gene sets as follows: a single module in the first organism should be linked to two different modules in the second organism of which the respective core parts do not share more than 30% of their genes. Such cases might point towards a condition-dependent differentiation in regulation that is observed in one species, but not in the other. Such differentiation in regulation seems to occur, for instance, for heat shock genes (EM44_45-BM44_45), most of which are chaperones and proteases known to protect cells against damage induced by protein unfolding. These genes were found to be co-expressed in *E. coli* as was also previously observed (Rasouly, Schonbrun et al. 2009). In *B. subtilis* the corresponding genes, all known to be regulated by HrcA are split up in two different modules (Schumann 2003). This observation indicates that HrcA induces a difference in expression behavior, depending on the type of transcription factors it is combined with. A potential interacting partner of HrcA could be CtsR, the transcription factor known to regulate the gene *clpE* (Schumann 2003) that belongs to one of the two evolutionary conserved modules in *B. subtilis*. Note that HrcA seems not to have a homologous counterpart in *E. coli*.

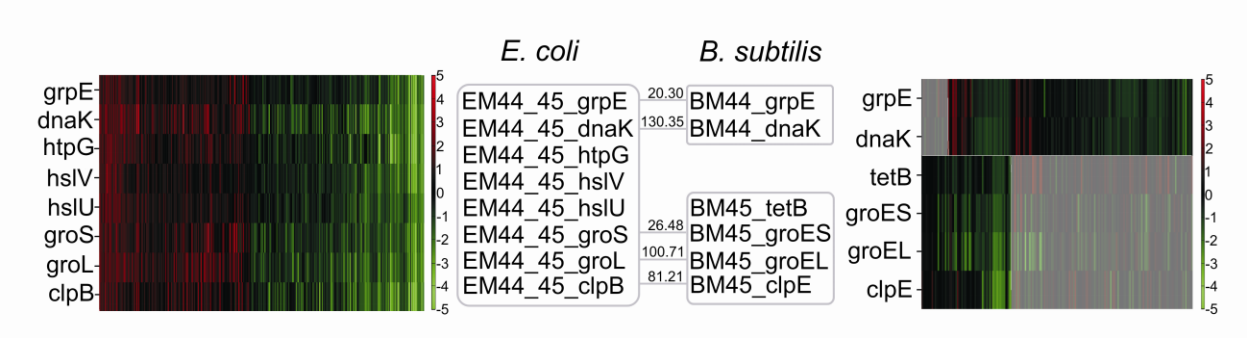


Figure 2.5. Differentiation in expression. The *E. coli* module EM44_45 (left panel) is covered by two different modules BM44 and BM45 in *B. subtilis* (right panel). Genes that belong to the same module are displayed in a gray box and homology relations are denoted by gray edges; numbers on the edges indicate Smith-Waterman alignment scores (z-values). Shaded areas in the right heatmap correspond to conditions where both *B. subtilis* modules do not overlap.

2.3.9. EXPRESSION BEHAVIOR OF LINKER GENES

The fact that the identification of evolutionary conserved module pairs was based on a many-to-many homology map allowed us to study the complex evolutionary history of several of the linker genes (Table S2.2).

At first we focused on linker genes that all showed a mutual homology. We found several of those linker genes modules, being connected by several one-to-many or many-to-many relations: e.g. paired modules that contained at least one gene in *E. coli* with multiple homologous counterparts in *B. subtilis* each of which was found in a different conserved module or the opposite way around. Those genes for which we found a divergence in mutual co-expression behavior between the homologous genes within one species could be an indication of their functional divergence as it is known that multiple copies of a particular gene in one species, resulting from horizontal gene transfer or duplication events tend to disappear unless they evolve into non-redundant copies by acquiring novel functions (neo-, subfunctionalization) (Rastogi and Liberles 2005). We found in total 19 cases of potential neo- and/or subfunctionalization (Table S2.2). For instance, the duplicated genes in *E. coli* with ribonucleotide reductase activity (Figure 2.6): each gene of the duplicated pairs *nrdA*/*nrdE* and *nrdE*/*nrdF* belongs to a different module (respectively EM39-BM39 and EM40-BM40), while the homologous counterparts of these genes

in *B. subtilis* (being *nrdE* and *nrdF*) belong to one single co-expression module. Although we found NrdR as the responsible regulator for both sets of paralogous ribonucleotide reductases genes in *E. coli*, genes within a duplicated pair exhibit a clear difference in expression behavior. Moreover, all three co-expressed genes of the *B. subtilis* conserved module (*nrdEF-ymaA*) were reported as essential genes (Kobayashi, Ehrlich et al. 2003), while their most closely related homologs (the *E. coli nrdEF* genes) were not (Zhang, Ou et al. 2004; Zhang and Lin 2009), but instead essentiality in *E. coli* was taken over by the less related homologs (*nrdAB*), reflecting a clear case of sub/neofunctionalization. Another example of complex transcriptional evolution of homologous gene families relates to the family involved in oligopeptide and dipeptide ABC transport. Figure 2.7 shows how in both organisms homologous genes are co-expressed in different conserved modules. A large fraction of homology links (indicated by blue, green and red lines) occur between members of the DppBCDF system in *E. coli* with members of the Opp and App transport system in *B. subtilis*. In each case, a gene in *E. coli* is linked to two or more genes in *B. subtilis* covering more than one co-expression module. For example, the *E. coli* gene *dppD* (EM26_27) is linked to respectively *B. subtilis oppD* (BM26_28) and *dppD, appD* (BM27_29_30). In *E. coli dppBCDF* genes form a dipeptide inner membrane ATP-binding cassette transporter involved in the uptake of heme iron (Letoffe, Delepelaire et al. 2006). In *B. subtilis* both the oligopeptide transport system Opp (Perego, Higgins et al. 1991) and the AppA system (Koide and Hoch 1994) are involved in competence development and sporulation with the App system being able to substitute the Opp system. Although both systems being functionally related in *B. subtilis,* they exhibit clear differences in their expression behavior pointing towards at least some further specialization (Figure 2.7).

For another set of homologous linker genes (16 cases) we found the multiple copies of the gene family within one species in the same module, indicating that their expression behavior was retained as a result of either recent multiplication events that did not yet result in further functional divergence, or the need of multiple gene copies for dosage effect.

In addition to these linker genes that all belonged to the same COG, we also found few examples (5 cases) where genes not exhibiting any mutual homology in one organism (not belonging to the same COG) were linked to the same gene in the other organism, implying that here two protein

domains occurring in one organism in separate genes got fused in the other organism into a single gene. One case for which the fusion was also supported by the literature was the linking gene set *purL*/*purL* and *purL*/*purQ* (Enright, Iliopoulos et al. 1999). The most interesting cases were those where the genes containing the separate or unfused domains belonged to different co-expression modules (*frwB*/*manP* and *frwC*/*manP*) as this indicates that there is a functional constraint to keep these genes unfused so that they can be differentially expressed.

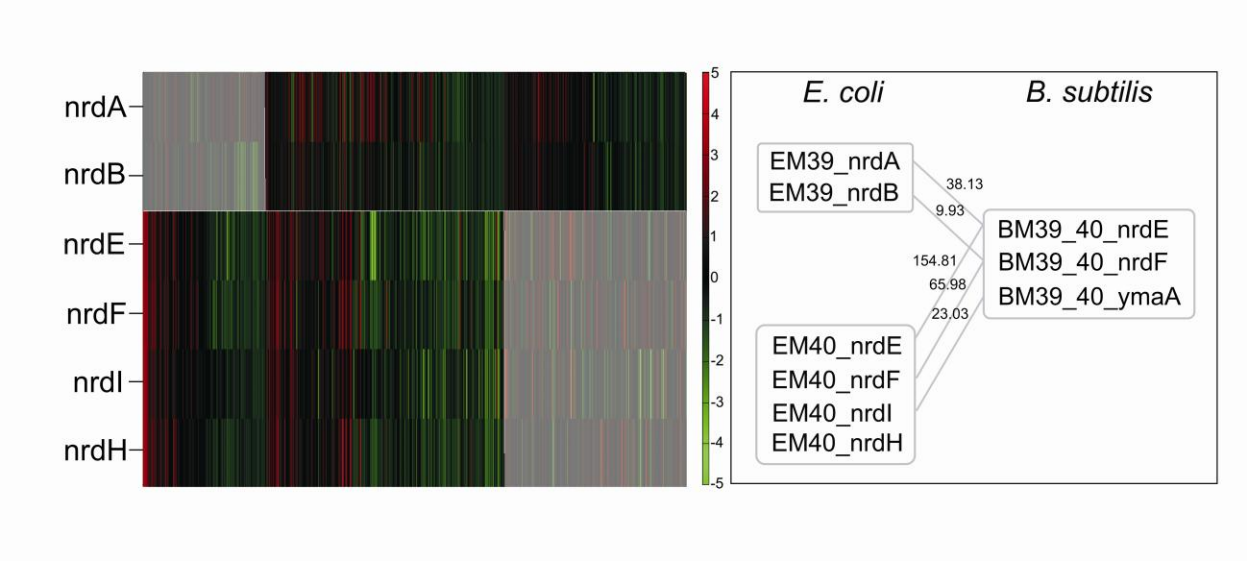


Figure 2.6. Expression divergence of duplicated genes in *E. coli*. Expression behavior of genes in modules EM39 (above the line in the heatmap) and EM40 (below the line in the heatmap) in *E. coli* (left panel). Shaded areas correspond to conditions not shared between modules. Homologous genes to the *B. subtilis nrdEF* operon (module BM39_40) were found in two different co-expression modules in *E. coli* (modules EM39 and EM40). Each module is surrounded by a gray box and homology relations are denoted by gray lines (right panel). Numbers over the lines represent Smith-Waterman alignment scores (z-values).

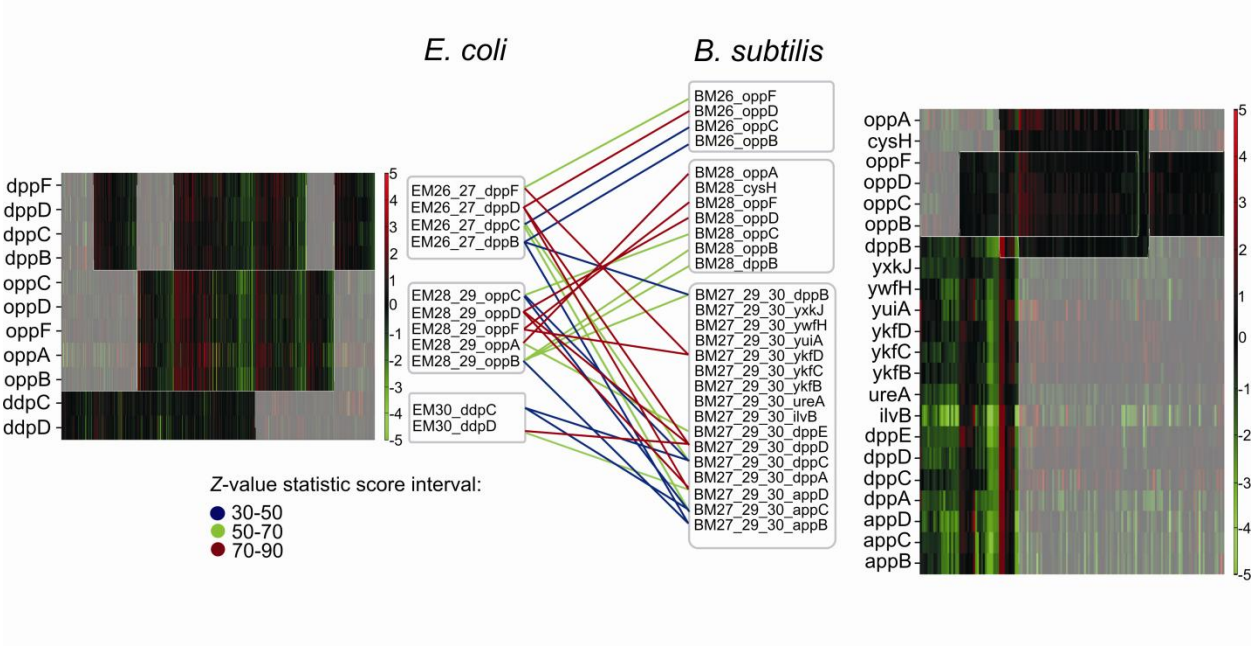


Figure 2.7. Visualization of the conserved module pairs EM26_27_28_29_30-BM26_27_28_29_30 (Table S2.1) consisting of the conserved modules involved in oligopeptide and dipeptide transport. Edges indicate homologous gene pairs (a color code is used to display the degree of homology (measured by the z-value of the Smith-Waterman alignment score). Heatmaps next to the modules illustrate the expression values of the corresponding genes for the selected conditions.

2.3.10. SENSITIVITY TOWARDS THE CHOICE OF THE PRESPECIFIED MAXIMAL CO-EXPRESSION STRINGENCY VALUE

Seed modules were identified by selecting groups of genes that showed locally a higher mutually co-expression level than their neighbors on the first subdiagonal of gene-gene threshold matrices (Figure 2.2). To prevent that we would obtain many very small seed modules containing only two genes, we set in the gene-gene threshold matrix all values larger than a prespecified maximal co-expression stringency to this value. Below we show how the choice of this prespecified maximal co-expression stringency value affected the results of our co-clustering procedure. We ran COMODO with 9 different prespecified maximal co-expression stringency values of 0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9 applied on respectively the *E. coli* and *B. subtilis* gene-gene threshold matrices. This means that for a prespecified maximal co-expression stringency

value of e.g. 0.5 all discretized Pearson correlation values in the gene-gene threshold matrix of respectively *E. coli* and *B. subtilis* that are larger than 0.5 are set to 0.5. For filtering we removed redundant module pairs (different matching module pairs that share 75% of homologous linker genes) and matching module pairs with chi-square value below 470.

Figures 2.8 and 2.9 show how the choice of the prespecified maximal co-expression stringency value affects the number of detected modules and gene coverage (genes covered by the identified matching modules pairs). Choosing a non-stringent prespecified maximal co-expression stringency value will result in fewer modules that by definition will be more loosely co-expressed, while choosing a more stringent value will result in more modules being more tightly co-expressed. As we were interested in the latter type of modules, a relatively stringent prespecified maximal co-expression stringency value is preferable. In addition, Figure 2.8 shows that gene coverage and the number of homologous pairs is stable applying the range of prespecified maximal co-expression stringency values from 0.7 to 0.9.

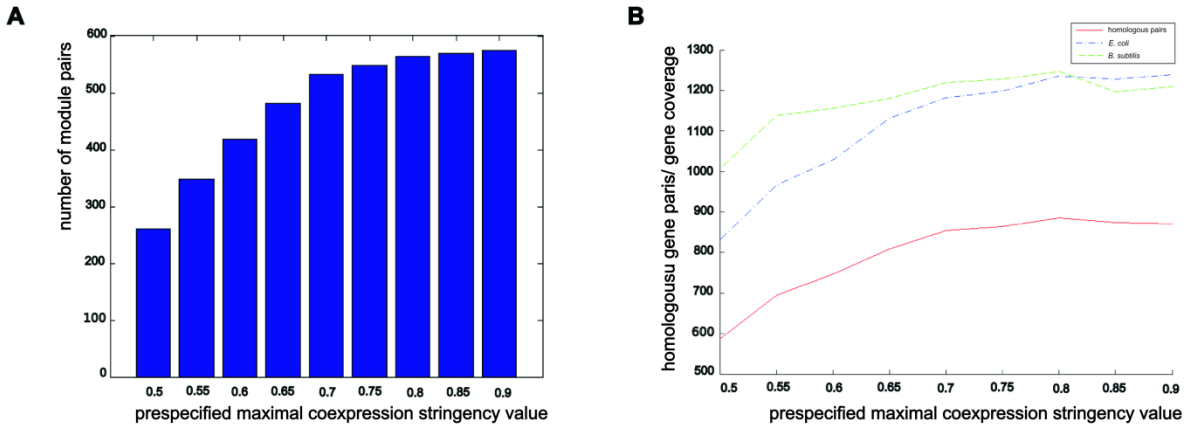


Figure 2.8. Sensitivity of the COMODO towards applying a range of different prespecified maximal co-expression stringency values. A. Number of final matching module pairs obtained by applying different values of the prespecified maximal co-expression stringency value. **B.** Number of homologous gene pairs and genes of either species covered by the obtained matching modules pairs.

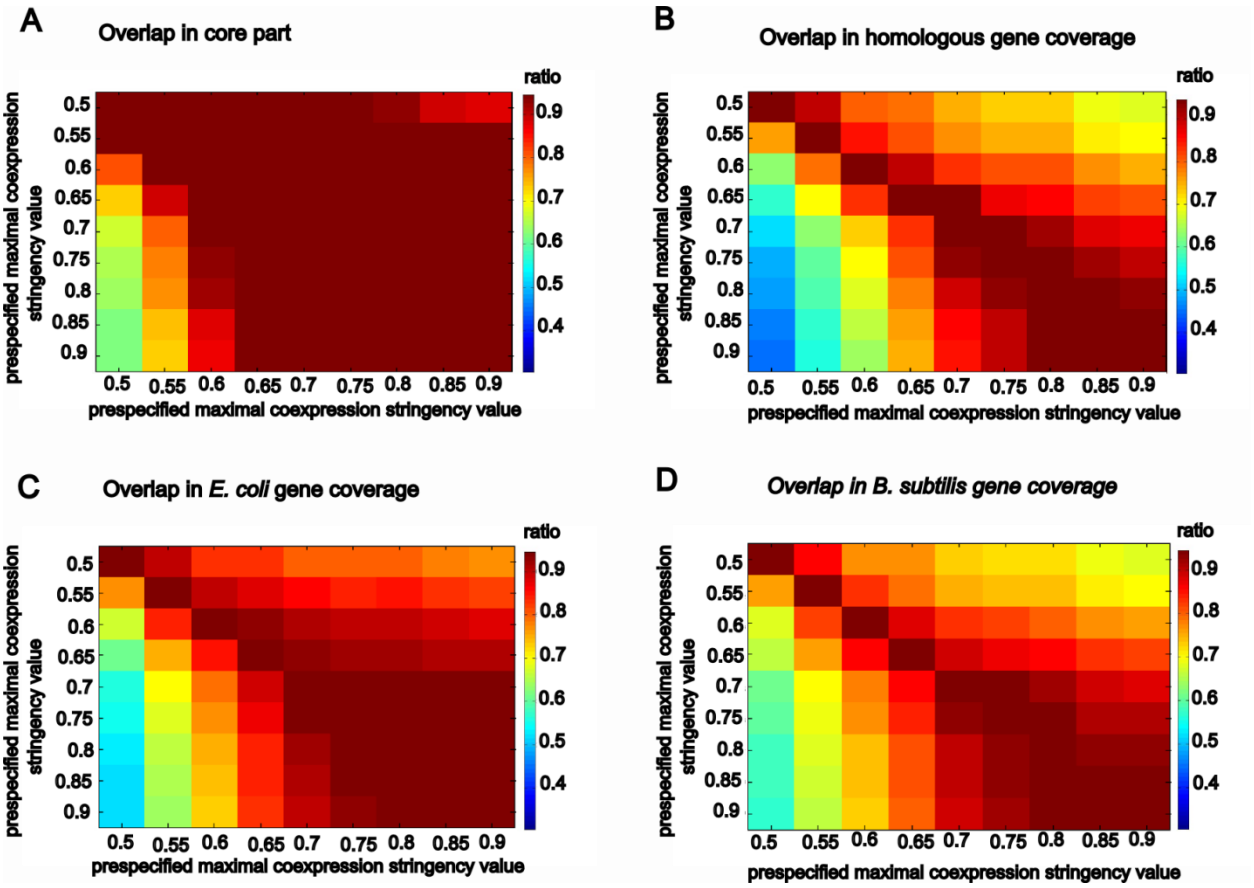


Figure 2.9. Comparison of COMODO output obtained with different prespecified maximal co-expression stringency values. These heatmaps compare the results of different COMODO runs obtained by applying the prespecified maximal co-expression stringency values indicated on respectively the X- and Y-axis. **A. Overlap in core part**: assesses to what extent the matching module pairs obtained by applying different prespecified maximal co-expression stringency values cover similar cores. Cores of different matching module pairs are considered to overlap if they are at least 75% similar in gene content. Each entry in the heatmap corresponds to the number of cores similar between runs obtained by applying respectively the prespecified maximal co-expression stringency values indicated on the X- and Y-axis, divided by the total number of cores obtained by the run with the prespecified maximal co-expression stringency value on the Y-axis. **B. Overlap in homologous gene coverage**: assesses the overlap in homologous gene pairs covered by matching module pairs obtained with different prespecified maximal co-expression stringency values. The entry in the heatmap corresponds to the total number of homologous gene pairs found in common in the cores of the matching module pairs obtained by applying respectively the prespecified maximal co-expression stringency values indicated on the X- and Y-axis divided by the total number of homologous gene pairs obtained by applying the prespecified maximal co-expression stringency value on the Y-axis. **C. Overlap in *E. coli* gene coverage:** assesses the number of genes in *E. coli* in common between matching modules pairs obtained with different prespecified maximal co-expression stringency values. The entry in the heatmap corresponds to the total number of *E. coli* genes found in common in the matching module pairs obtained by applying by

applying respectively the prespecified maximal co-expression stringency values indicated on the X- and Y-axis divided by the total number of genes obtained by applying the prespecified maximal co-expression stringency value on the Y-axis. **D. Overlap in *B. subtilis* gene coverage:** similar as C but for *B. subtilis*.

Figure 2.9 shows that in the range from 0.7 to 0.9 of prespecified maximal co-expression stringency value, the results of COMODO are quite robust against the choice of this value. However, setting this value too stringent might result in losing some modules as the seed modules might not contain a sufficient number of homologous linkers to be further extended. The results obtained by applying different less stringent prespecified maximal co-expression stringency values are less comparable as the proper threshold combination that gives rise to the best matching module pairs (with the highest chi-square values) can no longer be encountered.

2.4. DISCUSSION

COMODO is a method for cross-species co-clustering. It relies on the use of large scale co-expression compendia for each of the species to be compared. By using a bottom up approach and by exploiting homology relations to identify the optimal size and degree of co-expression in each of the modules that constitute a conserved module pair, COMODO allows identifying in each of the species the modules that best reflect the processes that are conserved in the core. The strength of COMODO relates to its ability of automatically prioritizing best matching module pairs that can cover a large range of different co-expression levels and module sizes. This feature allows the methodology to adapt to closely or evolutionary distant organisms and to identify both processes that are fully or partially conserved across evolution. Moreover, because COMODO can be used in combination with a many-to-many homology map, it is suitable to study functional relations between linker genes that mutually exhibit complex homology relations.

Applying COMODO to large scale expression compendia allowed comprehensively mapping the processes with conserved co-expression behavior in the divergent bacterial model organisms *E. coli* and *B. subtilis.* In contrast to previous studies Price *et al.* (Price, Huang et al. 2005) and Van

Noort *et al.* (van Noort, Snel et al. 2003), COMODO does not use any prior information on previously documented regulon structure or regulatory information and can thus map in an unbiased way modules with conserved co-expression between both species. Because COMODO adapts its module sizes in each species to maximize the relative number of linking homologs, it will not only identify conserved operons for which obviously the conserved co-expression signal is most pronounced, but it will also detect if they exist conserved modules comprising multiple operons.

As it was previously shown, that inferring true orthology is complicated by duplications and horizontal gene transfer (Price, Dehal et al. 2007), we combined the COG many-to-many map with our expression compendia to infer the most likely functional counterparts between *E. coli* and *B. subtilis*. Of the 5459 COG links between *E. coli* and *B. subtilis*, 355 were found in conserved module pairs. Of those 355 COG links that could be mapped to conserved module pairs, 149 represented reciprocal best hits. Those probably correspond to true functional counterparts. The other 206 most often were links of large gene families that got sub- or neofunctionalized. This figure also indicates that COG largely overestimates the number of true functional relations, although we cannot completely rule out that some of the functional links were not covered due to a lack of certain conditions in the expression datasets.

In general we found that most of the conserved modules were involved in elementary cellular processes needed to support bacterial cell duplication and inheritance of the genetic information, cell division and the provision of energy (Kobayashi, Ehrlich et al. 2003). The cores of these modules contained regulon members that were indeed shown by comparative studies to occur over a wide range of bacterial species (Babu, Teichmann et al. 2006; Lozada-Chavez, Janga et al. 2006). Modules involved in transcription, translation, and central carbon metabolism contained genes that were previously shown to be differential expressed during the global response to glucose in both *B. subtilis* and *E. coli* (Vazquez, Freyre-Gonzalez et al. 2009). Despite covering mainly elementary processes our conserved modules contained relatively few essential genes. This, together with the fact that the conserved modules covering elementary processes were rather small (restricted to a single or to maximally a few transcription units, except for those involved in ribosomal metabolism and translation) confirms the previous suggestion that

essential processes seem not to be primarily coordinated by the modulation of gene expression (Kobayashi, Ehrlich et al. 2003).

In addition to these smaller modules we also found larger conserved module pairs that were mainly involved in iron acquisition (Fur regulon) and flagella synthesis. While both processes are fairly conserved at the level of their gene content, mainly the process involved in iron acquisition has underwent major changes in regulon membership in either species.

The mechanism by which genes were transcriptionally co-regulated seemed to be much less conserved than their co-expression behavior itself: while the co-expression behavior of complete orthologous regulons was maintained over evolution, the transcription factors responsible for their regulation were only conserved in few cases as was also observed by Price *et al.* (Price, Dehal et al. 2007). However, in most cases the ortholog of a particular transcription factor known to be responsible for the co-expression behavior in one species did not exist in the other species, suggesting that the role of the disappeared transcription factor must have been taken over by an alternative, yet unknown but non-homologous transcription factor. Furthermore we observed that the variable part in *E. coli* or *B. subtilis* of the conserved modules largely consisted of genes specific for one organism, but not occurring in the other one, indicating that bacteria are also flexible in adding new members to an existing regulon (Babu, Teichmann et al. 2006; Lozada-Chavez, Janga et al. 2006; Price, Dehal et al. 2007). These observations suggest that despite the extreme potential of network rewiring, prokaryotes are extremely robust in preserving the co-expression behavior of some elementary pathways. Probably the operon structure contributes largely to this robustness against rewiring by maintaining a minimal level of co-expression (Lemmens, De Bie et al. 2009).

CHAPTER 3
EXTENDING COMODO TO THREE ORGANISMS: APPLICATION ON *S. ENTERICA*

3.1. INTRODUCTION

One of the key issues in system biology is to identify functional orthologous genes. These functional orthologous not only share sequence ancestry, but also perform the same function. Microarray expression data is a genome-scale high-throughput experiment which can identify genes with similar function with high accuracy (Chikina and Troyanskaya 2011) as genes with similar function tend to have more similar expression profiles.

In the previous chapter we introduced COMODO as a methodology which can detect co-expression conservation across two different organisms. As we mentioned in previous chapter, COMODO is initialized with co-expressed seeds or seed modules obtained in each of the species. These seeds are gradually expanded in each of the species until a pair of modules is obtained for which the number of shared homologs is statistically optimal relative to the size of the linked modules. The strength of COMODO relates to its ability of automatically prioritizing best matching module pairs that can cover a large range of different co-expression levels and module sizes.

In this chapter, we improve COMODO by first detecting a larger number of co-expressed seed modules in each organism. In each organism, seed modules are identified by applying a prespecified maximal stringency threshold (see 2.2.3). We enabled COMODO to apply a range of prespecified maximal stringency thresholds to detect more initial seed modules in each organism. In addition, we extended the optimization criteria to three organisms to detect co-expression conservation across three organisms.

Although previous cross-species comparison studies have revealed the co-expression conservation and also regulatory network conservation across prokaryotic (Lozada-Chavez, Janga et al. 2006; Okuda, Kawashima et al. 2007; Perez and Groisman 2009; Zarrineh, Fierro et al. 2011) and eukaryotic (Bergmann, Ihmels et al. 2004; Ihmels, Bergmann et al. 2005; Oldham, Horvath et al. 2006; Tirosh and Barkai 2007; Chikina and Troyanskaya 2011) organisms over long ranges of phylogenetic distances, still it is not clear to what extent life style influence the conservation of both co-expressed modules and the regulatory network across phylogenetically close organisms.

To explore the effect of life style, we studied the co-expression conservation of two evolutionary close prokaryotic model organisms: *E. coli* and *S. enterica*. Although these two gram-negative bacteria are evolutionary very close organisms, *S. enterica* is a dangerous human pathogen, and this made it interesting to investigate which evolutionary changes had caused the human pathogenecity in *S. enterica*. We applied COMODO to search for conserved modules in *E. coli* and *S. enterica*. To explore the general co-expression conservation across bacteria phyla, we also applied COMODO to search for conserved modules in three prokaryotic model organisms: *E. coli*, *B. subtilis*, and *S. enterica*. Moreover, we studied the genes involved in quorum sensing as the quorum sensing may be influenced by life style. Similarly, we also investigated genes, with pathogenesis function, as the pathogenecity is one of the major differences between *E. coli* and *S. enterica*, and it directly related to the life style.

3.2. MATERIALS AND METHODS

As described in the previous chapter, the homology map between different bacteria was derived from the COG database (Tatusov, Fedorova et al. 2003), and orthologous gene families were derived using smallest distance approach (Wall and Deluca 2007). The same microarray compendia was used for *E. coli* and *B. subtilis* as in the previous chapter, and the microarray compendium of *S. enterica* was obtained from COLOMBOS (Engelen, Fu et al. 2011) containing 657 conditions. The basic validation analysis like assigning gene ontology, metabolic pathway to

the modules and condition selection to visualize the co-expression of modules were performed as described in the previous chapter.

3.2.1. STATISTICS TO ASSESS CO-EXPRESSION CONSERVATION BETWEEN TWO OR THREE ORGANISMS

Two data sources, sequence similarity and expression compendia, are used to detect expressionally conserved genes across multiple organisms. Detecting genes with sequence similarity is more straightforward as different orthologous or homologous gene families are available through different databases (Tatusov, Fedorova et al. 2003; Wall, Fraser et al. 2003; Wall and Deluca 2007). The idea behind COMODO is to find proper threshold to detect co-expression modules in two or more organisms, in a way that maximizes the observed linked homologous genes using a proper statistical test.

To define proper statistics between two organisms, the homology relation can be considered as a bipartite graph where nodes on each side are the genes in each one of the organisms and edges represents the homology relation between the genes. Given two modules one in each organism, the p-value of observing such a module and homologous gene pairs can be calculated by Monte Carlo sampling method. To perform Monte Carlo sampling, in each step we shuffled two edges in a way that the distribution of degrees in the bipartite graph remained preserved. This shuffling is done by repeatedly selecting at random two edges and crossing them (replacing two homology relations). If two modules C_i and C_j are linked with $|T|$ homology relations, and shuffling procedure was performed n times, we calculate p-value as follow:

$$p = \frac{number\ of\ times\ two\ modules\ are\ linked\ with\ \ more\ than\ |T|\ links}{n}$$

To extend this formula to three organisms, the homology relation between three organisms can be considered as a tripartite graph where nodes are genes and edges are the homology relations, and Monte Carlo sampling can be done in a similar way for this tripartite graph. Each homology relation consists of three genes in three organisms linked all by homology. To perform Monte

Carlo sampling, in each step we choose two homology relations, in a way that each two genes of the same organism will be different. Now it is sufficient to reshuffle two links out of three links exists in homology relation. As an illustration, by considering two homology relations as (G_{11}-G_{21}-G_{31}) and (G_{12}-G_{22}-G_{32}), we can obtain three valid reshuffling: 1. (G_{12}-G_{21}-G_{31}) and (G_{11}-G_{22}-G_{32}) 2. (G_{11}-G_{22}-G_{31}) and (G_{12}-G_{21}-G_{32}) 3. (G_{11}-G_{21}-G_{32}) and (G_{12}-G_{22}-G_{31}) where the genes, written in a parenthesis, are all connected and ordered based on the organism number.

Therefore, if three modules C_i, C_j, and C_k are linked with |T| homology relations and shuffling procedure was performed n times, we calculate p-value as follow:

$$p = \frac{number\ of\ times\ three\ modules\ are\ linked\ with\ \ more\ than\ |T|\ links}{n}$$

Notice that the links just between two organisms are not considered, and we are just interested in the ones that link genes in all three organisms in the optimization part.

Although Monte Carlo sampling is the standard node-permutation method, in general when sample sizes are large, the Pearson's chi-square test will give accurate results to calculate p-values derived from permutation tests. Running Mont Carlo sampling in each iterative step of COMODO is not computationally feasible because in each step of COMOD Monte Carlo sampling should be preformed, and consequently the edge reshuffling procedure should be run for millions of times just for each step. Pearson's chi-square statistic test can be used instead as a proper test to approximate the p-value. The assumption behind the Pearson's chi-square statistic is that the homologous links are evenly distributed in the bipartite (in two organisms case) and tripartite (in three organisms case) graphs. In other words, nodes with large number of connections cause problem for estimating the real p-value with Pearson's chi-square statistic test. The fact that each gene does not have large number homologous pairs will fulfill the assumption of using Pearson's chi-square statistic to estimate the real p-value.

This Pearson's chi-square static was introduced in, the previous chapter for two organisms. To formulate the Pearson's chi-square test, consider N_1 genes in the genome of the first organism, N_2 genes in the genome of the second organism, N_3 genes in the genome of the third organism, and M linking homologous gene triples derived from the COG database. If we pick three genes

randomly, one from each organism, the probability that a homologous gene triple has been chosen is equal to $\frac{M}{N1 \times N2 \times N3}$. Therefore, the probability that these genes are not homologous is $1 - (\frac{M}{N1 \times N2 \times N3})$.

Given three modules (one for each organism) containing respectively g_1 genes from the first organism, g_2 genes from the second organism, and g_3 genes from the third organism (where g_1 , g_2, $g_3 << N_1$, N_2, N_3 respectively), the expected number of homologous gene triples that would appear assuming that the three modules are randomly selected modules can be estimated by:

$$E_{homologous} = g_1 \times g_2 \times g_3 \times (\frac{M}{N1 \times N2 \times N3})$$

The expected number of non-homologous gene triples appearing between them can be estimated by:

$$E_{non-homologous} = g_1 \times g_2 \times g_3 \times (1 - \left(\frac{M}{N1 \times N2 \times N3}\right))$$

We use the Pearson's chi-square test to assess whether the number of homologous and non-homologous gene triples in an observed module pair is significantly different from the expected one. A chi-square test with one degree of freedom is as follow:

$$\chi^2 = \frac{O_{homologous} - E_{homologous}^{\ 2}}{E_{homologous}} + \frac{O_{non_homologous} - E_{non_homologous}^{\ 2}}{E_{non_homologous}}$$

Where *O* and *E* stand for observed and expected values respectively. Note that as the *p*-value might get very close to zero, we use an optimization criterium that maximizes the actual chi-square values instead of minimizing the corresponding *p*-values.

3.2.2. APPLICATION OF THE METHODOLOGY TO THE E. COLI, B. SUBTILIS, AND S. ENTERICA DATASETS

The COMODO methodology was expanded to accept a range of prespecified maximal co-expression stringency values to detect more module seeds in each organism. We used five prespecified maximal co-expression stringency values, 0.9,0.8,0.7,0.6, and 0.5, in this study. As the previous chapter, the Pearson correlation across all conditions was used as the measure for co-expression. In theory, using five prespecified maximal co-expression stringency values results in five different module seeds, but in practice many of these module seeds are identical.

We used COMODO for two organisms to find co-expression conservation between *E. coli* and *S. enterica* with the same setting and filter procedure as the previous chapter experiments. Table S3.1. summarizes the conserved modules across these two species. In this table, the orthologous gene pairs are linked with red line to each other and homologous gene pairs with black line.

COMODO was also extended to find expressional conserved modules in three organisms. We applied COMODO to find conserved modules across three bacterial *E. coli*, *B. subtilis*, and *S. enterica*. For three organisms we also used the same setting and filter procedure as previous chapter experiments except we used 0.2 as the minimal fraction of homologous versus non-homologous genes for one of the stopping criteria and also we used the same fraction number in initial module selection step for least initial linker genes in each module. We used 0.2 instead of 0.1 for these two variables in our experiments to reduce the number of linked module triples which make more memory efficient as for three organisms the use of memory is much higher, and also to reach to the stopping criteria faster as searching in the best threshold for three modules (each for one organism) can be much slower than two. In addition, the highly conserved co-expressed modules contain much higher ratio of genes linked by homology. Table S3.2. summarizes the conserved modules across these three species. In this table, the orthologous gene triples are linked with red line to each other and homologous gene pairs with black line. Homologous gene pairs just in two organisms are also reported in separate columns, although they had no effect in the calculation of the optimization criteria.

3.3. RESULTS

3.3.1. IDENTIFYING EVOLUTIONARY CONSERVED AND NON-CONSERVED CO-EXPRESSED MODULES BETWEEN *E. COLI*, *B. SUBTILIS*, AND *S. ENTERICA*

In previous chapter, we applied our methodology to study the degree to which co-expression modules have been conserved between two bacterial model organisms *E. coli* and *B. subtilis*. In this chapter, we also applied COMODO over *E. coli* and *S. enterica*. Applying our method resulted in the identification of 211 conserved module pairs in *E. coli* and *S. enterica* that were linked through a statistically significant set of homologous genes Table S3.1). We also applied the extended COMODO for three organisms with over *E. coli*, *B. subtilis*, and *S. enterica*. Applying our method resulted in the identification of 110 conserved module triples in *E. coli*, *B. subtilis*, and *S. enterica* that were linked through a statistically significant set of homologous genes (Table S3.2).

As it was mentioned in previous chapter large evolutionary conserved modules are enriched for ribosomal metabolism and translation (Table S3.1 EM81_91_96_97-SM81_91_96_97 to 97 and Table S3.2 EM58_59-BM58_59-SM58_59), motility and flagella synthesis (Table S3.1 EM160-162-SM160_ and Table S3.2 EM83_89-BM83_89-SM83_89), iron acquisition (Table S3.1 EM172-SM172 and Table S3.2 EM95-BM95-SM95). Conversely, two large evolutionary conserved modules just in *E. coli* and *S. enterica* related to cellular respiration, both anaerobic respiration (Table S3.1 Module IDs EM44_48_51-SM44_48_51) and aerobic respiration (Table S3.1 EM159-SM159), seemed to be diverged in more distant bacterium *B. subtilis*.

Cellular respiration is not the only process with conserved expressional behavior between *E. coli* and *S. enterica*, and diverged expressional behavior in *B. subtilis*, many smaller modules in size are also conserved only across *E. coli* and *S. enterica*. Interestingly, some of them are related to signal transduction and response to stimuli regardless of different life style of these organisms. For example, response to various stimuli are specific to *E. coli* and *S. enterica* like response to stress (Table S3.1 EM20_21_71_72_76_112_118_120-SM20_21_71_72_76_112_118_120), response to external stimulus (Table S3.1 EM59-SM59), response to chemical stimulus (Table S3.1 EM121-SM121), and response to abiotic stimulus (Table S3.1 EM204-SM204). Signal

transduction modules with conserved expressional behavior between *E. coli* and *S. enterica* (Table S3.1 EM25_26_27_126_184_185-SM EM25_26_27_126_184_185) are other remarkable examples.

3.3.2. REGULATORY NETWORK CONSERVATION

Regulatory network evolves rapidly as various species need to adapt themselves to different environment. In previous chapter, we have highlighted evolutionary conserved transcription factors with conserved co-expressed targets like Fur, NrdR, LexA, BirA, ArgR/AhrC. These regulators are highly conserved across bacteria phyla, thus we also found them in three organisms, *E. coli*, *B. subtilis*, and *S. enterica*, experiment conserved (Table S3.2) except for BirA where its target are not highly co-expressed in our *S. enterica* due to the lack of conditions.

As mentioned in the previous chapter non-orthologous transcription factors may be responsible to regulate similar biological processes in *E. coli* and *B. subtilis*. Since *E. coli* and *S. enterica* are evolutionary close organisms and most of the genes have high sequence similarity, we can expect that in the latter case transcription factors are actual orthologous transcription factors with the same function. As an example, PurR, the transcription factor of purine biosynthesis processes (Table S3.2 EM75_105_106-BM75_105_106-SM75_105_106), do not show any sequence homology between *E. coli* and *B. subtilis*, while it shows very high sequence similarity between two closer organisms *E. coli* and *S. enterica*, and we know they are actual functional orthologous pairs (Yang, Lu et al. 2001; Yang, Lu et al. 2006). As the regulatory network does not exist for *S. enterica*, to validate the expected regulator we need to perform laboratory experiments like regulator mutation in PurR case (Yang, Lu et al. 2001), or we can check the upstream binding site motifs. If there is a high conservation at both protein sequence level of transcription factor and upstream binding site motifs of co-expressed target genes, it will be highly probable that this transcription factor is responsible for observed co-expression in the module.

Co-expression conservation of regulators themselves may also imply the similarity in regulatory interaction conservation. For example, conserved FliA/SigD, the flagellar sigma factor (sigma 28), and its anti sigma factor, FlgM are also highly conserved in co-expression across all the three bacteria as they co-expressed with motility and flagella synthesis (Table S3.2 EM89-

BM89-SM89). Conserved sigma factor FliZ, a protein which acts as a sigma(S) inhibitor (Pesavento, Becker et al. 2008), also co-expressed in this group, but this protein just exists in *E. coli* and *S. enterica*. LexA is another conserved regulator in all three bacteria which is also conserved in co-expressed (Table S3.2 EM108-BM108-SM108).

FliZ is not the only regulator which exhibits co-expression conservation only between *E. coli* and *S. enterica*. Self-regulatory transcription factors MtlR (Table S3.1 EM63-SM63), LldR (Table S3.1 EM183-SM183), IscR (Table S3.1 EM190-SM190), GlnG (Table S3.1 EM200-SM200), PhdR (Table S3.1 EM96-SM96), and Fis (Table S3.1 EM84_96-SM84_96) are also conserved in expression across *E. coli* and *S. enterica*. Therefore, we expect that these transcription factors are conserved across these two organisms. Two anti-sigma factors RseA and RseB are also conserved in expression (Table S3.1 module EM193-SM193), we expect the corresponding sigma factor RpoE must also be conserved in co-expression as they are in one operon in both *E. coli* and *S. enterica*, and perhaps the reason that it was not detected in *S. enterica* is the conditions set of S. *enterica* microarray (Figure 3.1.B). Furthermore, we know that in general sigma factors and anti-sigma factors are highly conserved across *E. coli* and *S. enterica* (Paget and Helmann 2003; Pons, Gonzalez et al. 2006; Menard, Santos et al. 2007).

In addition to the mentioned conserved orthologous regulators, there are cases of homologous regulators which show similar co-expression conservation meaning that the linked homologous appeared in the same linked co-expressed modules. This implies the possibility of taking over the regulatory function by homologous counterparts. For example, transcription factor STM0347 is homologous to self-regulator CsgD, and they are both co-expressed in linked co-expressed module (Table S3.1 EM166-SM166). In addition, CsgD is the main regulator of this linked module in *E. coli*. Although the direct orthologous gene CsgD exist in *S. enterica* and is known to be responsible for the co-expression in both organisms (Pesavento, Becker et al. 2008) and was reported to be the responsible regulator in both organisms(Pesavento, Becker et al. 2008), still STM0347 may also be responsible for the regulation of the co-expressed *S. enterica* genes. CsgD was probably not detected as co-expressed gene in *S. enterica* by COMODO because of condition set in this organism (Figure 3.1.A). As another example, two co-expressed *E. coli* transcription factors UidR and FeaR show homology and conservation in expression with two

co-expressed *S. enterica* transcription factors STM0580 and STM0581 respectively (Table S3.1 EM201-SM201). The fact that for both transcription factors in *E. coli* the orthologous counterpart in *S. enterica* were not detected by reciprocal smallest distance method (Wall and Deluca 2007), it will make our point stronger that the corresponding *S. enterica* transcription factors, STM0580 and STM0581, may have acquired the function, although the whole observation can be the artifact of using loose COG homology relations. Finally, we found two linked homologous transcription factors UhpA in *E. coli* and SsrB in *S. enterica* (Table S3.1 EM58-SM58) where the possibility of considering them as functional counterpart is very low because probably observation of this co-expression conservation is the artifact of loose COG homology relations.

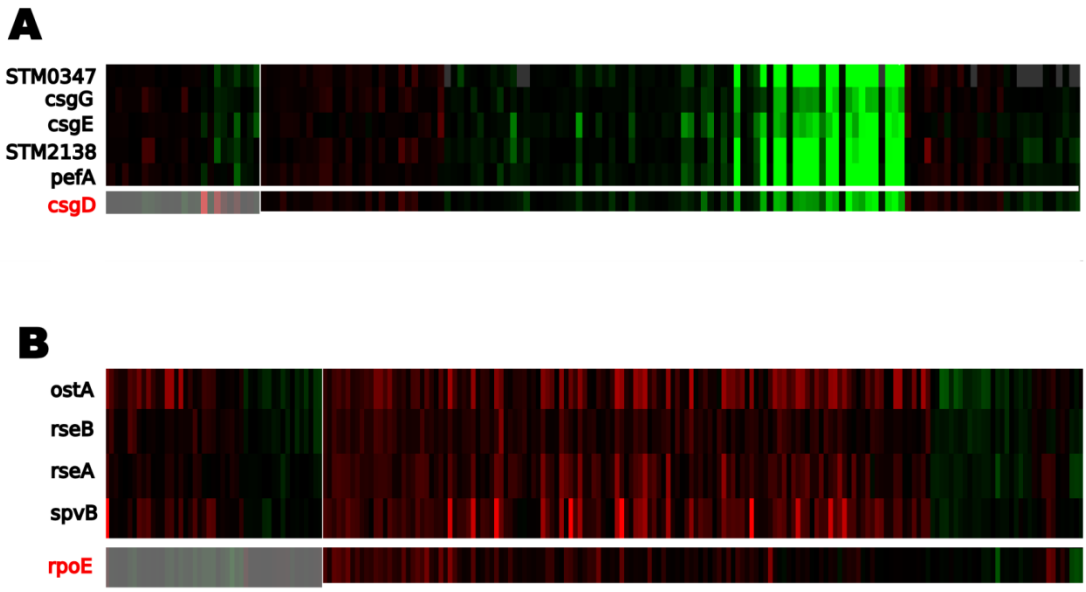


Figure 3.1. Expression behavior of genes in modules SM166 (Panel A) and SM193 (Panel B) in *S. enterica*. Genes in black are the genes which are found as the co-expressed modules by COMODO. While genes in red (csgD and rpoE) are the ones which are not found in the co-expressed modules, but their ortholgous pair are co-expressed with the *E. coli* counterpart modules. We expect that genes in red (csgD and rpoE) should also be part of their modules as they are in the same operon with some genes of their modules. Shaded areas correspond to conditions not shared for the genes which were not detected as co-expressed in *S. enterica* (red genes). The fact that these conditions are much smaller in number than the conditions genes in red (csgD and rpoE) show co-expression with the rest of the modules genes increases the probability that these genes are actually in those modules.

3.3.3. EXPRESSION COMPARISON OF GENES INVOLVED IN QUORUM SENSING AND PATHOGENECITY

Here we explore the co-expression of genes involved in quorum sensing and pathogenicity as quorum sensing and pathogenicity are influenced by the life style. Quorum sensing is the mechanisms that bacteria use to coordinate their behavior in various environments. Interestingly, from four gene products known to be responsible in quorum sensing (GO:0009372) in *E. coli* (Carbon, Ireland et al. 2009), two genes, LsrK (Table S3.1 EM5-6-SM5-6) and LuxS (Table S3.1 EM126-SM126) were found conserved in co-expression, and one gene MqsR does not have orthologous counterpart in *S. enterica*.

As *S. enterica* is a human pathogen we also looked at gene products involved in pathogenesis (GO:0009405). Six *E. coli* genes were listed as Pathogenesis (Carbon, Ireland et al. 2009), but none of them were found in any conserved module. From 60 gene products known to have function Pathogenesis in *S. enterica* (Carbon, Ireland et al. 2009), sseA (Table S3.1 SM17_18_19), slyA (Table S3.1 SM20), spvB (Table S3.1 SM193), and no pathogenesis gene was linked by homology to any gene in the counterpart linked module.

Perhaps the most interesting case of the mentioned pathogenesis genes in *S. enterica* is SpvB. This gene was co-expressed in a co-expressed link module where two anti-sigma factors, RseA and RseB, are found conserved (Table S3.1 EM193-SM193) and one sigma factor, RpoE, is seemed to be conserved (see 3.3.2). Therefore, we can presume that response to the extracytoplasmic/extreme heat stress factors have some relation with pathogenecity in *S. enterica* case. The fact that OstA, which is the protein responsible to osmotic stress, was found co-expressed just in *S. enterica* module (Table S3.1 SM193) makes our guess even stronger, that pathogenesis have relation with responding to the stresses in *S. enterica*.

3.4. DISCUSSION

Co-expression can highlight functional similarity of homologs across different species (Chikina and Troyanskaya 2011). We could extend COMODO to detect co-expression conservation across three species and we applied extended COMODO to detect conservation across *E. coli* and *S.*

enterica and *B. subtilis*. We could detect conserved biological processes across these three organisms which seems to be conserved across the whole bacteria phyla, as well as biological processes which are conserved across two closely related species *E. coli* and *S. enterica* such as aerobic and anaerobic respiration. Interestingly, many modules related to response to various stimuli and signal transductions were among the biological processes which were just conserved across two evolutionary closer species *E. coli* and *S. enterica*, even though some aspects of their life style are remarkably different (pathogenicity of *S. enterica*).

The conservation and divergence of the co-expressed genes illustrate the evolutionary path that each species might go through to adapt itself to the environment, but more importantly the regulatory network responsible for the observed expression should evolve rapidly not only to control the expression of genes involving in different biological processes, but also to enable the organism to interact to convey various signals from environment into the cell. Therefore, the structure of regulatory network is highly divergent even for two closely related organisms regardless of high conservation of observed co-expression (Lozada-Chavez, Janga et al. 2006). We observed the conservation of few biological processes even in all three organisms which is in line with previous knowledge, as the conservation of the target of these transcription factors and their upstream binding site motifs have been discussed in depth in separate focused papers (Fur (Andrews, Cartron et al. 2006; Chen, Lewis et al. 2007), NrdR (Rodionov and Gelfand 2005), LexA (Erill, Campoy et al. 2007), birA (Rodionov, Mironov et al. 2002), ArgR/AhrC (Gelfand, Makarova et al. 2001)). In addition, we could predict some regulatory network conservation just in *E. coli* and *S. enterica*. The conserved regulators are basically sigma factor and sigma factors which are known to be highly conserved (Chadsey, Karlinsey et al. 1998; Paget and Helmann 2003; Rhodius, Suh et al. 2006; Pesavento, Becker et al. 2008; Osterberg, Del Peso-Santos et al. 2011), and also some self regulatory transcription factors as they co-expressed with their targets.

The fact that we observe high co-expression conservation across *E. coli* and *S. enterica*, even conservation in various stimuli and signal transductions, and also we could predict some conservation in regulatory network although this network is not available for *S. enterica*, made this question even harder that what made their difference in life style (pathogenicity of *S. enterica*). Therefore, we investigated genes involved in quorum sensing and pathogenesis.

Amazingly, two genes out of four genes involved in quorum sensing in *E. coli* were conserved in co-expression. Finally, we could observe the major source of difference in two organisms by exploring genes involved in pathogenesis. The dominant majority of these genes were not co-expressed in the linked co-expressed modules, and SpvB in *S. enterica* was co-expressed in the module which is responsible to address stresses like extracytoplasmic, extreme heat stress and osmotic stress. This may imply the relation between pathogenesis and these stresses.

In conclusion, we investigated two phylogeneticly close species *E. coli* and *S. enterica* with some differences in their life style (pathogenicity of *S. enterica*), and we could observe high conservation in responses to various stimuli, transductions of different signals, quorum sensing. Even the comparison of the regulatory network structure based on the available knowledge show some conservation, and considering that the regulatory network is highly variable even in close species we could not conclude that different life style have a great impact on this network. The only large divergence that we could observe was the genes involved in pathogenesis. Therefore, we observed that the phylogenetic distance has far more effect on the co-expression conservation than life style.

CHAPTER 4

INFERRING CO-REGULATED GENES FROM REGULATORY NETWORK

4.1. INTRODUCTION

Gene expression is controlled by regulators such as sigma factors, transcription factors (TFs), and small RNAs (sRNAs). The interaction between regulators and their target genes/mRNA can be represented as a network. This interaction network is highly flexible through the evolution to empower organisms to adapt themselves to various environmental changes and perturbations (Hyduke and Palsson 2010). Therefore, Regulatory networks can be seen as the internal controlling system inside the cell, that have two major tasks, conveying the signal from the environment into the cell and controlling the expression of the genes. The fact that these two tasks are very different makes, it challenging to detect modularity in the regulatory network in a way that it can describe both of them. Although the interactions evolve rapidly in the regulatory network, the general structure of the circuits within this network follow 'general patterns' which is referred to as motifs (Yu and Gerstein 2006; Michoel, Joshi et al. 2011). Regulation of a gene may be controlled by more than one regulator which can be referred to as combinatorial regulation (Fadda, Fierro et al. 2009; Lemmens, De Bie et al. 2009; Kim, Bhardwaj et al. 2010). Both highlighting circuit motifs and combinatorial regulation have gained a lot of attention recently (Balaji, Babu et al. 2006; Yu and Gerstein 2006; Lemmens, De Bie et al. 2009; Gerstein, Bhardwaj et al. 2010; Kim, Bhardwaj et al. 2010).

Regulatory networks give rise to gene expression. Therefore, for genes exhibiting higher co-expression on microarray expression compendia, a higher similarity in regulatory circuits is expected. The regulators enriched as the common regulators of a co-expressed module are usually referred as combinatorial regulators, and usually a Fisher exact statistical test is used to detect the combinatorial regulators. Combinatorial regulators were detected in various organisms including two prokaryotic model organisms *E. coli* (Lemmens, De Bie et al. 2009) and *B. subtilis* (Fadda, Fierro et al. 2009).

Combinatorial regulators are detected by considering the co-expressed target genes, and it can be seen as a way to define modularity in the regulatory network. However, modularity in regulatory network was also studied just by considering the regulatory network structure. In this way, regulators that share large number of targets are detected and called collaborative regulators (Gerstein, Bhardwaj et al. 2010). Collaborative regulators are expected to give rise to the co-expression of their common targets. In fact by assuming modularity in the regulatory network, combinatorial regulators and collaborative regulators are becoming closely related definitions. The difference is that combinatorial regulators are detected considering co-expressed targets, while collaborative regulators are detected first from the regulatory network, and co-expression of the targets is the subsequence (Figure 4.1). A measure for the degree of collaboration for a pair of regulators was introduced in (Balaji, Babu et al. 2006). They simply calculate Jaccard coefficient targets of two regulators i and j as:

$$\frac{G_i \cap G_j}{G_i \cup G_j}$$

Where G_a is the set of targets of regulator a. Despite, these attempts to associate the observed co-expression with the modules inside the regulatory network, the consistency between the expression profile of the target genes and the interactions in the regulatory network was shown to be low (Gutierrez-Rios, Rosenblueth et al. 2003; Herrgard, Covert et al. 2003).

To overcome the mentioned problem, (Marr, Theis et al. 2010) tried to explore the relation between regulators and their targets. They defined subnet in regulatory networks as the subgraph induced by all nodes downstream of a root node, including the root node (Marr, Theis et al. 2010). A subnet can be seen as a root regulator and its direct and indirect targets, and the degree of co-expression on the microarray expression compendia for a subnet shows to what extent the root regulator has influence on the expression of its direct and indirect targets. Figure 4.1 demonstrates some differences between subnets and collaborative regulators and combinatorial regulators. One major advantage of subnet approach is that it traverses the whole regulatory network from a root regulator while in detecting collaborative regulators and combinatorial regulators the regulatory network was summarized as a bipartite network where regulators are in one side and targets in the other sides.

In addition to combinatorial regulators and collaborative regulators, recent studies also focused on the hierarchy of the regulatory network to decipher regulatory dynamics and network architecture (Gerstein, Bhardwaj et al. 2010; Kim, Bhardwaj et al. 2010). The importance of regulators is directly related to their positions in the hierarchy (Kim, Bhardwaj et al. 2010). One reason for this is that the feed forward loop in the regulatory network is abundant. Therefore, if a regulator regulates another regulator usually it regulates the targets of the lower regulator as well. Furthermore, upper-level regulators or global regulators have few functionally redundant copies, and had a shorter half-life (Kim, Bhardwaj et al. 2010). The upper-level regulators are believed to be noisier to be used to predict any function for their targets, as their targets may not exhibit high co-expression and also enriched for a specific function, while the lower-level, local, regulators have more direct effect on the expression of their targets and their targets usually are involved in similar functions (Lemmens, De Bie et al. 2009; Kim, Bhardwaj et al. 2010).

Recently, it was shown that the importance of regulator is not only related to the number of its targets but also its position in the regulatory network (chain-of-command) (Gerstein, Bhardwaj et al. 2010; Kim, Bhardwaj et al. 2010). Therefore, if a regulator regulates another regulator, it adds more value to the importance of the regulator on top. For example, LexA in *E. coli* inhibits sigma factor 70 and some global regulators, thus LexA has a high effect on the expression of a large number of genes in the organism (Gerstein, Bhardwaj et al. 2010; Kim, Bhardwaj et al. 2010). In addition, if some genes are dominantly controlled by a certain regulator (autonomous regulator); it adds more importance to this regulator compared to the genes controlled by a combination of regulators (collaborative regulators). As an example, LexA dominantly controls most of its targets and no other regulator regulates many LexA target genes (Gerstein, Bhardwaj et al. 2010; Kim, Bhardwaj et al. 2010).

Although our knowledge about the regulatory networks have been increased recently, but the modularity inside these networks and the way they evolve during the evolution have not been completely described. Detecting collaborative regulators is essential to find any modularity in regulatory network as these collaborative regulators resemble the combinatorial regulators (Figure 4.1), and observed co-expression of genes can be described by combinatorial regulation.

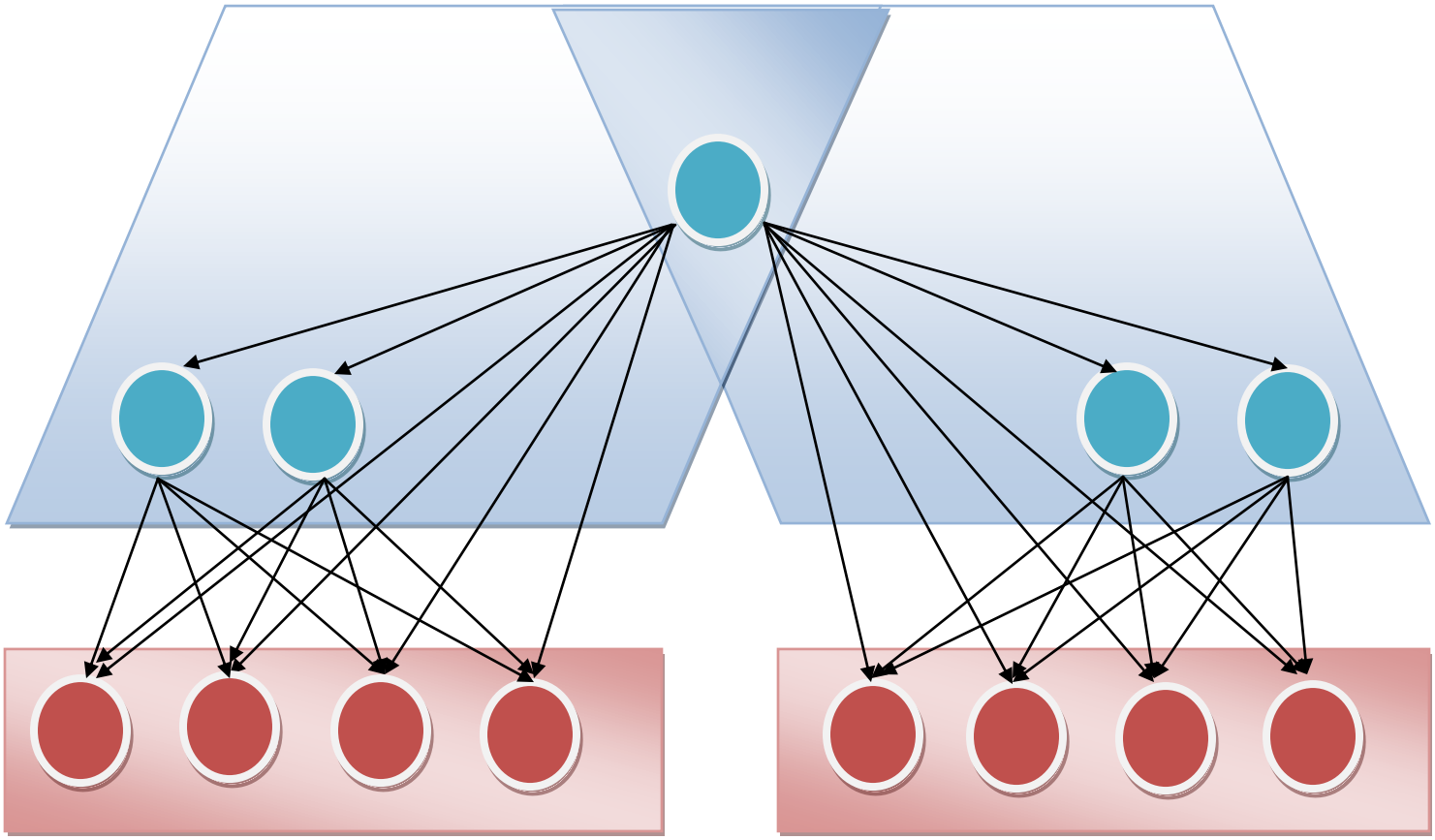


Figure 4.1. The relation between combinatorial regulators, collaborative regulators, and subnet in the regulatory network. A schematic of the regulatory network is presented where each blue circle is a regulator and each red circle is a non-regulator gene, and the regulatory interactions are shown by directed arrows. The two red rectangles illustrate the co-expressed modules. The regulators, enriched as combinatorial regulators, responsible for the co-expression of these two modules are shown as two blue parallelograms, and the node on top is enriched to be regulator of both groups. Unlike combinatorial regulators, collaborative regulators are just detected based on the structure of the regulatory network without considering the co-expression modules. By considering the number of targets that each pair of regulators shares, collaborative regulators can be clustered as two bleu parallelograms. This shows although combinatorial regulators and collaborative regulators are defined in different ways, but in ideal cases where a clear modularity exists in the regulatory network they exhibit similar groups. On the other hand, subnets in the regulatory network demonstrate different concept, as the definition of the subnet is very different. Subnet is defined as the subgraph induced by all nodes downstream of a root node, including the root node. Therefore, the whole network of this example is a subnet in which the regulator in the higher level is the root. Four other subnets also exist in this example, in each of these subnets one regulator in the second level is the root, and

four targets of the regulator are also in the subnet. In other words, subnet can be seen as a root regulator and its direct and indirect targets (this example does not contain any indirect target in any subnet), and the degree of co-expression on the microarray expression compendia for a subnet shows to what extent the root regulator has influence on the expression of its direct and indirect targets.

In this chapter, first we improved the measure to detect the collaborative regulators using Fisher exact test and Monte Carlo sampling methods instead of Jaccard coefficient introduced in (Balaji, Babu et al. 2006). In a similar attempt but in the other direction, we also tried to define a regulatory similarity between each two genes based on their common regulators considering the whole regulatory network structure. This similarity measure obviates the need to detect any modularity in the regulatory network. For each pair of genes, we measured the similarity in regulation by using the importance of their common regulators, and we called our defined similarity measure 'co-regulatory similarity measure'. We studied the mutual relation between this measure and the co-expression of genes on the microarray compendia.

4.2. MATERIALS AND METHODS

4.2.1. REGULATORY NETWORK (TRANSCRIPTIONAL AND POST-TRANSCRIPTIONAL INTERACTIONS)

We downloaded and annotated sigma factors, transcription factors (TFs), and small RNAs (sRNAs) and their targets from RegulonDB database (Gama-Castro, Jimenez-Jacinto et al. 2008).

4.2.2. CO-EXPRESSION MICROARRAY DATA

Like the previous chapters, the microarray compendium of *E. coli* was obtained from (Lemmens, De Bie et al. 2009). We used the Pearson correlation coefficient across all conditions as the measure of co-expression of two genes.

4.2.3. MONTE CARLO SAMPLING IN REGULATORY NETWORK TO ASSESS THE COLLABORATION OF REGULATORS

We considered a regulatory network as a bipartite graph (regulator, target), note that each regulator is represented twice on both side of this bipartite graph as two separate nodes because even the target part of the bipartite graph contains regulators as a regulator can be target of another regulator. Each two regulator may share certain number of their targets, and Fisher exact test can be used to assess the probability of observing equal or higher number of the shared targets the two regulators by chance. Consequently, lower p-value means two regulators are sharing more targets compare to what expected by chance. In other words, lower p-value implies two regulators are more collaborative. Monte Carlo sampling is the less biased way to measure this probability. Fisher exact test is more robust way to measure the p-value when the target number of regulators is more evenly distributed, while we know it is not the case for the regulatory network, and some regulators have large number of target genes while the others regulate few genes.

To perform Monte Carlo sampling, in each step we shuffled two edges in a way that the distribution of degrees in the bipartite graph remain preserved. This shuffling is done by repeatedly selecting at random two edges and crossing them (replacing the targets of two regulators). If two regulators R_i and R_j shares $|T|$ targets, and shuffling procedure was performed n times, we calculate p-value and q-value as follow:

$$p = \frac{number\ of\ times\ sharing\ more\ than\ |T|\ targets}{n}$$

$$q = \frac{|T| - E(common\ targets)}{\min\left(degree(R_i), degree(R_j)\right)}$$

Where function E stands for expected value. In contrast to p-value, higher q-value means two regulators are more collaborative.

One major advantage of using q-value is its faster convergence, and the fact that it can be used directly to find module is another advantage. On the other hand, p-value cannot be used directly and some transformation like using *–log(p-value)* or *1/p-value* is needed because more collaborative regulators gain lower p-values.

4.2.4. PAGERANK VALUE OF REGULATORS TO ASSESS IMPORTANCE OF A REGULATOR

PageRank assigns a numerical weight to each node in a network that measures its relative importance within the network (Page and Brin 1998). PageRank was originally developed for World Wide Web application to empower search engines to search the web pages efficiently. It considers the World Wide Web as a graph where each webpage is a node and each hyperlink is a directed edge. The importance of each webpage is measured by its PageRank value (Page and Brin 1998).

The importance of regulators in a regulatory network can be measured by PageRank value since regulatory network is very similar to World Wide Web. First of all, in World Wide Web important pages receive higher hits. Similarly, in regulatory network, important regulators regulate more targets. Furthermore, in World Wide Web, the importance of each page also relates to the pages that link to them (e.g. page directly linked by Google). In regulatory network if a regulator regulates a more global regulator, this regulator is more important.

We have used the original PageRank algorithm described by Lawrence Page and Sergey Brin (Page and Brin 1998) in several publications. It is given by:

$$PR(A) = (1-d) + d\,(PR(T1)/C(T1) + \ldots + PR(Tn)/C(Tn))$$

Where PR(A) is the PageRank of regulator A, PR(Ti) is the PageRank of target Ti which link to regulator A, C(Ti) is the number of outbound links on target Ti, and d is a damping factor which can be set between 0 and 1. We calculate this value by reversing the direction of regulation in the regulatory network as the importance of a regulator is a result of its target genes.

In World Wide Web an imaginary surfer is considered to traverse the network by randomly clicking links, and eventually this person stops clicking at one step. At any step, the probability

that the person continues clicking is called damping factor d. This damping factor is usually a high number (closer to 1), and we set it to 0.95.

4.2.5. CO-REGULATORY SIMILARITY MEASURE BETWEEN PAIR OF GENES AND PAIR OF MODULES BASED ON PAGERANK SIMILARITY OF COMMON REGULATORS

We define co-regulatory similarity of a pair of genes as the global similarity of their regulators on top of them. We measure the co-regulatory between each two genes G_1 and G_2 based on the PageRank values of their common regulators R_i as follow:

$$Similarity(G_1, G_2) = \sum_{i=1}^{n} \frac{1}{PageRank(R_i)}$$

Where n is the number of common regulators between G_1 and G_2. As it can be seen from the formula, each common regulator on top of the gene pair G_1 and G_2 adds a value to the co-regulatory similarity of them. This value is equal to the inverse of the PageRank of the regulator because more local regulators have lower PageRank value, thus it contributes more to the co-regulation of its targets. In other words, the highly co-regulated genes are more likely to have common local regulator(s) on top.

The defined co-regulatory similarity measure can be easily extended to the module level. For two modules M_1 and M_2 of size l_1 and l_2 we used average similarity of each pair of genes across two modules as co-regulatory similarity measure:

$$Similarity(M_1, M_2) = \frac{\sum_{i=1}^{l_1} \sum_{j=1}^{l_2} Similarity(G_i G_j)}{l_1 \times l_2}$$

Where $G_i \in M_1$ and $G_j \in M_2$ are the genes inside the modules.

4.2.6. FINDING MODULES IN A NETWORK USING OSLOM

OSLOM is a recently developed method to find modules in a network (Lancichinetti, Radicchi et al. 2011). The internal optimization criteria of OSLOM empowers it to detect just densely connected subnetwork. OSLOM operates based on the local optimization of a fitness function

expressing the statistical significance of clusters with respect to random fluctuations, which is estimated with tools of Extreme and Order Statistics (Lancichinetti, Radicchi et al. 2011). As low connected subnetworks cannot gain high statistical significance in comparison to random fluctuations, OSLOM could detect the proper modules for our aim.

As described by the main authors, OSLOM uses the significance as a fitness measure in order to evaluate the clusters by defining the probability of finding the cluster in a random null model, i. e. in a class of graphs without community structure. The configuration model is chosen as the null model. This is a model designed to build random networks with a given distribution of the number of neighbors of a vertex (degree). The networks are generated by joining randomly vertices under the constraint that each vertex has a fixed number of neighbors, taken from the pre-assigned degree distribution.

In comparison to the other methods, OSLOM is superior in detecting overlapping clusters and clusters with hierarchical structure, i.e. a clusters which include (or are included by) other clusters (Lancichinetti, Radicchi et al. 2011). As hub genes (genes with high connectivity) are generally highly presented in various biological networks, and these genes are involved in several biological processes, a suitable clustering method should be able to detect overlapping clusters in which the overlapping nodes are usually the hub genes. In addition, many biological networks have hierarchical structure. For example, in metabolic network a number of metabolic pathways can be part of a superpathway, or in protein-protein interaction network a combination of protein complexes may construct a larger structural complex.

4.3. RESULTS

4.3.1. DETECTING COLLABORATIVE REGULATORS

If we can detect modules consisting of collaborative regulators, we can verify the co-expression of the targets of each module. If we observe modularity in the regulatory network and co-expression of the target genes of the regulators in one module, it means that the modules of collaborative regulators were indeed acting combinatorially.

We tried three different measures to detect collaborative regulators in *E. coli*, and consequently deciphering modularity in the regulatory network of *E. coli*. First instead of the simple Jaccard coefficient, we calculate the more sensitive Fisher exact test p-value to calculate to what extent two regulators mutually share their target genes, considering the regulatory network as a bipartite graph (regulator- target). This Fisher exact test computes the probability of observing the number of common targets with respect to the number of genes that each regulator regulates and the size of the whole genome. After calculating p-values for each pair of regulators, we observed that if we plot log transform of p-values derived from Fisher exact test and log transform of regulator number, the outcome would be a straight line (Figure 4.2). This may imply that log transform of the p-values may have power law distribution, thus it possibly will be hard to cluster the regulators based on their common targets.

We have also developed two measures (p-value and q-value) based on Monte Carlo sampling method to sample a bipartite graph (regulator - target) with the same node degree (see 4.2.3). Here, the idea is to build a network consists of only regulators as nodes, and edges in this network may represent to what extent two regulators are collaborative. We built the network by using q-value or transformed p-values as the value for the edges. We applied OSLOM method to find modules in these networks, and OSLOM could not detect high modularity in any of the networks. It may imply regardless of which measure we use (q-value or transformed p-value) that it is difficult to find modules in the network.

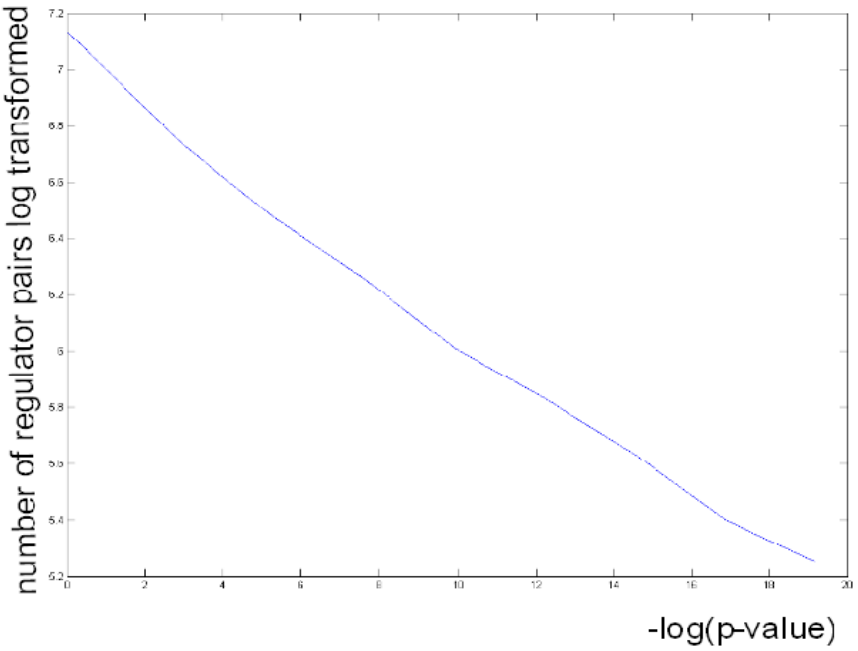


Figure 4.2. -log transformation of Fisher exact test p-values calculated for each regulator pairs by comparing the observation common targets to what can be expected by chance, versus the logarithm regulator pairs number.

4.3.2. CO-REGULATORY SIMILARITY A MEASURE TO PREDICT CO-EXPRESSION

We used the PageRank measure to capture the importance of regulators in a regulatory network. This measure can fulfill both the idea of chain-of-command, as well as the target number because as entailed in the formula each target adds a value to the PageRank of the regulator (target number, and the targets with higher PageRank add higher value to the PageRank for the regulator (chain-of-command). In *E. coli*, regulators with higher PageRank are sigma factors (RpoD, RpoE, RpoS, RpoN, RpoH), global transcription factors (e.g. Crp, ArcA, Fnr), and transcription factors with a high value due to the chain-of-command idea like LexA, since LexA inhibits some global regulators and RpoD sigma factor 70. In other words, although LexA does not regulate large number of targets, regulating important targets has sharply increased its PageRank value. In addition, normal stress related target genes of LexA do not have many regulators which cause their high contribution to the PageRank value of LexA as in the formula the contribution of targets are divided between all the regulators on top of them. Because of this property of the

formula, more autonomous regulators as defined in (Gerstein, Bhardwaj et al. 2010; Kim, Bhardwaj et al. 2010) gain higher PageRank than collaborative regulators as defined in (Gerstein, Bhardwaj et al. 2010; Kim, Bhardwaj et al. 2010) gain.

According to how we define the co-regulation of genes (see 4.2.5), two genes are more co-regulated if they have more regulators in common and their common regulators are more local (lower PageRank). Therefore, we define co-regulatory similarity between two genes as the sum of the inverse PageRank value of their common regulators (see 4.2.5). This value is directly related to random walk starting from one gene and tracing the regulatory network to reach the second gene. This measure can be easily expanded to calculate the regulatory similarity between two groups of genes (see 4.2.5).

Since we have defined co-regulatory similarity as similarity in the regulatory circuits controlling the genes, the co-regulatory similarity between two genes must be observable on microarray expression compendia. Therefore, the detected co-regulatory similarity should be traceable as co-expression on microarray data to prove the accuracy of our proposed co-regulation similarity measure. Pearson correlation coefficient across all conditions measures co-expression between two genes. Lack of conditions in expression data and also unmeasured regulatory interaction can reduce the credibility of comparison between co-expression and our defined co-regulatory similarity measure. Therefore, we excluded the genes, without at least one common regulator or a Pearson correlation co-expression value over 0.5 with at least one other gene, from our analysis. Figure 4.3.A represents the distribution of co-regulation similarity between all selected pairs of genes, defined based on PageRank values of their common regulators, and Figure 4.3.B shows the distribution of Pearson correlation coefficient across all conditions on microarray data between all selected pairs of genes. Now the question is to show that the gene pairs with the highest co-regulatory similarity value (Figure 4.3.A right tail) exhibit high co-expression (Figure 4.3.B both right and left tails).

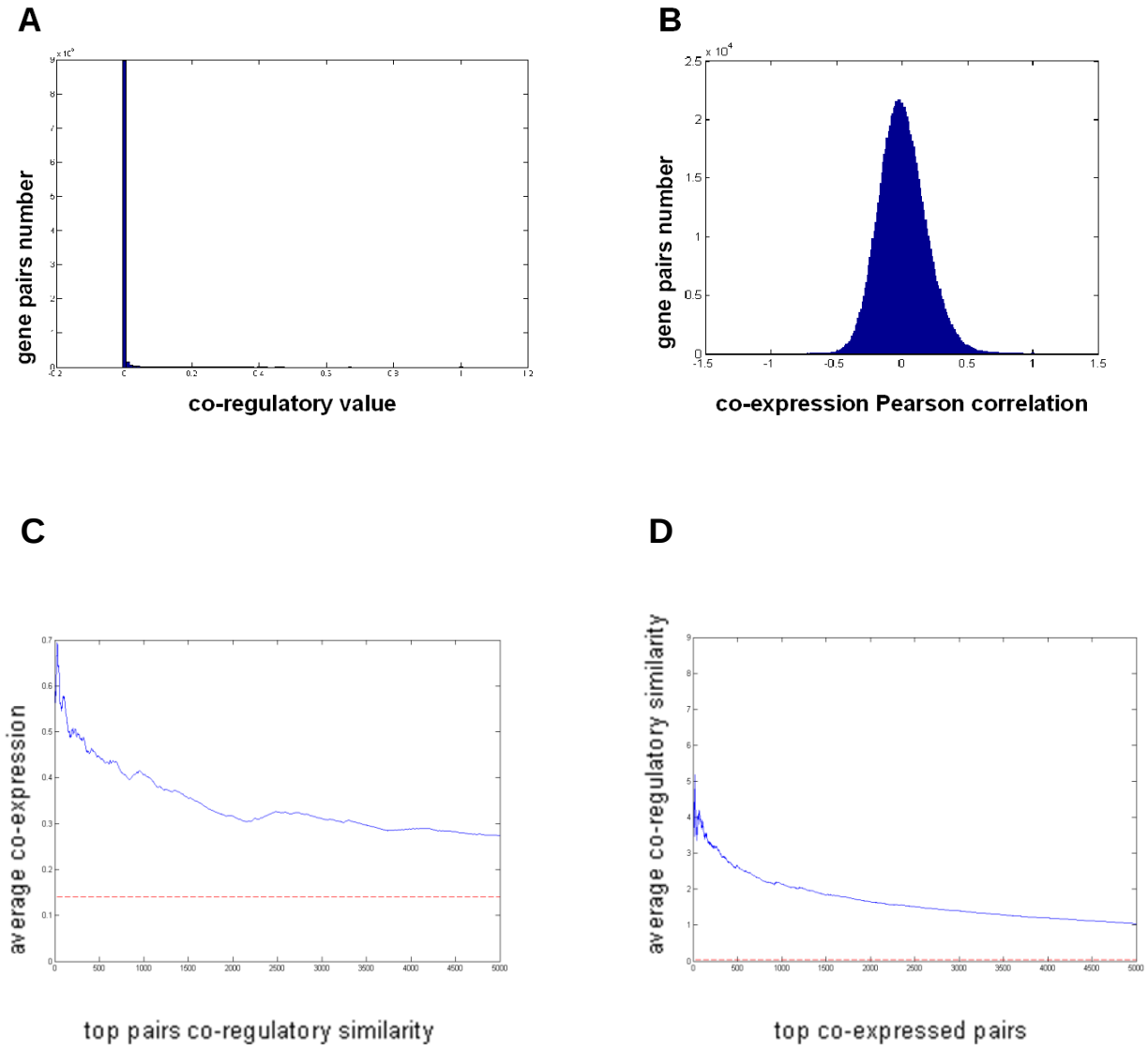


Figure 4.3. Co-expression versus co-regulatory similarity. **A.** Distribution of co-regulation similarity between all selected pairs of genes, defined based on PageRank values of their common regulators. **B.** Distribution of Pearson correlation coefficient across all conditions on microarray data between all selected pairs of genes. Genes without at least one common regulator or Pearson correlation co-expression value over 0.5 with at least one other gene were excluded from the analysis. **C.** Average co-expression of gene pairs with the highest co-regulatory similarity (blue solid curve) versus average co-expression of all genes pairs (dashed red line). The number of gene pairs with the highest co-regulatory similarity is represented on the X-axis, and the average of co-expression of the corresponding top gene pairs with the highest co-regulatory similarity is represented on the Y-axis. The absolute value of Pearson correlation across all conditions was used as the measure for the co-expression. **D.** Average co-regulatory similarity of gene pairs with the highest co-expression (blue solid curve) versus average co-regulatory similarity of all genes

pairs (dashed red line). The number of gene pairs with the highest co-expression is represented on the X-axis, and the average of co-regulatory similarity of the corresponding top gene pairs with the highest co-expression is represented on the Y-axis.

Figure 4.3.C shows that gene pairs that gain higher co-regulatory similarity based on our defined measure exhibit much higher co-expression than an average gene pair, and the genes with higher co-regulatory similarity exhibits higher co-expression as the blue curve decreases as more gene pairs added. Here the absolute value of Pearson correlation across all conditions was used as the measure for the co-expression. Similarly, as it was shown in figure 4.3.D gene pairs, exhibiting higher co-expression, gain higher co-regulatory similarity value.

Kolmogorov-Smirnov statistic is the standard statistic to compare the order of two lists and was used to interpret gene expression data in combination with other data sources such as phenotype (Subramanian, Tamayo et al. 2005; Keller, Backes et al. 2007). We used the Kolmogorov-Smirnov test to verify that the gene pairs with large co-regulatory similarity values exhibit higher co-expression using the absolute value of Pearson correlation coefficient as the co-expression measure. In fact, if the entire gene pairs with non-zero co-regulatory similarity is chosen (pairs with at least one common regulator), Kolmogorov-Smirnov test verify that these pairs exhibit higher co-expression than what is expected by chance with the significance level 5%. This means if two genes contain a common regulator they exhibit higher co-expression than two random genes which is expectable by the definition of the gene regulation, while comparison between our defined co-regulatory similarity measure and co-expression demonstrate stronger relation than what can be deduced by Kolmogorov-Smirnov test (Figure 4.3.C and 4.3.D).

4.4. DISCUSSION

The concept collaborative regulators may have direct relation with the combinatorial regulatory which is responsible for observed co-expression on the microarray compendia. In this study we introduced new measures, Fisher exact test and also Monte Carlo sampling, to detect collaborative regulators. These measures are more sensitive than the originally proposed Jaccard coefficient (Balaji, Babu et al. 2006; Gerstein, Bhardwaj et al. 2010) because they provide a p-value of observing the number of shared targets for two regulators compare to what is expected by chance. However, we could not successfully apply any of these methods to highlight collaborative regulators which could decipher the regulatory network modularity in *E. coli*.

As we failed to highlight the collaborative regulators, which can describe the observed co-expression, we introduced a regulatory similarity measure for a pair of gene considering the whole structure of the regulatory network. The introduced regulatory measure for a pair of genes considers the importance of common regulators. This importance is related to the position of the regulator in the hierarchy and was measured by using PageRank method. The higher regulators in the hierarchy are more global regulators and have more targets while the lower regulators are the more local ones. One advantage of using PageRank is that it can also fulfill the concept of chain-of-command (Gerstein, Bhardwaj et al. 2010; Kim, Bhardwaj et al. 2010) in the regulatory network.

Using our definition of co-regulatory similarity measure for a pair of genes, we could observe the direct relation between the regulatory network and the observe co-expression on microarray expression compendia. As two co-expressed genes are probably involved in the similar biological processes, we can expect genes with high co-regulatory similarity value are also involved in similar function. Therefore, the defined co-regulatory similarity measure can be seen as a suitable measure for data integration.

One major advantage of our defined co-regulatory similarity measure is that it obviates the need to detect any modularity in the regulatory network. Therefore, the observed co-expression between a group of genes on the microarray expression compendia can be explained as the effect of the whole regulatory network, and not a certain group of regulators. The fact that the

regulatory network may not be modular is not in contrast to the combinatorial regulatory concept. Based on our definition of the co-regulatory similarity measure, a group of genes which have higher co-regulatory similarity are more likely to have common regulators and some local regulator are expected to be among these common regulators. Nevertheless, our definition of co-regulatory similarity does not imply that a group of co-expressed genes has to share the majority of their common regulator. In contrast, finding modularity in the regulatory network based on collaborative regulators contain this hidden presumption that there exists a detectable modularity in the regulatory network that have a one to one relation with the co-expression modularity of genes (as an example consider schematic Figure 4.1).

Considering our defined co-regulatory similarity measure, we can explain the fast evolution of the regulatory network. This network should evolve much faster in comparison to other interaction networks and cellular pathways to enable the organisms to adapt themselves to the new environment, but what should remain conserved is the co-regulatory similarity of the genes. Considering our definition of co-regulatory similarity measure, we can explain why the majority of rewiring happens in higher regulators in the hierarchy (more global regulators) (Jothi, Balaji et al. 2009) because the targets of lower level regulators in the hierarchy, or local regulators, gain higher co-regulatory similarity value. Consequently, the target genes of local regulators are more likely to be co-expressed. On the other hand, more global regulator seem to assist the conveying signals from the environment to the cell, but have less effect on the observed co-expression of the genes.

CHAPTER 5
THE RELATION BETWEEN PHYSICAL INTERACTION NETWORKS AND FUNCTIONAL DATA SOURCES: APPLICATION TO THE *E. COLI* GENOME

5.1. INTRODUCTION

Integrating different types of data including physical interactions (protein-protein, protein-DNA, etc.), genetic interactions (synthetic sickness and synthetic lethality), and expression data can lead to a better functional annotation of genes and a better understanding of the cell behavior (Kelley and Ideker 2005; Beyer, Bandyopadhyay et al. 2007; Andres Leon, Ezkurdia et al. 2009; Hu, Janga et al. 2009; Huang and Fraenkel 2009).

Integrating different data sources derived from high-throughput to assign new function to genes with unknown genes have been applied over different species especially *E. coli* (Andres Leon, Ezkurdia et al. 2009; Hu, Janga et al. 2009) and yeast (Zhu, Zhang et al. 2008; Myers, Chiriac et al. 2009; Narayanan, Vetta et al. 2010). In these studies, the mutual relations between different data sources were carefully explored. The study of (Zhu, Zhang et al. 2008) is a good example of combining different data sources including genotypic, expression, transcription factor binding site (TFBS), and protein-protein interaction network (PPI) data from a number of yeast experiments. They observed that some protein complexes are either traceable as a clique in protein-protein interaction network, or as highly co-expressed genes in microarray expression compendia. Therefore, they concluded the complementary relation between protein-protein interaction network and microarray expression data. As another example, (Kelley and Ideker 2005) studied the mutual relation between the combined network of physical interactions (protein-protein and protein-DNA) and genetic interactions, and they concluded that genetic interactions occur mostly between different pathways, retained from physical interaction data, rather than within pathways. As another example, in (Chandrasekaran and Price 2010), it was shown that using transcriptional regulatory network improve the metabolic network prediction in *E. coli*. To predict metabolic changes that result from genetic and environmental perturbation,

they integrated transcriptional regulatory network with corresponding metabolic network using high-throughput measured data (Chandrasekaran and Price 2010). The high accuracy of their result in constructing regulatory-metabolic network in *E. coli* made their constraint-based probabilistic modeling proper to study less studied organisms like *M. tuberculosis* (Chandrasekaran and Price 2010).

Predicting the behavior of cell and its response to perturbations (Ishii, Nakahigashi et al. 2007) with respect to the cellular architecture of regulatory circuits (Cho, Zengler et al. 2009; Thiele, Jamshidi et al. 2009) as well as signaling pathways (Hyduke and Palsson 2010) is another major application of data integration. These kinds of studies can reveal the repeating pattern (motifs) over interaction circuits of networks with controlling role such as regulatory network and phosphorylation network which are evolutionary more favorable to control cellular behavior(Michoel, Joshi et al. 2011), and the way that these circuits are developed to handle perturbations (Ishii, Nakahigashi et al. 2007). For example, (Ishii, Nakahigashi et al. 2007) attempted to assess if there is a redundancy structure of the metabolic network or if there are multiple regulatory circuits in *E. coli* to address perturbations. They observed in response to enzymatic gene disruptions (mutation), metabolic levels remains stable, and they could conclude that the reason of this stability is due to the parallel routes in metabolic network, and when one route is not available due to the mutation of enzyme in this route, the parallel route is still accessible. In contrast, they observed in response to changes in growth rate perturbation (exposing high glucose), *E. coli* regulates enzyme levels to maintain a stable metabolic state, implying that changes in regulation is the complementary strategy that *E. coli* uses to keep metabolic network robust to response to the environmental perturbations.

Although current data integration methods based on network could predict new function for many genes of different genome successfully (Myers, Chiriac et al. 2009; Narayanan, Vetta et al. 2010) or successfully described the organism response to stimuli (Ishii, Nakahigashi et al. 2007; Yeger-Lotem, Riva et al. 2009), still the mutual relation between physical interaction networks with controlling role inside the cell such as regulatory network and other physical interaction networks and also other functional data sources is not completely explored. The regulatory network evolves rapidly to enable different species to adapt themselves to various environmental

conditions, but they also control the expression of non-controlling networks inside the cell, and these non-controlling networks such as protein-protein interactions and metabolic pathways are evolutionary conserved in close species (Shou, Bhardwaj et al. 2011). The fact that the regulatory network evolves differently from the other networks, which are controlled by this network, makes it difficult to deduce functional relation between genes from the regulatory network. For example, more global regulators have been shown to regulate genes with completely different biological functions (Jothi, Balaji et al. 2009; Kim, Bhardwaj et al. 2010). In addition, it not clear to what extent the regulatory hierarchy, which is derived from the regulatory network, is similar to the functional hierarchy of the genes, which is defined by GO terms hierarchy. The hierarchy which can be derived from a regulatory network may be highly dynamic as the regulatory networks evolve rapidly. In contrast, the hierarchy which can be observed using GO terms seems to be static as it is directly related to the GO terms hierarchy, and the hierarchy of GO terms is a fixed hierarchy for all species.

In this chapter first, we assessed the relation between networks with controlling role and non-controlling roles. We used our co-regulatory similarity measure to evaluate the regulatory similarity of genes, which are involved in the same biological processes. We considered genes in two types of modules as the genes which are involved the same biological processes; genes involved in the same metabolic or signaling pathways, and also genes which are detected as modules in the protein-protein interaction network. We also compared the regulatory hierarchy of the genes, which are involved in the same biological processes, with their functional hierarchy. We used the co-regulatory similarity measure at the module level (see 4.2.5) to build the hierarchy of modules. To measure the functional similarity across different biological processes, we introduced a new species-specific functional relation measure between each two modules based on their shared gene ontology terms.

5.2. MATERIALS AND METHODS

In this part we discuss the available high confident data sources available in *E. coli*. Note that unlike yeast, the phosphorylation and genetic interaction networks are not available for *E. coli*.

5.2.1. CURRENT AVAILABLE PHYSICAL INTERACTION DATA SOURCES IN E. COLI

5.2.1.1. Protein-protein interactions network

We downloaded 7603 interactions from combined interaction dataset provided by (Peregrin-Alvarez, Xiong et al. 2009) from the following link:

http://www.compsysbio.org/bacteriome/download.php

The combined interaction dataset was derived from combining the (Hu, Janga et al. 2009) (3888 interactions between 918 proteins with high confidence score) and functional interactions (Peregrin-Alvarez, Xiong et al. 2009) (3989 interactions between 1941 proteins deduced by Peregrin-Alvarez). The later one was built by combining 3 experimental datasets (large scale pull down (Arifuzzaman, Maeda et al. 2006), small scale assays (Xenarios, Salwinski et al. 2002; Salwinski, Miller et al. 2004), large scale Tap (Butland, Peregrin-Alvarez et al. 2005)) and 6 computational datasets (conserved co-expression (Bergmann, Ihmels et al. 2004), phylogenetic profiles (Bowers, Pellegrini et al. 2004), literature mining (Hoffmann and Valencia 2005), gene proximity (Bowers, Pellegrini et al. 2004), Rosetta stone (Bowers, Pellegrini et al. 2004), Interlogs (Rain, Selig et al. 2001; Gerstein, Yu et al. 2004)). Although for some of these functional interactions there is no evidence of direct interaction, but there is fairly high confidence that they are involved in same biological process.

We combined those 7603 interactions provided by (Peregrin-Alvarez, Xiong et al. 2009) with 1528 interactions derived and annotated from literature-based curated EcoCyc (Keseler, Bonavides-Martinez et al. 2009) protein complexes. This resulted eventually in 8454 high-confidence protein-protein interactions.

5.2.1.2. Regulatory network: Transcriptional and post-transcriptional interactions

Like previous chapters, we considered transcriptional and post-transcriptional interactions as the regulatory network. We downloaded sigma factors, transcription factors (TFs), and small RNAs (sRNAs) and their targets from RegulonDB database (Gama-Castro, Jimenez-Jacinto et al. 2008). Unfortunately unlike yeast, post-translational interactions such as phosphorylation interactions are not available for *E. coli*.

5.2.1.3. Metabolic and signaling pathways

Like chapter 2, we downloaded highly accurate literature-based curate metabolic and signaling pathways (mostly two-components pathways) from EcoCyc (Keseler, Bonavides-Martinez et al. 2009).

5.2.2. FUNCTIONAL DATA SOURCES

5.2.2.1. Gene Ontology terms

Like previous chapters, Gene Ontology (GO) terms of *E. coli* were downloaded from EcoCyc (Keseler, Bonavides-Martinez et al. 2009).

5.2.2.2. Co-expression microarray compendia

Like in previous chapters, the microarray compendium of *E. coli* was obtained from Lemmens et al. (Lemmens, De Bie et al. 2009). We used the Pearson correlation coefficient across all conditions as the measure of mutual co-expression between genes. To measure the co-expression of genes inside a module, we calculated the average of these Pearson correlation coefficients of all gene pairs in the module.

5.2.3. JACCARD SIMILARITY COEFFICIENT

The Jaccard coefficient measures similarity between two modules of genes is defined as the number of common genes divided by the total number of the genes:

$$\frac{G_i \cap G_j}{G_i \cup G_j}$$

Where Ga is the set of genes in module a.

5.2.4. DETECTING MODULES IN EACH PHYSICAL INTERACTION DATA SOURCE

Different physical interaction data sources are available for *E. coli*. Each one measures a certain type of interaction. Metabolic (292 modules) and signaling modules (22 modules) were directly downloaded from EcoCyc pathway database (Keseler, Bonavides-Martinez et al. 2009).

A recently developed community detection method called, OSLOM (Order Statistics Local Optimization method) (Lancichinetti, Radicchi et al. 2011), was applied over high-confidence PPI network and could detect 114 modules.

To find modules consisting of genes with similar regulators, OSLOM was applied over the co-regulatory similarity values of all pair of genes and detects 68 modules.

5.2.5. FUNCTIONAL SIMILARITY MEASURE BETWEEN TWO MODULES

We assign a vector *V* with a length equal to the number of GO terms to each module. For *i* th module, the *j* element related to the *j* th GO term is calculated as: $V_i(j) = \frac{G_i \cap G_j}{G_j}$

Then if the *j* th GO term includes large number of genes, we expect a low value for $V_i(j)$.

We define similarity between two modules l and k based on cosine similarity of their vectors as follow:

$$Similarity(V_l, V_k) = \frac{V_l \cdot V_k}{|V_l| \times |V_k|} = \frac{\sum_{i=1}^{n} V_l(i) \cdot V_k(i)}{\sqrt{\sum_{i=1}^{n} V_l(i)^2} \times \sqrt{\sum_{i=1}^{n} V_k(i)^2}}$$

Where n is the total number of GO terms. For example, consider just five Go terms exist, and each one includes (1000, 50, 25, 10, 2) genes. Imagine two gene modules including 20 and 15 genes and they share (20, 19, 15, 6 ,1) and (10, 8 ,5 ,3, 0) genes with the GO terms. Then:

$V_1 = (20/1000, 19/50, 15/25, 6/10, 1/2)$

$V_2 = (10/1000, 8/50, 5/25, 3/10, 0/2)$

And $Similarity(V_1,V_2) = 0.3610 / (1.0558 * 0.3946) = 0.8665$

The GO terms, including only one gene, are not indicating any functional relation and consequently are not informative and are excluded from this analysis.

5.3. RESULTS

5.3.1. STUDYING MUTUAL RELATION BETWEEN PHYSICAL INTERACTION DATA SOURCES AND FUNCTIONAL DATA SOURCES

To gain a global view on mutual relation between various interaction networks and functional hierarchy of GO terms in *E. coli*, we perform two steps summarized in Table 5.1. First, we studied the mutual relation between the regulatory network and modules involved in a certain biological process (Table 5.1 task 1). We considered the modules derived from protein-protein interaction network (PPI) and each pathway metabolic or signaling pathway as a module involved in a certain biological process. We checked the similarity in regulatory circuits (see 4.2.5) of the genes involved in a certain biological module using our defined co-regulatory similarity measure. Later, we built the regulatory hierarchy of the modules involved in the same biological processes, and we also built the functional hierarchy of the modules involved in the biological processes using our defined species-specific functional similarity measure of modules (task 2). The comparison of these two hierarchies could reveal the degree of similarity in the regulatory hierarchy and functional hierarchy in *E. coli*.

Task 1: Studying the mutual relation between the regulatory network and non-controlling interaction networks

- *Detecting biological modules in non-controlling networks (Modules in PPI network + cellular pathways)*
- *Checking the average co-regulatory similarity values of genes pairs in the same module*
- *Detecting modules with high co-regulatory similarity values using the regulatory network*
- *Comparison between the detected biological processes modules and modules with high co-regulatory similarity values (derived from regulatory network)*

Task 2: Comparing the hierarchy of biological process modules built using regulatory networks with the one built using functional GO terms

- *Build the hierarchy of biological process modules by using co-regulatory similarity measure for modules*
- *Build the hierarchy of biological process modules by using the defined functional similarity measure for modules*
- *Compare two hierarchy*

Table 5.1. Overview of the different analysis performed in this chapter. First, we studied the mutual relation between the regulatory network and non-controlling interaction networks (task 1). Later, we built the regulatory hierarchy of the modules involved in the same biological processes, and we also built the functional hierarchy of the modules involved in the biological processes using our defined species-specific functional similarity measure of modules (task 2).

5.3.2. DETECTING MODULES OF GENES INVOLVED IN THE SAME BIOLOGICAL PROCESSES

(Lee, Gianchandani et al. 2008) investigated the mutual relation between genes involved in signaling, metabolic, and transcriptional regulatory networks in yeast using dynamic flux balance analysis (Lee, Gianchandani et al. 2008). Inspired by (Lee, Gianchandani et al. 2008), we explored the mutual relation of the physical interactions of genes involved in four biological processes including regulatory interactions, signaling pathways, metabolic pathways, structural and functional components.

Regulatory interactions consist of transcriptional, post-transcriptional, and post-translational interactions and control every cellular process and regulate all other categories. Unlike yeast, the post-translational interactions such as phosphorylation interactions are not available for *E. coli*. As we mentioned in the previous chapter, we were not able to detect modularity in the regulatory network (see 4.3.1). For *E. coli* metabolic and signaling pathways are available in EcoCyc

(Keseler, Bonavides-Martinez et al. 2009), and we considered each pathway or super-pathway as a module. We defined a fourth category as "structural and functional components" to contain the genes involving in different biological processes such as building large components (e.g. flagellum or ribosomal big and small subunits) or transport activities. Based on this definition, the building blocks of these "structural and functional components" are interacting proteins, and protein-protein interaction network is a proper data source to detect them.

A measured protein-protein interaction can be the interaction between: **1.** Two transcription factors that regulates their targets together **2.** A kinase and substrate (phosphorylation activity) **3.** A protein complex constituting an enzyme **4.** Structural or functional component. Here we are more interested to detect modules in protein-protein interaction networks consisting of structural or functional components. A structural or functional component is a densely connected subnetwork in protein-protein interaction network, hence a module detection method which can highlight more densely connected subnetworks is more suitable for our framework. Therefore, we used OSLOM (Lancichinetti, Radicchi et al. 2011) to find modules in protein-protein interaction network as this method detect densely connected subnetworks due to its internal optimization criteria.

Naturally the detected modules in protein-protein interaction networks are from two last categories, a protein complex encoding an enzyme and structural or functional component, because two interacting transcription factors or kinase and substrate (phosphorylation activity) are low connected nodes in protein-protein interaction network by their nature. To show that protein complexes encoding enzymes were not highly appeared in OSLOM results, we compared the detected modules by OSLOM with EcoCyc cellular metabolic and signaling pathways. Figure 5.2 shows that the detected modules in protein-protein interaction network do not share many genes with cellular metabolic and signaling pathways. In other words, genes which are encoding enzymes are not among the modules detected in protein-protein interaction network. The reason can be these genes have lower connectivity on average than genes in protein-protein interaction network. Average connectivity in our protein-protein interaction network is 1.8904 while average connectivity genes encoding enzymes in this network is 1.3447 (Self loops were not considered).

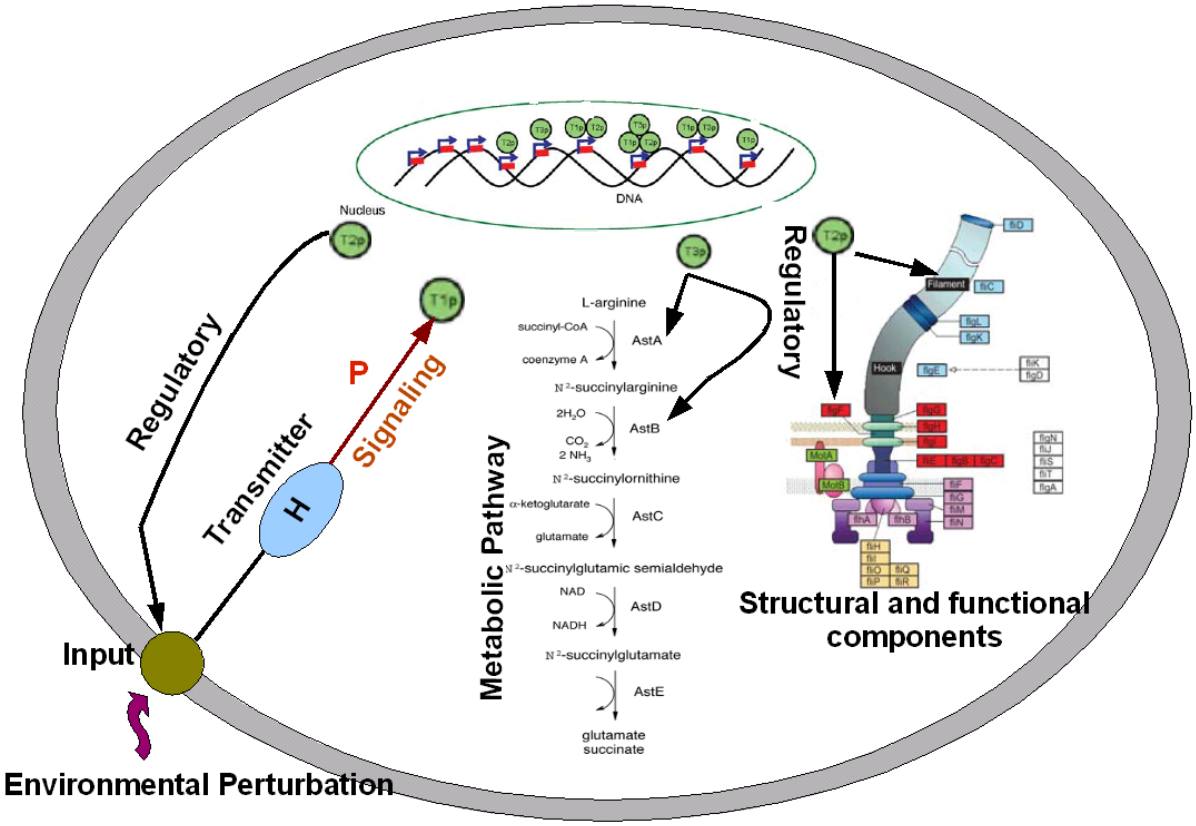


Figure 5.1. Four biological processes categories inside a cell. 1. Regulatory interactions consist of transcriptional, post-transcriptional, and post-translational interactions and control every cellular process and regulate all other categories. **2. Signaling pathways** which may be activated by an environmental change and usually activates certain regulators. **3. Metabolic pathways** where each pathways is a series of chemical reactions occurring within a cell. **4. Structural and functional components** accomplish a certain function (e.g. here flagellum is responsible for locomotion).

We have also compared the modules derived from protein-protein interaction network with the modules which were derived from protein-protein interaction network of (Hu, Janga et al. 2009) and the original modules (Peregrin-Alvarez, Xiong et al. 2009) both using Markov Cluster Algorithm (MCL) clustering. Our detected modules were highly different from that of (Hu, Janga et al. 2009), but rather similar to the one of (Peregrin-Alvarez, Xiong et al. 2009) because we used protein-protein interaction network of (Peregrin-Alvarez, Xiong et al. 2009). As it was reported in (Peregrin-Alvarez, Xiong et al. 2009) most of the connections reported in from (Hu,

Janga et al. 2009) are functional rather than actual physical protein-protein interaction causing large difference in the modules derived from these two data source, and Peregrin-Alvarez could show that Peregrin-Alvarez, Xiong et al. 2009) data source and the detected modules are more reliable, as they have more actual physical interaction and also they integrated more functional data sources to predict interacting proteins. Interestingly, the detected OSLOM modules with the MCL-based (Peregrin-Alvarez, Xiong et al. 2009) original modules show high similarity (Figure 5.3). Our qualitative analysis shows that modules of OSLOM are usually larger and included more than one MCL module which are related to each other (see Appendix A). In addition, it can cluster highly connected proteins (hubs) in more than one module while in many cases MCL failed to do it (see Appendix A). The reason behind these dissimilarities between the modules detected by OSLOM and MCL is the actual differences of the algorithms, MCL is flow based network clustering method, while OSLOM detects subnetworks which are densely connected and highly different from what is expected by chance. OSLOM results does not include low connected proteins which prevents finding paths on network while it enables OSLOM to find highly connected structural and functional component.

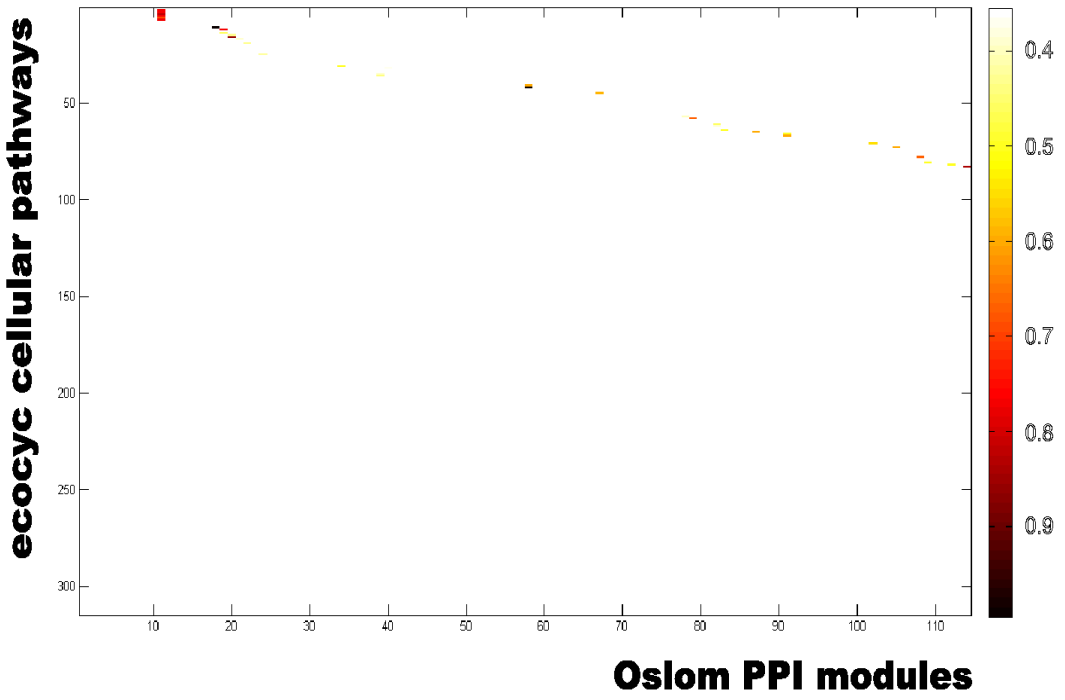


Figure 5.2. The Jaccard coefficient was used to compare modules detected by OSLOM in the protein-protein interaction network with cellular signaling and metabolic pathways derived from EcoCyc. Modules detected in the protein-protein interaction network are shown on the X-axis. These modules were ordered based on the order in which there were organized by OSLOM. On Y-axis EcoCyc cellular pathways are presented, and these pathways were ordered in a way that pathways with larger Jaccard coefficient values would appear on the diagonal. The intensity of dots on the heatmap is related to the value of Jaccard coefficient. As it can be seen in this heatmap, the OSLOM PPI modules do not have high Jaccard coefficient with cellular pathways.

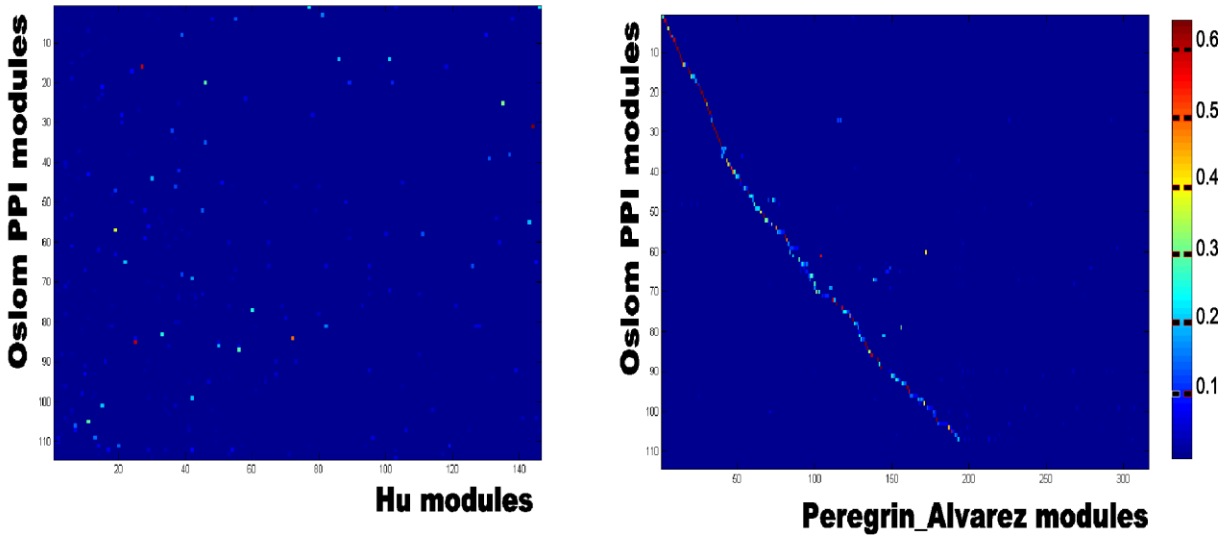


Figure 5.3. The Jaccard coefficient was used to compare OSLOM PPI modules with (Hu, Janga et al. 2009) (left panel), and (Peregrin-Alvarez, Xiong et al. 2009) PPI modules (right panel). The intensity of dots on each heatmap shows the Jaccard coefficient **Left panel:** There are few modules with high Jaccard coefficient and as the dots are scattered all around the heatmap, and we could not reorder them on a diagonal implying the results of from (Hu, Janga et al. 2009) is completely different as the data sources were completely different. **Right panel:** Comparison of the modules detected by OSLOM with from (Hu, Janga et al. 2009) original modules (applying MCL algorithm over the same data source) shows a close relation between the modules. We reordered the modules in both dimension to bring the highly related ones on the diagonal. As it can be seen on the heatmap the majority of OSLOM modules have large Jaccard coefficient with from (Hu, Janga et al. 2009) modules. OSLOM did not find modules with lower connectivity.

5.3.3. EXPLORING THE MUTUAL RELATION BETWEEN GENES INVOLVED IN SIMILAR BIOLOGICAL PROCESSES AND THE REGULATORY NETWORK

The regulatory network controls the expression of genes involved in similar biological processes such as "structural and functional modules" and cellular pathways. Therefore, we expect high similarity in the regulatory interactions controlling the genes which are involved in the similar biological processes. In this part first we use our defined co-regulatory similarity measure and microarray expression compendia to show that genes, involved in the same biological processes, are more likely to have similar controlling regulatory interactions circuits and tend to be co-

expressed with each other. After that we also show that genes, which have more similarity in controlling regulatory interactions, are the ones involved in similar biological processes.

We calculated the average Pearson correlation coefficient across all conditions of gene pairs involved in the same biological process as the co-expression measure (Figure 5.4). Similarly, we used our defined co-regulatory similarity measure to assess the similarity in controlling regulatory circuits of the modules (Figure 5.5). As we expected we could observe both the average co-expression and average co-regulatory similarity are much higher than what can be expected for a random module. Although 36% of signalling pathways, 24% of metabolic pathways, and 20% of PPI modules did gain not average co-regulation higher than the expected co-regulation value 0.008, the average co-regulatory value was striking higher than average for the majority of the modules (Figure 5.5). The major limitation in using co-regulatory similarity measure is the lack of known regulatory interactions for many genes, as all the regulatory interactions exists in the regulatory interaction network of *E. coli* was not measured with high-confidence experiences yet, and if we do not identify any regulators for the genes involved in a biological module (which was the case for almost all of the modules exhibits lower co-regulatory similarity than random) the assigned co-regulatory similarity value to this module will remain zero.

The fact that genes involved in a similar biological process are more likely to be co-expressed and their controlling regulatory interaction circuits are very similar is not a new finding, and we need to show the same phenomena in the other direction meaning that with a given regulatory interaction network, we can still detect the similar modules of genes which are involved in the same biological processes. For this aim, we also used OSLOM to detect modules just using the regulatory network based on the introduced co-regulatory similarity measure, and we compared these modules to the modules in which genes were involved in the same biological processes by using Jaccard coefficient. Interestingly we could observe high similarity between these two types of modules (Figure 5.6). As it can be seen in Figure 5.6, modules detected from regulatory network are larger in size and one of them can be linked to more than one corresponding modules with genes involved in the same biological process.

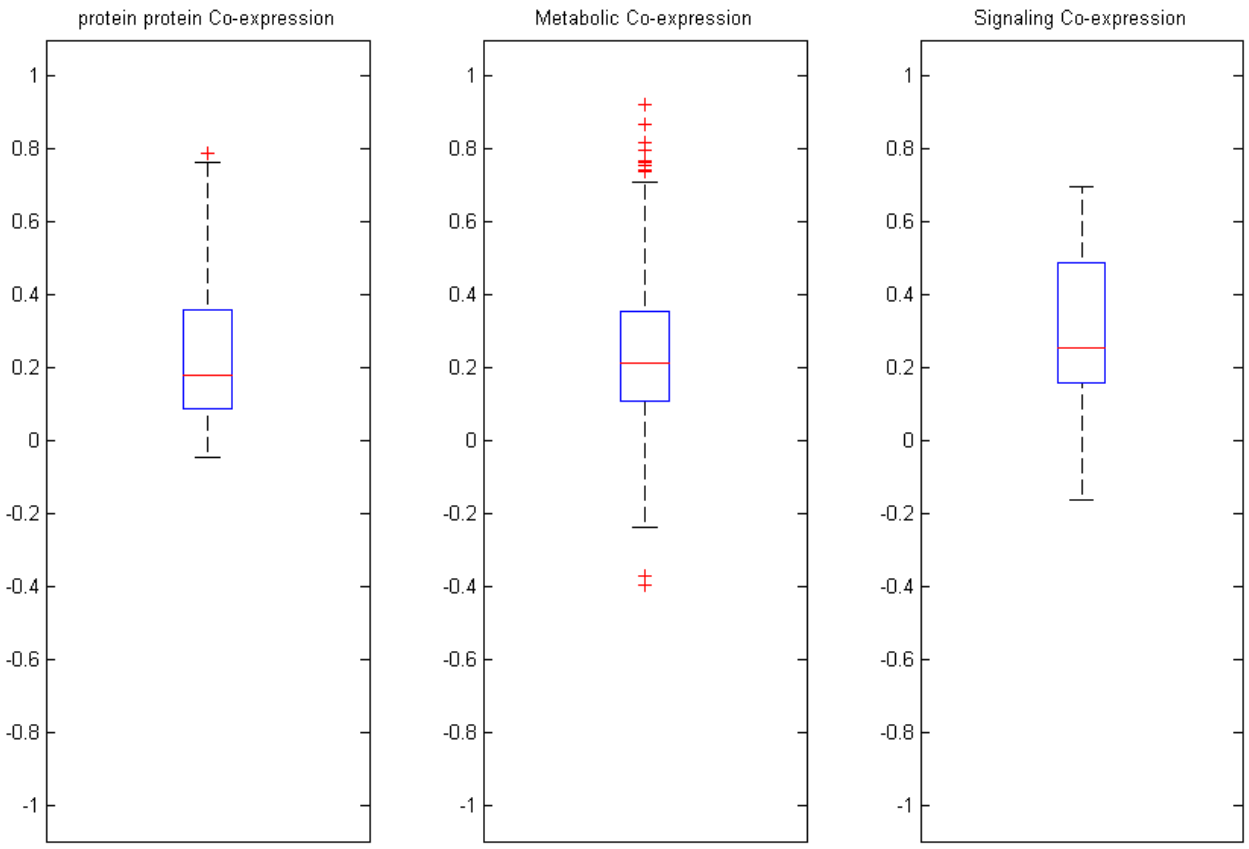


Figure 5.4. The distribution of the average co-expression value calculated by the Pearson correlation coefficient across all conditions for different modules involved in the same biological processes. These modules are divided to three categories modules derived from protein-protein interaction network (left panel), metabolic pathways (middle panel), and signalling pathways (right panel). These average Pearson correlation coefficients were usually much higher than 0, which is expected value for a random module.

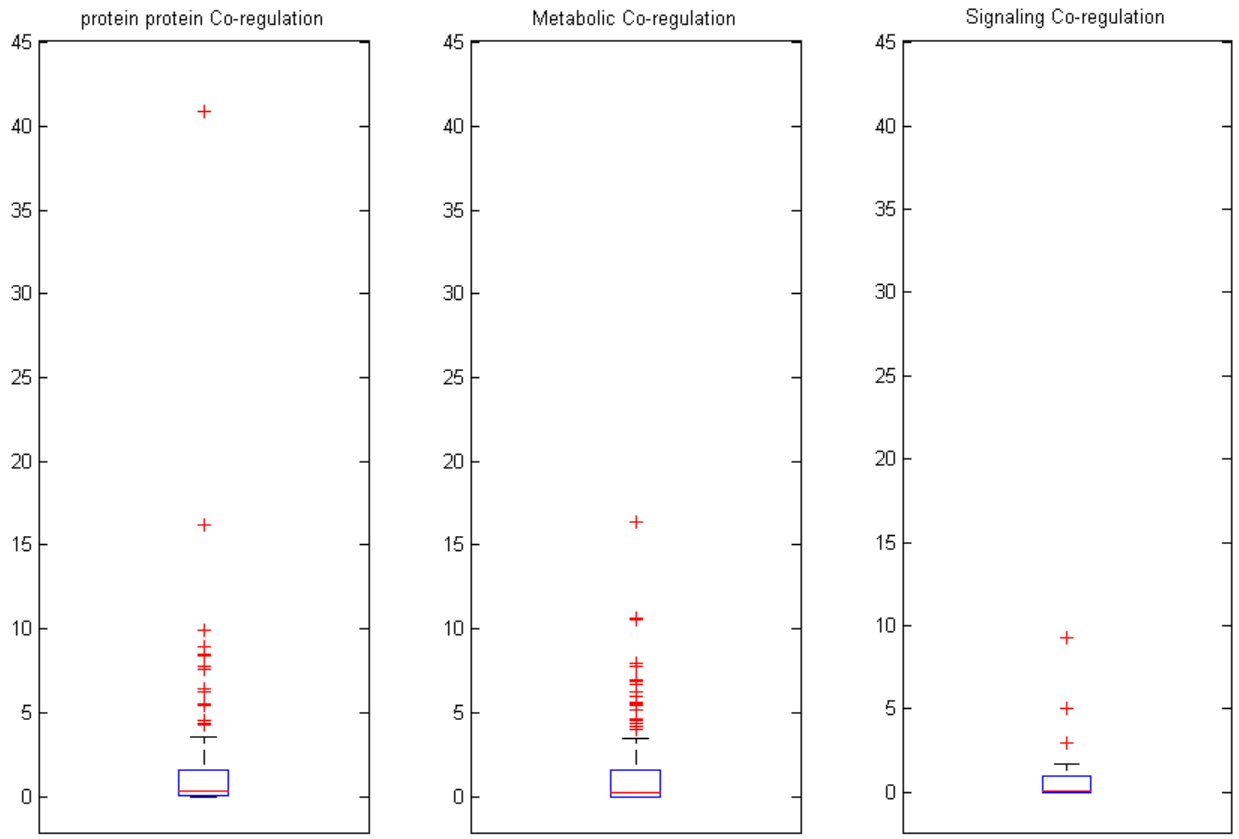


Figure 5.5. The distribution of the average co-regulation similarity measure for different modules involved in the same biological processes. These modules are divided to three categories modules derived from protein-protein interaction network (left panel), metabolic pathways (middle panel), and signalling pathways (right panel). These average co-regulatory similarity measures were usually much higher than 0.008, which is expected value for a random module.

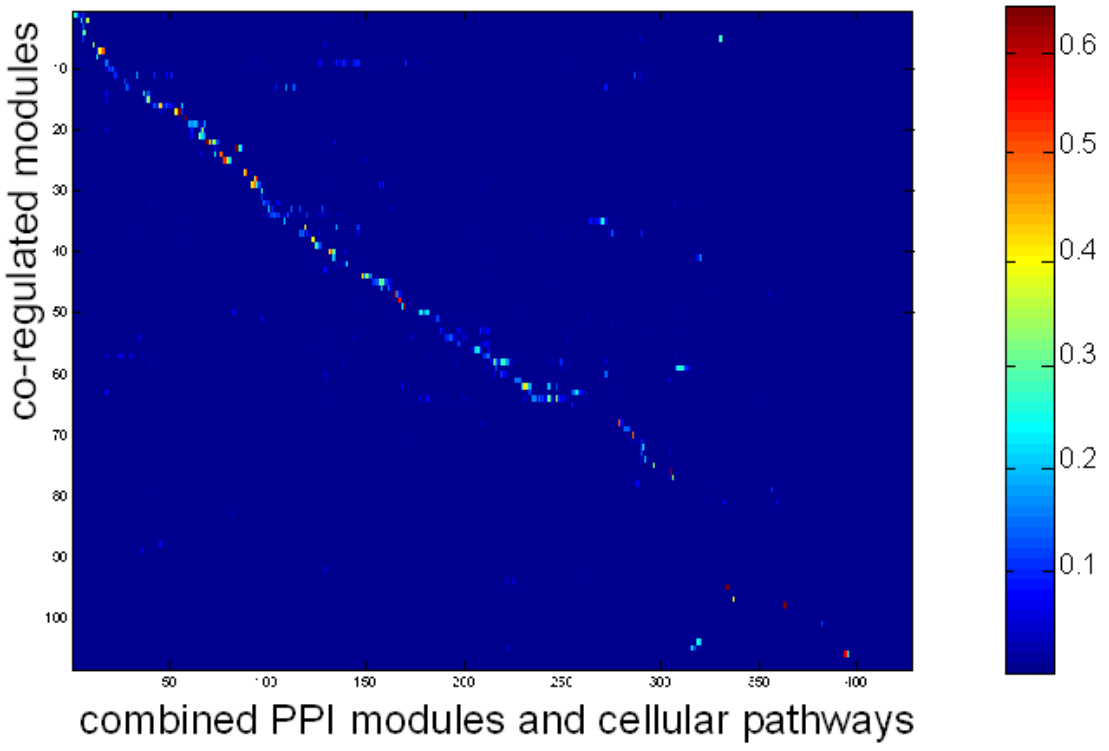


Figure 5.6. Jaccard coefficient was used to compare OSLOM detected modules using co-similarity measure with the modules involved in the same biological process. Modules involved in the same biological processes, including combined detected modules in protein-protein interaction network, and cellular metabolic and signaling pathways are represented on the X-axis. A hierarchical clustering were applied on these combined modules using co-regulatory similarity as the similarity measure, and these modules are ordered as the leafs of hierarchical clustering. With this representation modules with more similarity in regulation are located closer to each other on the X-axis. Modules derived by applying OSLOM on gene co-regulatory similarity are represented on Y-axis and reordered based on the order of the other axis to bring the higher Jaccard coefficient on the diagonal. The intensity of dots on the heatmap is related to the value of Jaccard coefficient.

5.3.4. COMPARING REGULATORY NETWORK HIERARCHY AND GO TERMS HIERARCHY

We have already shown the relation between modules of genes involved in the same biological processes and the regulatory network interaction which controls these modules. In this part we explore the relation between the regulatory hierarchy of the modules of genes involved in the same biological processes and their functional hierarchy. For this aim, first for each two modules

of genes involved in the same biological processes, the co-regulatory similarity measure of the modules was calculated. At the next step, for these modules, the functional similarity was calculated using the similarity measure between the modules (see 5.2.5). Finally the correlation between these two similarities measures across all pairs of modules was calculated using Pearson correlation coefficient. When we use all three GO terms domains, this Pearson correlation coefficient was equal to 0.2868 which is a considerably higher than what is expected by chance. This Pearson correlation coefficient was equal to 0.2682 considering just biological processes, 0.2575 considering just molecular function, 0.0154 considering just cellular components (This Pearson correlation coefficient was equal to 0.2851 for combined biological process and molecular function GO-terms).

Although the observed Pearson correlation coefficient (0.2868) clearly shows that there is a relation between regulatory hierarchy and functional hierarchy, still we looked at the extreme cases to investigate the reason of not obtaining very high correlation value. We observed that most of the signaling pathways gained a high functional similarity value with each other. Nevertheless, it is known that genes involved in different two-component signaling pathways do not have any relation with each other, signaling pathways have different regulatory programs. As a result we observe a problem that arises because of the GO-term misclassification. Although for an expert user the meaning of two-component signaling pathway is clear and the person will not consider them similar in function, for automatic analysis this classification is misleading. As another extreme case example, carbon transport related module has high module co-regulator similarity with carbon source related module while this similarity were not captured in GO-terms because the hierarchy of these two groups is different in all 3 domains although for an expert the functional similarity is observable. Still we cannot say there is no difference between functional hierarchy and regulatory hierarchy as paralogous gene in duplicated operons may not have similar regulatory pattern even though they can be involved in similar function or close species have many orthologous genes in common meaning that there functional hierarchy is similar, but regulatory hierarchy may be very different even for closer species.

5.4. DISCUSSION

We could exhibit the mutual relation between the regulatory network as a controlling network and other non-controlling networks such as protein-protein interaction network and cellular pathways. We divided biological processes into four general categories known as Regulatory interactions, signaling pathways, metabolic pathways, and structural and functional components (Figure 5.1). At the next step we tried to detect modules in each category. In previous chapter, we showed that finding modules in regulatory network is difficult (see 4.3.1). We downloaded signaling and metabolic pathways in a modular format from EcoCyc (Keseler, Bonavides-Martinez et al. 2009) (each pathway or super-pathway is considered a module). We detected modules which we called structural and functional components, by finding densely connected subnetwork in protein-protein interaction network using OSLOM. After detecting modules consisting of genes involved in similar biological processes, we could show that a high similarity in regulatory circuits exists for the genes in each module by using the defined co-regulatory similarity measure. In addition, if we tried to find modules just by looking into the regulatory network structure and using our defined co-similarity measure, we can observe modules very similar to the modules consisting of genes involved in similar biological processes.

We introduced a new species-specific functional similarity measure for two modules that has several advantageous over the methods that had been developed based on semantic similarity. Semantic similarity provides quantitative ways to compute similarities between GO terms, genes, and gene groups by considering the structure of GO terms DAG (Couto, Silva et al. 2007; Wang, Du et al. 2007; Du, Li et al. 2009). One major problem with semantic similarity is that it just considers the topology of GO terms, but it does not consider the number of genes inside the GO terms, while the number of genes inside a GO term is a species-specific feature of GO terms. In addition, it is believed that one use of semantic similarity is to highlight informative GO terms located at certain level(s), and usually the GO terms close to the root are considered as noninformative (Du, Li et al. 2009). As an example, (Hu, Jiang et al. 2010) took just 32 informative GO terms from biological process domain. Figure 5.7 shows the number of genes in GO terms located in first 4 layers of biological processes domain of *E. coli*. It is clear that the

distribution of genes in GO terms is not balance. Many GO terms contains several genes even in lower levels while there are always GO terms containing low numbers of genes in the very top of the DAG. For example, 8 out of 14 GO terms in the uppermost level of the biological process domain in *E. coli* includes less than 80 gene products (Table 5.1). This means that most of GO terms in the uppermost levels are still informative in *E. coli* case, while as it can be seen in Figure 5.7, the less-informative GO terms, including hundreds of genes, may exist even in lower levels. Furthermore, a certain gene may just be annotated for very general less-informative GO term in a certain organism. Therefore, ignoring less-informative GO terms will lead to lose all the available data for this gene.

We could also demonstrate the relation between regulatory hierarchy of the different modules and their functional hierarchy. For this aim we used our defined co-regulatory similarity measure to build the regulatory hierarchy of the modules, and we introduced a new species-specific functional similarity measure between modules, and we used it to build the functional hierarchy of the modules. We could observe correlation between co-regulatory similarity and functional similarity across different modules. Our results imply that although regulatory network evolves rapidly while functional GO terms is a static definition with firm structure of GO terms relations regardless of the organism, still a non-neglectable relation exists between regulatory hierarchy and functional hierarchy of modules.

The fact, that our defined co-regulatory similarity measure was successful to explain the relation between the controlling regulatory network and the other interaction networks, implies that the observed expression is the effect of the total structure of the regulatory network, not only direct regulators of the co-expressed genes. Based on the results of this chapter and previous chapter, we can claim that co-regulatory similarity is a proper measure to assess the regulatory similarity of genes as this measure is in agreement with the observed co-expression on the microarray compendia, and it could explain the relation between the highly flexible regulatory networks with the evolutionary conserved targets such as cellular pathways and protein complexes.

Regulatory networks evolve rapidly to allow the different species to adapt themselves to various environments, while cellular pathways and protein complexes are rather conserved across evolutionary related species (Shou, Bhardwaj et al. 2011) . Therefore, explaining the trends of

evolution of regulatory had remained as a difficult problem. The trends that rewiring may happen in regulatory networks were presented in several paper (Lozada-Chavez, Janga et al. 2006; Perez and Groisman 2009; Kim, Bhardwaj et al. 2010). Based on the properties of the defined co-regulatory measure, we expect that the evolutionary trends of regulatory network should be more global to ensure the co-regulatory similarity of related genes by considering the total structure of the network. Considering the co-regulatory formula we can expect the rewiring happens mostly in more global regulators as the more local regulators have lower contribution to the co-regulatory similarity value between two genes, but may contribute more to the co-regulatory similarity between two modules of genes because two modules are less like to share a local regulator. Therefore, we can explain why the mutant of many regulators, especially higher regulator in the hierarchy like Fnr, Fis, and ArcA, do not cause lethality in *E. coli* (Covert, Knight et al. 2004; Perrenoud and Sauer 2005; Blot, Mavathur et al. 2006; Bradley, Beach et al. 2007; Seshasayee, Fraser et al. 2009). Knocking out one regulator, especially a global one will not perturb the expression possible modules of genes involved in a certain biological process, but it may change the internal signal routing passes and disable the organism to address certain environmental perturbations.

Finally, functional similarity measure and co-regulatory similarity measure can be used in the data integration researches. Our defined functional similarity measure is species-specific and it fetches information from all the GO terms in the hierarchy. Likewise, our defined co-regulatory similarity measure is far more sensitive that assigning same function to all targets of a regulator or kinase. Although the Phosphorylation network currently is not available for *E. coli*, but while it will become available it can be integrated to the regulatory network to enrich the controlling interactions.

Gene Oontology ID	Gene Ontology Name	Gene Product number
GO:0001906	cell killing	2
GO:0040007	growth	5
GO:0032501	multicellular organismal process	6
GO:0016032	viral reproduction	20
GO:0000003	reproduction	30
GO:0022610	biological adhesion	44
GO:0032502	developmental process	46
GO:0051704	multi-organism process	73
GO:0071840	cellular component organization or biogenesis	265
GO:0065007	biological regulation	532
GO:0050896	response to stimulus	569
GO:0051179	localization	894
GO:0008152	metabolic process	2286
GO:0009987	cellular process	2374

Table 5.1. GO terms in the uppermost level of the biological process domain, directly linked to the root node GO:0008150 biological_process, and their number of annotated gene products in *E. coli*.

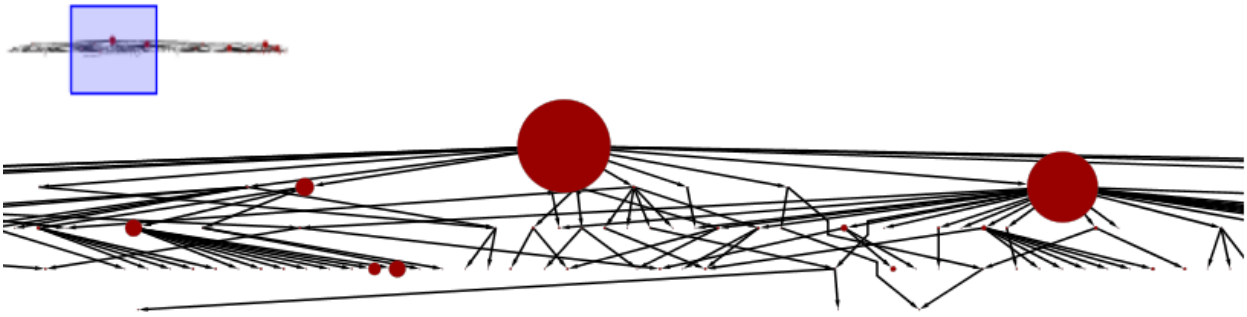


Figure 5.7. The zoom in of four first layers gene ontology structure of biological process domain in *E. coli*. Each biological domain is a tree-like directed acyclic graph (DAG) in which each node is a GO term and the direction of edges shows the parents GO terms. GO terms are selected by applying breath first search (BFS) algorithm on this DAG by setting the maximum distance from the root, GO:0008150 biological_process node, to three. The size of GO terms is related to the number of gene products annotated for this term. Therefore, the larger nodes are more general GO terms and are less-informative.

CHAPTER 6

CONCLUSIONS AND PERSPECTIVES

6.1. CONCLUSIONS

The advent of new technologies resulted is a revolutionary boost in molecular biology. System biology has facilitated integrating and analyzing various omics data from different experiments to uncover function of genes and interaction of them in different level. Several genome scale high-throughput experiments which measure different kind of interactions are available for model organisms like yeast and *E. coli*. Analyzing this data will lead to deep understanding of the cell life as a biological system. Therefore, developing new data mining methodology which can integrate different data sources with considering the biological complexities of the cell life is inevitable. In addition, this gained knowledge can be expanded to the other organisms by comparing proper data sources across various species.

Function of many genes originally predicted, based on transferring annotation of genes in other organisms with high sequence similarity, but sequence similarity is not the only data available data for cross-species comparison. Microarray co-expression data is known to be the most appropriate data source which can be coupled to sequence similarity to enrich the cross-species comparison of homologous genes (Chikina and Troyanskaya 2011). We could develop a new co-clustering methodology to detect co-expression conservation across two different organisms. In chapter 2, we applied this methodology on two well-studied prokaryotic model organisms: *Escherichia coli* and *Bacillus subtilis* as a proof of concept. Comparison with available knowledge of expression conservation across these two organisms, mostly based on operon conservation (Snel, van Noort et al. 2004; Okuda, Kawashima et al. 2005) or regulon conservation (Okuda, Kawashima et al. 2007), reveals the high performance of COMODO. COMODO could even detect co-expression conservation beyond both operon and regulon level such as genes which are involved in translation. In chapter 3, we applied COMODO to highlight the conservation and divergence of various biological processes across *Escherichia coli*, *S.*

enterica, and *Bacillus subtilis*. We could observe high co-expression conservation across *E. coli* and *S. enterica*, even conservation in response to various stimuli, signal transductions, and quorum sensing. The only non-conservation that we could detect was related to the genes involved in pathogenicity which are considered the main causes of difference in life style of the two phylogenetically close species *Escherichia coli* and *Salmonella enterica*. We could also demonstrate the application of the predicted conserved co-expressed modules to predict the possible conservation of the regulatory interactions. Comparing binding sites of these predicted regulators seems to be the most straight-forward way to increase the confidence of these predictions.

Integrating various data sources in one organism is another way to expand the current knowledge. One major problem with data integration is how to integrate the regulatory network as the controlling network with other data sources because the mutual relation between the regulatory network and other physical interactions which are controlled by the regulatory network was not explained before. Usually for data integration purposes genes, which are regulated by the same regulator(s), are considered functionally related (Myers, Chiriac et al. 2009; Narayanan, Vetta et al. 2010), but it is known that more global regulators have been shown to regulate genes with completely different biological functions (Jothi, Balaji et al. 2009; Kim, Bhardwaj et al. 2010). Unlike the previous papers that defined co-regulated genes as a pair of genes regulated by similar regulator(s) (Myers, Chiriac et al. 2009; Narayanan, Vetta et al. 2010), we defined a new co-regulatory similarity measures which considered the place of the regulator in the regulatory hierarchy to assign a co-regulatory similarity value to each pair of genes. In our defined measure, a lower regulator in the hierarchy, more local regulator, adds more to the co-regulatory similarity measure of its target genes compare to the higher and more global regulator. We calculated PageRank values for all the regulators in the regulatory network as a measure to find the place of each regulator in the regulatory hierarchy because this measure is in line with both connectivity and chain-of-command idea's which are known to be the main criteria to reflect the importance of regulators in the hierarchy (Myers, Chiriac et al. 2009; Narayanan, Vetta et al. 2010). We showed that our defined co-regulatory similarity measure is highly correlated with the observed co-expression on the microarray compendia in *E. coli*. In addition, we showed that genes, involved in the similar biological processes, demonstrate higher

co-regulatory similarity using our defined measure. Finally, we observed correlation between the regulatory hierarchy and the functional hierarchy, derived from GO categories, in *E. coli*.

Co-regulatory similarity measure is not the only measure that we introduced which can be used in data integration. We also introduced a new species-specific measure to assess functional similarity between a pair of genes considering GO terms. Our introduced measure is more sensitive compared to semantic similarity methods because first of all it is a species specific measure, and it considers the number genes in a GO term to assess to what extent a GO term is informative. In addition, it does not reduce any GO term as non-informative GO term.

6.2. PERSPECTIVES

6.2.1. DATA INTEGRATION

Although we have introduced two new measures, co-regulatory similarity measure (see 4.2.5) and functional similarity measure (see 5.2.5), suitable for data integration, we did not propose any data integration framework. The obvious application of our study is to propose a new data integration framework. In this part first we discuss the application of data integration in reconstructing biological networks and then we discuss how we can integrate various data sources into four biological processes categories inside a cell that we introduced in (5.3.2).

6.2.1.2. Reconstructing biological networks

The most intuitive way of viewing various omics data (see 1.1.1) is by adopting a graph based representation. In such view nodes represent different molecular entities whereas edges represent interactions. Edges can have weighs that reflect the degree of belief in a certain interaction. Integrating proper networks lead to a more comprehensive insight of cell behaviour, and usually in these networks, nodes represent either genes or proteins and edges represent either physical interactions or functional relations. Table 6.1 summarizes available data for *E. coli* in some of the biological networks which were highly used in data integration.

Although *E. coli* is the best-studied prokaryotic model organism, the available data is still poor for this organism. No genetic or post-translational (phosphorylation) interaction network exist for *E. coli*, and even the number of the high-confidence protein-protein interactions is limited. Validating interactions derived from high-throughput experiments as well as predicting possible new interactions based on the other data sources seem to be a promising approach to increase our knowledge about the biological network. Recently several papers were published regarding to the reconstruction of high-confidence protein-protein interaction network in *E. coli* (Hu, Janga et al. 2009; Peregrin-Alvarez, Xiong et al. 2009), genetic interaction network in yeast (Pandey, Zhang et al. 2010), and post-translation interaction network in yeast (Yachie, Saito et al. 2011). To assess the confidence of the predictions, the predicted interactions are benchmarked against the available validated interactions (derived from laboratory experiments) as a positive set and a set of non-interacting gene pairs as a negative set.

Our introduced co-regulatory similarity can be used to validate and predict interactions in co-expression, protein-protein, metabolic, and post-translational interaction networks (Table 6.1). At least validating and predicting interactions in protein-protein and post-translation interaction networks are more promising because a large high quality protein-protein interaction network does not exist in *E. coli*, and post-translational network does not exist in this organism at all. Although over 7600 high-confidence protein-protein interactions predicted in (Peregrin-Alvarez, Xiong et al. 2009), but this number is very low compare to the possible interactions that may exist. In addition, although (Hu, Janga et al. 2009) could build a very large protein-protein interaction network, but confidence of their predictions is not high. If we can build a large and reliable protein-protein interaction network, we can use it as one of the data sources to build post-translational interaction network.

NETWORK TYPE	INTERACTION DEFINTION	DATA SOURCE	INTEGRALE SOURCE
Functional interactions			
Co-expression network	Degree of co-expression between genes	**Expression compendia** (Faith, Hayete et al. 2007; Engelen, Fu et al. 2011)	Co-regulatory similarity[1]
Genetic Interaction network	Observed Phenotype defects in double mutants: aggravating or alleviating	**eSGA** (synthetic gene array) - proof of concept (Butland, Babu et al. 2008) Analysis Technology (**GIANT coli**) Proof of concept (Gross, Typas et al. 2008) **Low-throughput** experiments (about 200 interaction) (Babu, Musso et al. 2009)	
Physical interactions			
Metabolic Network	Flow of flux through enzyme reactions	**EcoCyc** (Keseler, Bonavides-Martinez et al. 2009) **iAF1260** (Feist, Henry et al. 2007)	Co-expression Text mining (co-occurrence) Gene neighborhood Co-regulatory similarity[1]
Protein-protein interaction network	Direct Physical interactions between two proteins or proteins co-eluting in a complex	**Maldi TOF, LC-MS/Ms** (6,234 interactions) (Butland, Peregrin-Alvarez et al. 2005) **Large-scale Pull down** (11,511 interactions) (Arifuzzaman, Maeda et al. 2006) **Large-scale SPA** (5,993 interactions) (Hu, Janga et al. 2009) **Dip database** (12,893 interactions in 2010/10/10) (Xenarios, Salwinski et al. 2002) **EcoCyc protein complexes** (Keseler, Bonavides-Martinez et al. 2009)	Co-expression Text mining (co-occurrence) Gene neighborhood Co-regulatory similarity[1] Conserved protein-protein interaction in other organisms

(post)transcriptional network	Physical interactions between a sigma/transcription factor (SRNA) and DNA	**RegulonDB** (Gama-Castro, Jimenez-Jacinto et al. 2008) **chip-chip**, **Chip-seq**	Expression compendia
post-translation network	Phosphorylation of a substrate by a kinase	**EcoCyc** (Hist-Asp two-components systems) (Keseler, Bonavides-Martinez et al. 2009) **Ser/Thr/Tyr phosphorylation** (Macek, Gnad et al. 2008)	Protein-protein interaction network Co-expression Gene neighborhood Co-regulatory similarity[1] GO terms: GO:0016301 kinase activity (177 gene products) -GO:0004871 signal transducer activity (102 gene products)

Table 6.1. Different functional and physical interaction networks and available data for each network in *E. coli*. The name of the network is written in the first column, and the definition of the interactions in the network can be found in the second column. For each network, both validated experiments and high-throughput experiments are mentioned in the third column. The other data sources which can be integrated to predict or verify interaction in the network are listed in the last column. One data source which can be used in reconstruction of some of the networks is our co-regulatory similarity measure indicated by 1. Co-regulatory similarity measure is not one data source but it is a functional relation measure derived from transcriptional, post-transcriptional , and post-translational interaction networks.

6.2.1.3. Integrating different data sources

In (5.3.2), we introduced four biological processes categories inside a cell and we called them regulatory interactions, signaling pathway, metabolic pathways, and structural and functional components. Still we did not introduce any data integration framework, and we just detected some structural and functional components by using the protein-protein interaction network. Different biological networks can be used to find modules in different biological processes (Figure 6.1).

Regulatory interaction network in our framework is simply an aggregation of all the controlling interaction networks including transcription, post-transcriptional, and post-translational network. We can improve the current available EcoCyc metabolic and signaling pathways for *E. coli* by using iAF1260 model for metabolic interaction network available in BIGG database (Feist, Henry et al. 2007). If we can build a post-translational network it can also assist to detect better signaling pathways. Microarray expression compendia and genetic interactions can be coupled to other data sources to boost the predictions in all four categories, but they should be interpreted differently for each category as the meaning of interactions is different for each category. For example, a genetic interaction between two regulators has completely different biological interpretation than a genetic interaction between two genes involved in the same or different metabolic pathway(s). Finally, our defined co-regulatory similarity measure can be used to improve the prediction in structural and functional components as well as both signaling and metabolic pathways.

One issue that should be considered in integrating different data sources is that if they have built independently or one data source has been used as a data source to build the other one. For example, both protein-protein interaction network and microarray expression compendia can be used to detect structural and functional components, but if microarray expression compendia had been used to detect interactions in protein-protein interaction network, then the result may be become biased towards the genes that show high co-expression with each other.

6.2.2. CROSS-SPECIES COMPARISON

We have developed COMODO that use microarray expression compendia and sequence homology as an input to verify the genes conserved in co-expression. One obvious way to improve this cross-species comparison is to add other data sources. Protein-protein interactions and metabolic pathways tend to be conserved across phylogenetically close species, and have already been used for cross-species comparison across different organisms (Karp, Ouzounis et al. 2005; Bandyopadhyay, Sharan et al. 2006; Li, Coghlan et al. 2006; Karimpour-Fard, Detweiler et al. 2007; Singh, Xu et al. 2008; Zaslavskiy, Bach et al. 2009). In contrast to the mentioned data sources, regulatory network has not highly been used as a source of data in cross-species comparison as inferring functional relation between genes based on this network is not straightforward. The introduced co-regulatory similarity measure facilitates using this data source in a cross-species comparison framework.

Although the developed cross-species comparison methods usually tend to identify homologous genes with conserved function, detecting conserved interactions in different networks may be one interesting application of cross-species comparison methods. Even though expanding the available interactions from better studied organism to other organisms or strains where the available interactions is limited seemed to be easy in protein-protein and metabolic interaction networks, this expanding is not straightforward for the regulatory network. The introduced co-regulatory similarity measure can also be used to expand or predict the regulatory interactions.

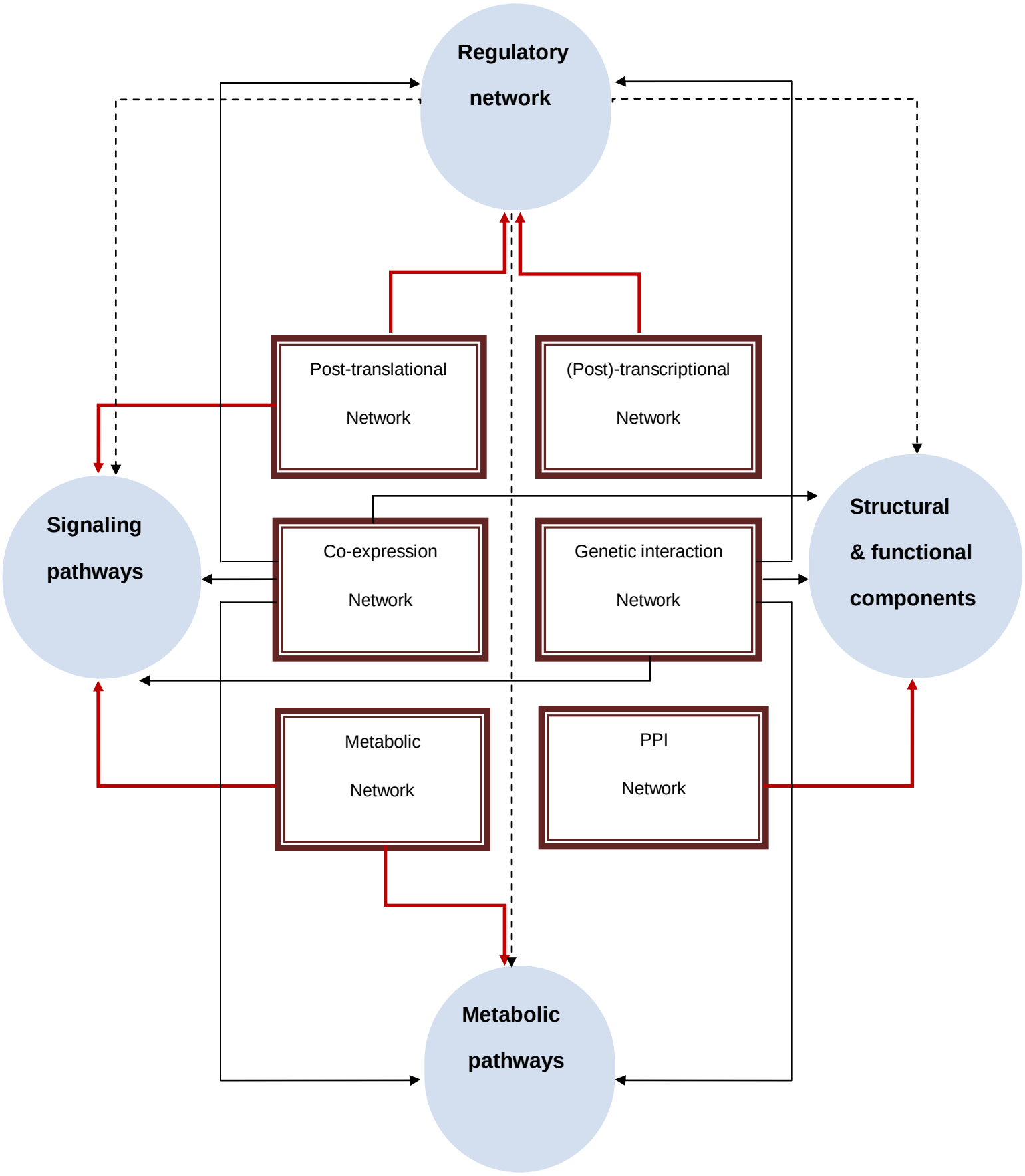

Regulatory network
Post-translational Network
(Post)-transcriptional Network
Signaling pathways
Co-expression Network
Genetic interaction Network
Structural & functional components
Metabolic Network
PPI Network
Metabolic pathways

Figure 6.1. The relation between different biological network and four biological process categories that we have introduced. Different biological networks can be used to detect modules in four biological categories that we introduced in (5.3.2). Different biological networks are shown by rectangles and the four categories are depicted by squares. The networks which are directly related to the modules in our defined categories are connected to their related categories by red arrows. Regulatory interaction network in our framework is simply an aggregation of all the controlling interaction networks including (post)-transcriptional, and post-translational network. Metabolic and signaling pathways can be deduced from metabolic network, and post-translation network can be used as a coupling data source to detect signaling pathways. Structural and functional components can be detected by using the protein-protein interaction network (see chapter 5). In addition, the networks which are connected by black solid arrows to their categories are the networks which can be coupled to the other data sources to boost the predictions accuracy. For example, co-expression network can be used to predict more accurate metabolic and signaling pathways since genes, involved in the same biological pathways, are expected to exhibit higher co-expression. The same argument is also true regarding to use the co-expression network to predict more accurate structural and functional components because proteins, involved in the same structural and functional components, tend to exhibit higher co-expression. Co-expression network can also be used to predict regulatory network interactions. In chapter 4 we discussed the relation between these two networks. Finally, the dashed arrows represent the co-regulatory measure that derived from the regulatory network and can be used for better module prediction in other three categories. We showed in (5.3.3) that genes, involved in the same biological metabolic and signaling pathways or same structural and functional components, exhibit higher co-regulation similarity based on the measure that we introduced for co-regulation similarity (see also Figure 5.5).

REFERENCES

Andres Leon, E., I. Ezkurdia, et al. (2009). "EcID. A database for the inference of functional interactions in E. coli." Nucleic Acids Res **37**(Database issue): D629-635.

Andrews, S. C., M. L. Cartron, et al. (2006). "Feo - Transport of ferrous iron into bacteria." Biometals **19**(2): 143-157.

Arifuzzaman, M., M. Maeda, et al. (2006). "Large-scale identification of protein-protein interaction of Escherichia coli K-12." Genome Res **16**(5): 686-691.

Baba, T., T. Ara, et al. (2006). "Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection." Mol Syst Biol **2**: 2006 0008.

Babu, M., G. Musso, et al. (2009). "Systems-level approaches for identifying and analyzing genetic interaction networks in Escherichia coli and extensions to other prokaryotes." Mol Biosyst **5**(12): 1439-1455.

Babu, M. M., S. A. Teichmann, et al. (2006). "Evolutionary dynamics of prokaryotic transcriptional regulatory networks." Journal of Molecular Biology **358**(2): 614-633.

Balaji, S., M. M. Babu, et al. (2006). "Comprehensive analysis of combinatorial regulation using the transcriptional regulatory network of yeast." Journal of Molecular Biology **360**(1): 213-227.

Bandyopadhyay, S., R. Sharan, et al. (2006). "Systematic identification of functional orthologs based on protein network comparison." Genome Res **16**(3): 428-435.

Barrett, T., D. B. Troup, et al. (2007). "NCBI GEO: mining tens of millions of expression profiles--database and tools update." Nucleic Acids Res **35**(Database issue): D760-765.

Benjamini, Y. and Y. Hochberg (1995). "Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing." Journal of the Royal Statistical Society Series B-Methodological **57**(1): 289-300.

Bergmann, S., J. Ihmels, et al. (2003). "Iterative signature algorithm for the analysis of large-scale gene expression data." Phys Rev E Stat Nonlin Soft Matter Phys **67**(3 Pt 1): 031902.

Bergmann, S., J. Ihmels, et al. (2004). "Similarities and differences in genome-wide expression data of six organisms." Plos Biology **2**(1): 85-93.

Beyer, A., S. Bandyopadhyay, et al. (2007). "Integrating physical and genetic maps: from genomes to interaction networks." Nature Reviews Genetics **8**(9): 699-710.

Blot, N., R. Mavathur, et al. (2006). "Homeostatic regulation of supercoiling sensitivity coordinates transcription of the bacterial genome." EMBO Rep **7**(7): 710-715.

Bowers, P. M., M. Pellegrini, et al. (2004). "Prolinks: a database of protein functional linkages derived from coevolution." Genome Biol **5**(5): R35.

Bradley, M. D., M. B. Beach, et al. (2007). "Effects of Fis on Escherichia coli gene expression during different growth stages." Microbiology **153**(Pt 9): 2922-2940.

Butland, G., M. Babu, et al. (2008). "eSGA: E. coli synthetic genetic array analysis." Nat Methods **5**(9): 789-795.

Butland, G., J. M. Peregrin-Alvarez, et al. (2005). "Interaction network containing conserved and essential protein complexes in Escherichia coli." Nature **433**(7025): 531-537.

Cai, J., D. Xie, et al. "Modeling co-expression across species for complex traits: insights to the difference of human and mouse embryonic stem cells." PLoS computational biology **6**(3): e1000707.

Cai, J., D. Xie, et al. (2010). "Modeling co-expression across species for complex traits: insights to the difference of human and mouse embryonic stem cells." PLoS Comput Biol **6**(3): e1000707.

Calvio, C., F. Celandroni, et al. (2005). "Swarming differentiation and swimming motility in Bacillus subtilis are controlled by swrA, a newly identified dicistronic operon." J Bacteriol **187**(15): 5356-5366.

Calvio, C., C. Osera, et al. (2008). "Autoregulation of swrAA and motility in Bacillus subtilis." J Bacteriol **190**(16): 5720-5728.
Carbon, S., A. Ireland, et al. (2009). "AmiGO: online access to ontology and annotation data." Bioinformatics **25**(2): 288-289.
Chadsey, M. S., J. E. Karlinsey, et al. (1998). "The flagellar anti-sigma factor FlgM actively dissociates Salmonella typhimurium sigma28 RNA polymerase holoenzyme." Genes Dev **12**(19): 3123-3136.
Chandrasekaran, S. and N. D. Price (2010). "Probabilistic integrative modeling of genome-scale metabolic and regulatory networks in Escherichia coli and Mycobacterium tuberculosis." Proc Natl Acad Sci U S A **107**(41): 17845-17850.
Chen, Z., K. A. Lewis, et al. (2007). "Discovery of Fur binding site clusters in Escherichia coli by information theory models." Nucleic Acids Res **35**(20): 6762-6777.
Chikina, M. D. and O. G. Troyanskaya (2011). "Accurate Quantification of Functional Analogy among Close Homologs." Plos Computational Biology **7**(2).
Cho, B. K., K. Zengler, et al. (2009). "The transcription unit architecture of the Escherichia coli genome." Nat Biotechnol **27**(11): 1043-1049.
Couto, F. M., M. J. Silva, et al. (2007). "Measuring semantic similarity between Gene Ontology terms." Data & Knowledge Engineering **61**(1): 137-152.
Covert, M. W., E. M. Knight, et al. (2004). "Integrating high-throughput and computational data elucidates bacterial networks." Nature **429**(6987): 92-96.
Danot, O., D. Vidal-Ingigliardi, et al. (1996). "Two amino acid residues from the DNA-binding domain of MalT play a crucial role in transcriptional activation." Journal of molecular biology **262**(1): 1-11.
Demeter, J., C. Beauheim, et al. (2007). "The Stanford Microarray Database: implementation of new analysis tools and open source release of software." Nucleic Acids Res **35**(Database issue): D766-770.
Du, Z., L. Li, et al. (2009). "G-SESAME: web tools for GO-term-based gene similarity analysis and knowledge discovery." Nucleic Acids Res **37**(Web Server issue): W345-349.
Engelen, K., Q. Fu, et al. (2011). "COLOMBOS: Access Port for Cross-Platform Bacterial Expression Compendia." PLoS One **6**(7): e20938.
Enright, A. J., I. Iliopoulos, et al. (1999). "Protein interaction maps for complete genomes based on gene fusion events." Nature **402**(6757): 86-90.
Eppler, T., P. Postma, et al. (2002). "Glycerol-3-phosphate-induced catabolite repression in Escherichia coli." Journal of bacteriology **184**(11): 3044-3052.
Erill, I., S. Campoy, et al. (2007). "Aeons of distress: an evolutionary perspective on the bacterial SOS response." FEMS Microbiol Rev **31**(6): 637-656.
Fadda, A., A. C. Fierro, et al. (2009). "Inferring the transcriptional network of Bacillus subtilis." Molecular Biosystems **5**(12): 1840-1852.
Faith, J. J., B. Hayete, et al. (2007). "Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles." PLoS Biol **5**(1): e8.
Feist, A. M., C. S. Henry, et al. (2007). "A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information." Mol Syst Biol **3**: 121.
Fierro, A. C., F. Vandenbussche, et al. (2008). "Meta Analysis of Gene Expression Data within and Across Species." Current Genomics **9**(8): 525-534.
Fisher, S. H., M. A. Strauch, et al. (1994). "Modulation of Bacillus subtilis catabolite repression by transition state regulatory protein AbrB." Journal of bacteriology **176**(7): 1903-1912.
Fukami-Kobayashi, K., Y. Tateno, et al. (2003). "Parallel evolution of ligand specificity between LacI/GalR family repressors and periplasmic sugar-binding proteins." Molecular Biology and Evolution **20**(2): 267-277.

Gama-Castro, S., V. Jimenez-Jacinto, et al. (2008). "RegulonDB (version 6.0): gene regulation model of Escherichia coli K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation." Nucleic Acids Research **36**: D120-D124.
Gelfand, M. S., K. S. Makarova, et al. (2001). "Conservation of the binding site for the arginine repressor in all bacterial lineages." Genome Biology **2**(4).
Gerdes, S. Y., M. D. Scholle, et al. (2003). "Experimental determination and system level analysis of essential genes in Escherichia coli MG1655." Journal of Bacteriology **185**(19): 5673-5684.
Gerstein, M., H. Y. Yu, et al. (2004). "Annotation transfer between genomes: Protein-protein interologs and protein-DNA regulogs." Genome Research **14**(6): 1107-1118.
Gerstein, M. B., N. Bhardwaj, et al. (2010). "Analysis of diverse regulatory networks in a hierarchical context shows consistent tendencies for collaboration in the middle levels." Proceedings of the National Academy of Sciences of the United States of America **107**(15): 6841-6846.
Gross, C. A., A. Typas, et al. (2008). "High-throughput, quantitative analyses of genetic interactions in E. coli." Nature Methods **5**(9): 781-787.
Grunberg-Manago, M. (1999). "Messenger RNA stability and its role in control of gene expression in bacteria and phages." Annual Review of Genetics **33**: 193-227.
Guillouard, I., S. Auger, et al. (2002). "Identification of Bacillus subtilis CysL, a regulator of the cysJI operon, which encodes sulfite reductase." Journal of bacteriology **184**(17): 4681-4689.
Gutierrez-Rios, R. M., D. A. Rosenblueth, et al. (2003). "Regulatory network of Escherichia coli: Consistency between literature knowledge and microarray profiles." Genome Research **13**(11): 2435-2443.
Hamze, K., D. Julkowska, et al. (2009). "Identification of genes required for different stages of dendritic swarming in Bacillus subtilis, with a novel role for phrC." Microbiology-Sgm **155**: 398-412.
Hartig, E., A. Hartmann, et al. (2006). "The Bacillus subtilis nrdEF genes, encoding a class Ib ribonucleotide reductase, are essential for aerobic and anaerobic growth." Applied and Environmental Microbiology **72**(8): 5260-5265.
Herrgard, M. J., M. W. Covert, et al. (2003). "Reconciling gene expression data with known genome-scale regulatory network structures." Genome Research **13**(11): 2423-2434.
Hoffmann, R. and A. Valencia (2005). "Implementing the iHOP concept for navigation of biomedical literature." Bioinformatics **21 Suppl 2**: ii252-258.
Horlacher, R. and W. Boos (1997). "Characterization of TreR, the major regulator of the Escherichia coli trehalose system." The Journal of biological chemistry **272**(20): 13026-13032.
Hu, J. C., P. D. Karp, et al. (2009). "What we can learn about Escherichia coli through application of Gene Ontology." Trends in Microbiology **17**(7): 269-278.
Hu, P., S. C. Janga, et al. (2009). "Global functional atlas of Escherichia coli encompassing previously uncharacterized proteins." PLoS Biol **7**(4): e96.
Hu, P., H. Jiang, et al. (2010). "Predicting protein functions by relaxation labelling protein interaction network." Bmc Bioinformatics **11 Suppl 1**: S64.
Huang, S. S. and E. Fraenkel (2009). "Integrating proteomic, transcriptional, and interactome data reveals hidden components of signaling and regulatory networks." Science Signaling **2**(81): ra40.
Hyduke, D. R. and B. O. Palsson (2010). "Towards genome-scale signalling-network reconstructions." Nature Reviews Genetics **11**(4): 297-307.
Ihmels, J., S. Bergmann, et al. (2004). "Defining transcription modules using large-scale gene expression data." Bioinformatics **20**(13): 1993-2003.
Ihmels, J., S. Bergmann, et al. (2005). "Comparative gene expression analysis by a differential clustering approach: Application to the Candida albicans transcription program." Plos Genetics **1**(3): 380-393.

Ishii, N., K. Nakahigashi, et al. (2007). "Multiple high-throughput analyses monitor the response of E. coli to perturbations." Science **316**(5824): 593-597.
Jothi, R., S. Balaji, et al. (2009). "Genomic analysis reveals a tight link between transcription factor dynamics and regulatory network architecture." Mol Syst Biol **5**: 294.
Jothi, R., S. Balaji, et al. (2009). "Genomic analysis reveals a tight link between transcription factor dynamics and regulatory network architecture." Molecular Systems Biology **5**.
Karimpour-Fard, A., C. S. Detweiler, et al. (2007). "Cross-species cluster co-conservation: a new method for generating protein interaction networks." Genome Biology **8**(9): -.
Karp, P. D., C. A. Ouzounis, et al. (2005). "Expansion of the BioCyc collection of pathway/genome databases to 160 genomes." Nucleic Acids Research **33**(19): 6083-6089.
Keller, A., C. Backes, et al. (2007). "Computation of significance scores of unweighted Gene Set Enrichment Analyses." Bmc Bioinformatics **8**: 290.
Kelley, R. and T. Ideker (2005). "Systematic interpretation of genetic interactions using protein networks." Nat Biotechnol **23**(5): 561-566.
Keseler, I. M., C. Bonavides-Martinez, et al. (2009). "EcoCyc: A comprehensive view of Escherichia coli biology." Nucleic Acids Research **37**: D464-D470.
Kim, P. M., N. Bhardwaj, et al. (2010). "Rewiring of Transcriptional Regulatory Networks: Hierarchy, Rather than Connectivity, Better Reflects the Importance of Regulators." Science Signaling **3**(146).
Kobayashi, K., S. D. Ehrlich, et al. (2003). "Essential Bacillus subtilis genes." Proceedings of the National Academy of Sciences of the United States of America **100**(8): 4678-4683.
Koide, A. and J. A. Hoch (1994). "Identification of a second oligopeptide transport system in Bacillus subtilis and determination of its role in sporulation." Molecular microbiology **13**(3): 417-426.
Lancichinetti, A., F. Radicchi, et al. (2011). "Finding statistically significant communities in networks." PLoS One **6**(4): e18961.
Lee, J. M., E. P. Gianchandani, et al. (2008). "Dynamic analysis of integrated signaling, metabolic, and regulatory networks." PLoS Comput Biol **4**(5): e1000086.
Lefebvre, C., J. C. Aude, et al. (2005). "Balancing protein similarity and gene co-expression reveals new links between genetic conservation and developmental diversity in invertebrates." Bioinformatics **21**(8): 1550-1558.
Lelandais, G., V. Tanty, et al. (2008). "Genome adaptation to chemical stress: clues from comparative transcriptomics in Saccharomyces cerevisiae and Candida glabrata." Genome Biology **9**(11): -.
Lemmens, K., T. De Bie, et al. (2009). "DISTILLER: a data integration framework to reveal condition dependency of complex regulons in Escherichia coli." Genome Biology **10**(3): -.
Letoffe, S., P. Delepelaire, et al. (2006). "The housekeeping dipeptide permease is the Escherichia coli heme transporter and functions with two optional peptide binding proteins." Proceedings of the National Academy of Sciences of the United States of America **103**(34): 12891-12896.
Li, H., A. Coghlan, et al. (2006). "TreeFam: a curated database of phylogenetic trees of animal gene families." Nucleic Acids Res **34**(Database issue): D572-580.
Lozada-Chavez, I., S. C. Janga, et al. (2006). "Bacterial regulatory networks are extremely flexible in evolution (vol 34, pg 3434, 2006)." Nucleic Acids Research **34**(16): 4654-4654.
Lu, Y., X. He, et al. (2007). "Cross-species microarray analysis with the OSCAR system suggests an INSR -> Pax6 -> NQO1 neuro-protective pathway in aging and Alzheimer's disease." Nucleic Acids Research **35**: W105-W114.
Lu, Y., P. Huggins, et al. (2009). "Cross species analysis of microarray expression data." Bioinformatics **25**(12): 1476-1483.

Macek, B., F. Gnad, et al. (2008). "Phosphoproteome analysis of E. coli reveals evolutionary conservation of bacterial Ser/Thr/Tyr phosphorylation." Mol Cell Proteomics **7**(2): 299-307.
Mao, F. L., P. Dam, et al. (2009). "DOOR: a database for prokaryotic operons." Nucleic Acids Research **37**: D459-D463.
Marr, C., F. J. Theis, et al. (2010). "Patterns of Subnet Usage Reveal Distinct Scales of Regulation in the Transcriptional Regulatory Network of Escherichia coli." Plos Computational Biology **6**(7).
Menard, A., P. E. D. L. Santos, et al. (2007). "Architecture of Burkholderia cepacia complex sigma(70) gene family: evidence of alternative primary and clade-specific factors, and genomic instability." Bmc Genomics **8**: -.
Michoel, T., A. Joshi, et al. (2011). "Enrichment and aggregation of topological motifs are independent organizational principles of integrated interaction networks." Mol Biosyst **7**(10): 2769-2778.
Myers, C. L., C. Chiriac, et al. (2009). "Discovering biological networks from diverse functional genomic data." Methods Mol Biol **563**: 157-175.
Narayanan, M., A. Vetta, et al. (2010). "Simultaneous clustering of multiple gene expression and physical interaction datasets." PLoS Comput Biol **6**(4): e1000742.
Okuda, S., S. Kawashima, et al. (2005). "Conservation of gene co-regulation between two prokaryotes: Bacillus subtilis and Escherichia coli." Genome Inform **16**(1): 116-124.
Okuda, S., S. Kawashima, et al. (2007). "Characterization of relationships between transcriptional units and operon structures in Bacillus subtilis and Escherichia coli." Bmc Genomics **8**: -.
Oldham, M. C., S. Horvath, et al. (2006). "Conservation and evolution of gene coexpression networks in human and chimpanzee brains." Proc Natl Acad Sci U S A **103**(47): 17973-17978.
Ollinger, J., K. B. Song, et al. (2006). "Role of the fur regulon in iron transport in Bacillus subtilis." Journal of Bacteriology **188**(10): 3664-3673.
Osterberg, S., T. Del Peso-Santos, et al. (2011). "Regulation of Alternative Sigma Factor Use." Annu Rev Microbiol.
Page, L. and S. Brin (1998). "The anatomy of a large-scale hypertextual Web search engine." Computer Networks and Isdn Systems **30**(1-7): 107-117.
Paget, M. S. B. and J. D. Helmann (2003). "Protein family review - The sigma(70) family of sigma factors." Genome Biology **4**(1): -.
Pandey, G., B. Zhang, et al. (2010). "An integrative multi-network and multi-classifier approach to predict genetic interactions." PLoS Comput Biol **6**(9).
Parkinson, H., M. Kapushesky, et al. (2007). "ArrayExpress--a public database of microarray experiments and gene expression profiles." Nucleic Acids Res **35**(Database issue): D747-750.
Perego, M., C. F. Higgins, et al. (1991). "The Oligopeptide Transport-System of Bacillus-Subtilis Plays a Role in the Initiation of Sporulation." Molecular Microbiology **5**(1): 173-185.
Peregrin-Alvarez, J. M., X. Xiong, et al. (2009). "The Modular Organization of Protein Interactions in Escherichia coli." PLoS Comput Biol **5**(10): e1000523.
Perez, J. C. and E. A. Groisman (2009). "Evolution of Transcriptional Regulatory Circuits in Bacteria." Cell **138**(2): 233-244.
Perrenoud, A. and U. Sauer (2005). "Impact of global transcriptional regulation by ArcA, ArcB, Cra, Crp, Cya, Fnr, and Mlc on glucose catabolism in Escherichia coli." J Bacteriol **187**(9): 3171-3179.
Pesavento, C., G. Becker, et al. (2008). "Inverse regulatory coordination of motility and curli-mediated adhesion in Escherichia coli." Genes Dev **22**(17): 2434-2446.
Peterson, J. D., L. A. Umayam, et al. (2001). "The Comprehensive Microbial Resource." Nucleic Acids Res **29**(1): 123-125.

Pons, T., B. Gonzalez, et al. (2006). "FlgM anti-sigma factors: identification of novel members of the family, evolutionary analysis, homology modeling, and analysis of sequence-structure-function relationships." Journal of Molecular Modeling **12**(6): 973-983.
Price, M. N., P. S. Dehal, et al. (2007). "Orthologous transcription factors in bacteria have different functions and regulate different genes." Plos Computational Biology **3**(9): 1739-1750.
Price, M. N., K. H. Huang, et al. (2005). "A novel method for accurate operon predictions in all sequenced prokaryotes." Nucleic Acids Research **33**(3): 880-892.
Rain, J. C., L. Selig, et al. (2001). "The protein-protein interaction map of Helicobacter pylori." Nature **409**(6817): 211-215.
Rajagopala, S. V., B. Titz, et al. (2007). "The protein network of bacterial motility." Mol Syst Biol **3**: 128.
Rasouly, A., M. Schonbrun, et al. (2009). "YbeY, a Heat Shock Protein Involved in Translation in Escherichia coli." Journal of Bacteriology **191**(8): 2649-2655.
Rastogi, S. and D. A. Liberles (2005). "Subfunctionalization of duplicated genes as a transition state to neofunctionalization." BMC evolutionary biology **5**(1): 28.
Rhodius, V. A., W. C. Suh, et al. (2006). "Conserved and variable functions of the sigmaE stress response in related genomes." PLoS Biol **4**(1): e2.
Robertson, J. B., M. Gocht, et al. (1989). "AbrB, a regulator of gene expression in Bacillus, interacts with the transcription initiation regions of a sporulation gene and an antibiotic biosynthesis gene." Proceedings of the National Academy of Sciences of the United States of America **86**(21): 8457-8461.
Rodionov, D. A. and M. S. Gelfand (2005). "Identification of a bacterial regulatory system for ribonucleotide reductases by phylogenetic profiling." Trends in Genetics **21**(7): 385-389.
Rodionov, D. A., A. A. Mironov, et al. (2002). "Conservation of the biotin regulon and the BirA regulatory signal in Eubacteria and Archaea." Genome Res **12**(10): 1507-1516.
Rolfes, R. J. and H. Zalkin (1988). "Escherichia-Coli Gene Purr Encoding a Repressor Protein for Purine Nucleotide Synthesis - Cloning, Nucleotide-Sequence, and Interaction with the Purf Operator." Journal of Biological Chemistry **263**(36): 19653-19661.
Salwinski, L., C. S. Miller, et al. (2004). "The Database of Interacting Proteins: 2004 update." Nucleic Acids Res **32**(Database issue): D449-451.
Schock, F. and M. K. Dahl (1996). "Expression of the tre operon of Bacillus subtilis 168 is regulated by the repressor TreR." Journal of bacteriology **178**(15): 4576-4581.
Schumann, W. (2003). "The Bacillus subtilis heat shock stimulon." Cell Stress & Chaperones **8**(3): 207-217.
Sekowska, A., H. F. Kung, et al. (2000). "Sulfur metabolism in Escherichia coli and related bacteria: facts and fiction." Journal of molecular microbiology and biotechnology **2**(2): 145-177.
Seshasayee, A. S., G. M. Fraser, et al. (2009). "Principles of transcriptional regulation and evolution of the metabolic system in E. coli." Genome Res **19**(1): 79-91.
Shou, C., N. Bhardwaj, et al. (2011). "Measuring the evolutionary rewiring of biological networks." PLoS Comput Biol **7**(1): e1001050.
Sierro, N., Y. Makita, et al. (2008). "DBTBS: a database of transcriptional regulation in Bacillus subtilis containing upstream intergenic conservation information." Nucleic Acids Research **36**: D93-D96.
Singh, R., J. Xu, et al. (2008). "Global alignment of multiple protein interaction networks with application to functional orthology detection." Proc Natl Acad Sci U S A **105**(35): 12763-12768.
Smith, T. G. and T. R. Hoover (2009). "Deciphering bacterial flagellar gene regulatory networks in the genomic era." Adv Appl Microbiol **67**: 257-295.
Snel, B., V. van Noort, et al. (2004). "Gene co-regulation is highly conserved in the evolution of eukaryotes and prokaryotes." Nucleic Acids Research **32**(16): 4725-4731.

Stuart, J. M., E. Segal, et al. (2003). "A gene-coexpression network for global discovery of conserved genetic modules." Science **302**(5643): 249-255.
Subramanian, A., P. Tamayo, et al. (2005). "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles." Proc Natl Acad Sci U S A **102**(43): 15545-15550.
Sun, G., E. Sharkova, et al. (1996). "Regulators of aerobic and anaerobic respiration in Bacillus subtilis." Journal of bacteriology **178**(5): 1374-1385.
Tatusov, R. L., N. D. Fedorova, et al. (2003). "The COG database: an updated version includes eukaryotes." Bmc Bioinformatics **4**: -.
Tatusov, R. L., E. V. Koonin, et al. (1997). "A genomic perspective on protein families." Science **278**(5338): 631-637.
Thiele, I., N. Jamshidi, et al. (2009). "Genome-scale reconstruction of Escherichia coli's transcriptional and translational machinery: a knowledge base, its mathematical formulation, and its functional characterization." PLoS Comput Biol **5**(3): e1000312.
Tirosh, I. and N. Barkai (2007). "Comparative analysis indicates regulatory neofunctionalization of yeast duplicates." Genome Biology **8**(4): -.
Tirosh, I., Y. Bilu, et al. (2007). "Comparative biology: beyond sequence analysis." Current Opinion in Biotechnology **18**(4): 371-377.
Torrents, E., I. Grinberg, et al. (2007). "NrdR controls differential expression of the Escherichia coli ribonucleotide reductase genes." Journal of Bacteriology **189**(14): 5012-5021.
Van den Bulcke, T., K. Lemmens, et al. (2006). "Inferring transcriptional networks by mining 'Omics' data." Current Bioinformatics **1**(3): 301-313.
van Noort, V., B. Snel, et al. (2003). "Predicting gene function by conserved co-expression." Trends in Genetics **19**(5): 238-242.
Vazquez, C. D., J. A. Freyre-Gonzalez, et al. (2009). "Identification of network topological units coordinating the global expression response to glucose in Bacillus subtilis and its comparison to Escherichia coli." Bmc Microbiology **9**: -.
Wade, J. T., D. C. Roa, et al. (2006). "Extensive functional overlap between sigma factors in Escherichia coli." Nature Structural & Molecular Biology **13**(9): 806-814.
Wall, D. P. and T. Deluca (2007). "Ortholog detection using the reciprocal smallest distance algorithm." Methods Mol Biol **396**: 95-110.
Wall, D. P., H. B. Fraser, et al. (2003). "Detecting putative orthologs." Bioinformatics **19**(13): 1710-1711.
Wang, J. Z., Z. Du, et al. (2007). "A new method to measure the semantic similarity of GO terms." Bioinformatics **23**(10): 1274-1281.
Weng, M., P. L. Nagy, et al. (1995). "Identification of the Bacillus subtilis pur operon repressor." Proceedings of the National Academy of Sciences of the United States of America **92**(16): 7455-7459.
Xenarios, I., L. Salwinski, et al. (2002). "DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions." Nucleic Acids Res **30**(1): 303-305.
Yachie, N., R. Saito, et al. (2011). "Integrative features of the yeast phosphoproteome and protein-protein interaction map." PLoS Comput Biol **7**(1): e1001064.
Yang, Z., Z. Lu, et al. (2001). "Study of adaptive mutations in Salmonella typhimurium by using a super-repressing mutant of a trans regulatory gene purR." Mutat Res **484**(1-2): 95-102.
Yang, Z., Z. Lu, et al. (2006). "Adaptive mutations in Salmonella typhimurium phenotypic of purR super-repression." Mutat Res **595**(1-2): 107-116.
Yeger-Lotem, E., L. Riva, et al. (2009). "Bridging high-throughput genetic and transcriptional data reveals cellular responses to alpha-synuclein toxicity." Nat Genet **41**(3): 316-323.

Yu, H. Y. and M. Gerstein (2006). "Genomic analysis of the hierarchical structure of regulatory networks." Proceedings of the National Academy of Sciences of the United States of America **103**(40): 14724-14731.
Zarrineh, P., A. C. Fierro, et al. "COMODO: an adaptive coclustering strategy to identify conserved coexpression modules between organisms." Nucleic Acids Res **39**(7): e41.
Zarrineh, P., A. C. Fierro, et al. (2011). "COMODO: an adaptive coclustering strategy to identify conserved coexpression modules between organisms." Nucleic Acids Res **39**(7): e41.
Zaslavskiy, M., F. Bach, et al. (2009). "Global alignment of protein-protein interaction networks by graph matching methods." Bioinformatics **25**(12): i259-267.
Zhang, R. and Y. Lin (2009). "DEG 5.0, a database of essential genes in both prokaryotes and eukaryotes." Nucleic Acids Research **37**: D455-D458.
Zhang, R., H. Y. Ou, et al. (2004). "DEG: a database of essential genes." Nucleic Acids Research **32**: D271-D272.
Zhu, J., B. Zhang, et al. (2008). "Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks." Nat Genet **40**(7): 854-861.

SUPPLEMENTARY TABLES

All the supplementary tables are downloadable online, and the explanation of the tables available on top of the tables.

Table S2.1: Table S2.1 is downloadable from the following link:

http://homes.esat.kuleuven.be/~kmarchal/Supplementary_Information_PhD_Peyman_2011/TableS2_1.xls

Table S2.2: Table S2.2 is downloadable from the following link:

http://homes.esat.kuleuven.be/~kmarchal/Supplementary_Information_PhD_Peyman_2011/TableS2_2.xls

Table S3.1: Table S3.1 is downloadable from the following link:

http://homes.esat.kuleuven.be/~kmarchal/Supplementary_Information_PhD_Peyman_2011/TableS3_1.xls

Table S3.2: Table S3.2 is downloadable from the following link:

http://homes.esat.kuleuven.be/~kmarchal/Supplementary_Information_PhD_Peyman_2011/TableS3_2.xls

APPENDIX A

Qualitative comparison of OSLOM PPI modules with MCL PPI modules

We have applied a recently developed community detection method called OSLOM (Lancichinetti, Radicchi et al. 2011) on protein-protein interactions data, described in (Peregrin-Alvarez, Xiong et al. 2009) to detect modules. Here we performed a qualitative comparison between the modules that the main authors have highlighted in their paper derived by MCL method and our corresponding modules detected by OSLOM. These modules include chemotaxis and flagella assembly (Figure A1), leucin, isoleucin and valine biosynthesis (Figure A2), pili assembly (Figure A3), and cell wall biosynthesis and cell division (Figure A4).

Chemotaxis and Flagella assembly

Components of chemotaxis of flagella assembly are organized within two distinct modules (3 and 15) in Peregrin-Alvarez results (Figure A1). OSLOM also found these two modules (module 84 and module 85) (Figure A1). OSLOM could detect three highly connected proteins cheY, motA, and motB (Figure A1 black nodes) in both modules. Note that fhiA and mbhA, are both pseudo-genes and we had removed them from the network before applying OSLOM.

Leucin, Isoleucin and Valine Biosynthesis

Leucin, isoleucin and valine biosynthesis are organized within three distinct modules (45, 66 and 203) in Peregrin-Alvarez results (Figure A2). OSLOM found the proteins involved in leucin, isoleucin and valine biosynthesis just in one module (module 61) (Figure A2). Interestingly, OSLOM could find few more proteins related to other carbohydrate processes and transport (edd, gcl, icd, oxc, uhpC) and a two-component signaling pathway in the mentioned module (uhpABT). OSLOM did not detect two lower connected proteins yfdU and ybhJ.

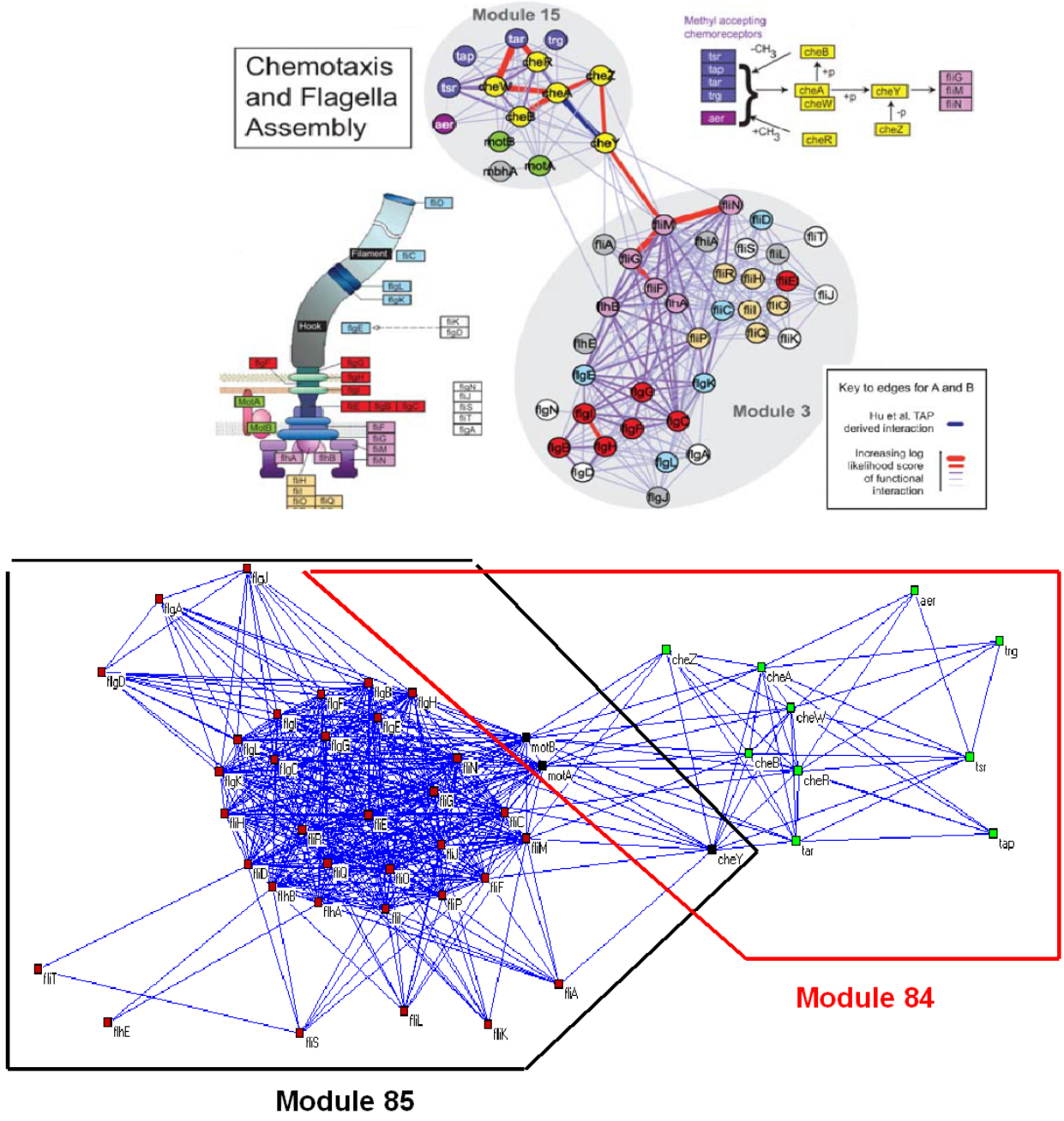


Figure A1. Modules related to cheomatix and Flagella assembly reported in Peregrin-Alvarez results (top) and OSLOM results (bottom). OSLOM could find three highly connected proteins cheY, motA, and motB (black nodes) in both modules. fhiA and mbhA, are both pseudo-genes and we had removed them from the network before

applying OSLOM. Peregrin-Alvarez modules visualization (top) are taken from their paper (Peregrin-Alvarez, Xiong et al. 2009).

Pili Assembly

Pili Assembly is organized within two distinct modules (21 and 35) in Peregrin-Alvarez results (Figure A3), while OSLOM could detect them in one module (module 86). This module contains yhcD which is predicted outer membrane protein. Two Pili-like proteins were reported in one of the Pili Assembly module (module 21) of Peregrin-Alvarez. In contrast, OSLOM could detect four Pili-like proteins in a separate module (module 78) (Figure A3). The fact that OSLOM could separate these two groups proves its higher accuracy.

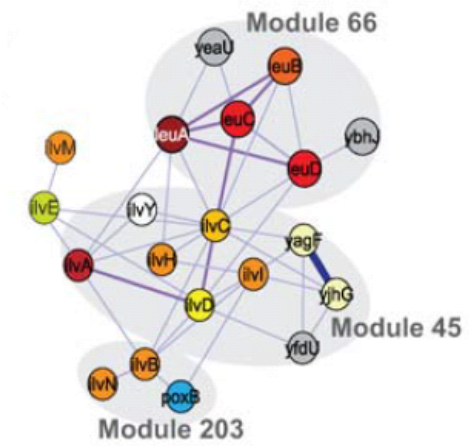
Cell Wall Biosynthesis and Cell Division

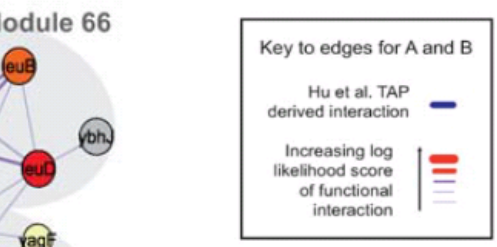
Cell wall biosynthesis and cell division related genes are organized within 10 distinct modules (2, 9, 38, 63, 132, 205, 206, 207, 209, and 247) in Peregrin-Alvarez results (Figure A4). The related proteins are organized in five distinct modules (1, 62, 64, 102, and 111) in OSLOM results (Figure A4). Modules 9, 38, 205, and 207 of Peregrin-Alvarez contain proteins similar to module 111 of OSLOM. OSLOM could detect ftsE, ftsX, and secA more than Peregrin-Alvarez while it did not include low connected relA and atp which do not seem relevant. Modules 132 of Pregerin-Alvarez and 62 OSLOM are identical. Module 132 of Pregerin-alvarez is completely entailed in module 101 of OSLOM. Loosely connected module 207 of Pregerin-alvarez were not detected by OSLOM. Module 38 of Pregerin-alvarez and module 1 of OSLOM are fairly similar but OSLOM find prokaryotic protein translocation apparatus which comprise secA, secB, secD, secE, secF, secG and secY all together. In this example also OSLOM shows that it does not detect loosely connected nodes in the network, and the results of OSLOM were more relevant.

General conclusions

Comparing the detected proteins by two methods, we can conclude that in general OSLOM modules are larger, and sometimes two or three MCL modules are related to one OSLOM module. In addition, OSLOM modules are highly connected, and loosely connected proteins

usually do not appear in its results. Furthermore, highly connected proteins can appear in more than one module.





Leucine, Isoleucine and Valine Biosynthesis

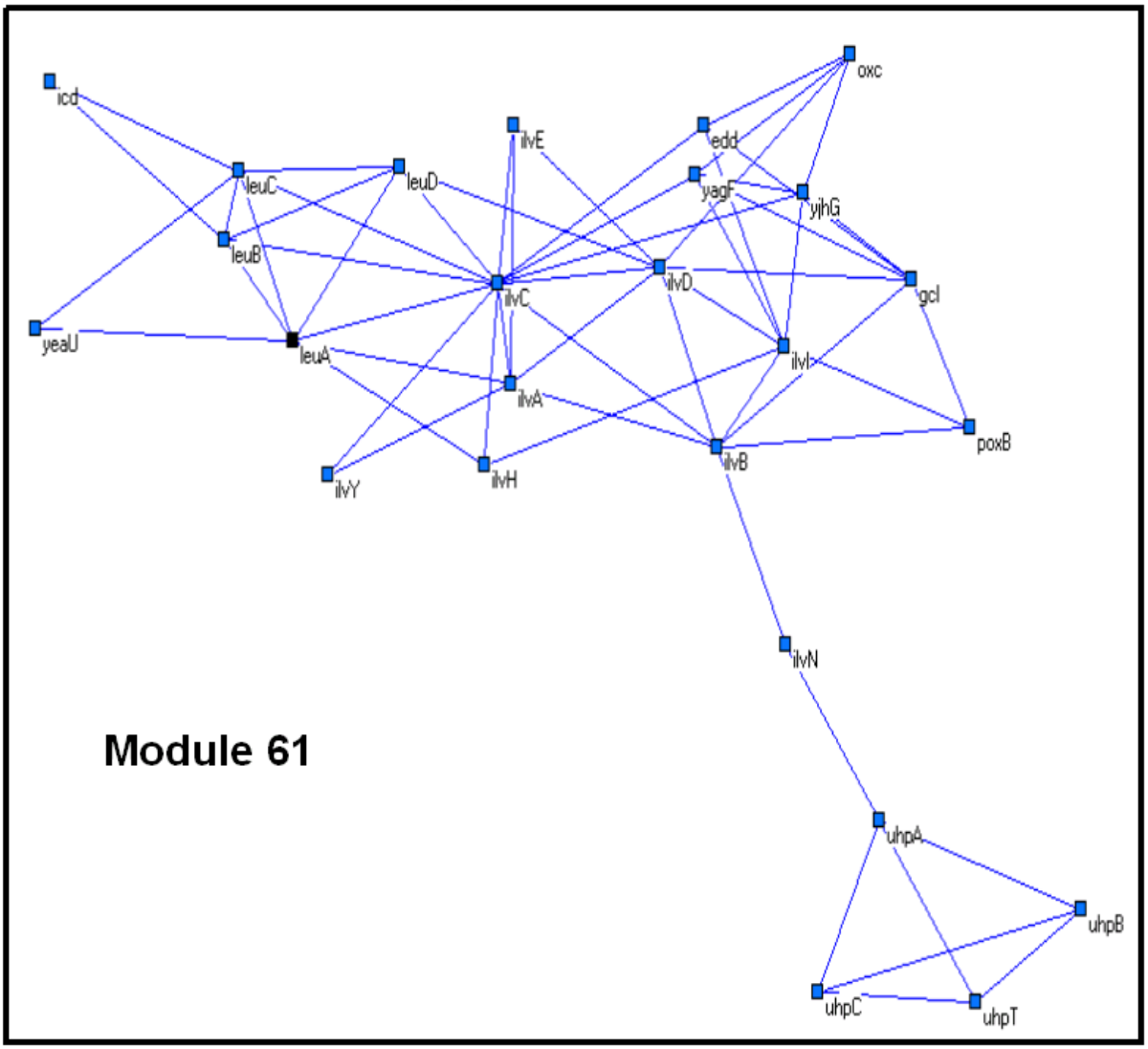


Figure A2. Leucin, Isoleucin and Valine Biosynthesis related modules (modules 45, 66 and 203) in Peregrin-Alvarez results (top) and similar module detected by OSLOM (module 61-bottom). OSLOM could find few more proteins related to other carbohydrate processes and transport (edd, gcl, icd, oxc, uhpC) and a two-component signaling pathway in the mentioned module (uhpABT). OSLOM could detect ilvE, but OSLOM did not detect two lower connected proteins yfdU and ybhJ. Peregrin-Alvarez modules visualization (top) are taken from their paper (Peregrin-Alvarez, Xiong et al. 2009).

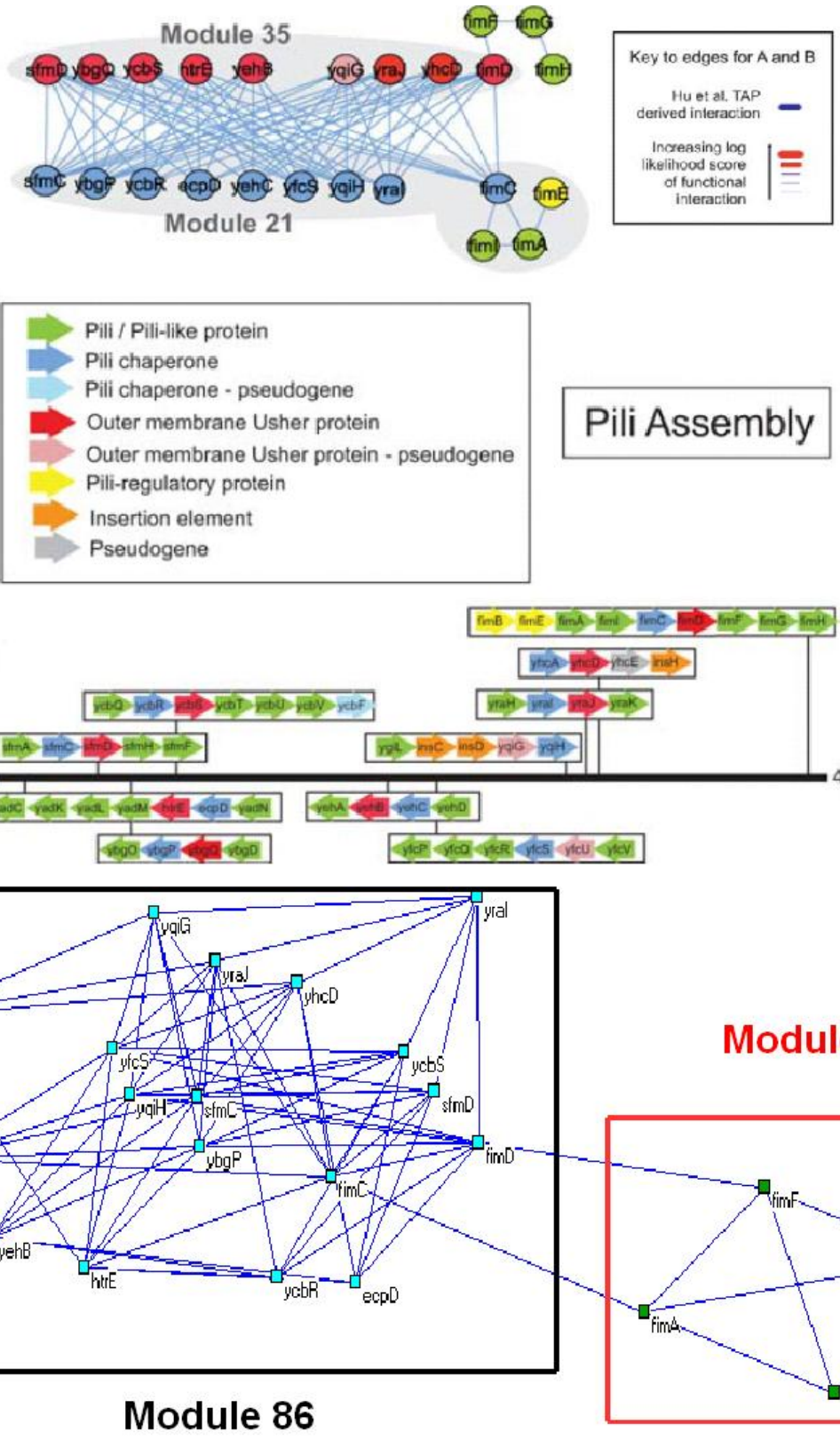
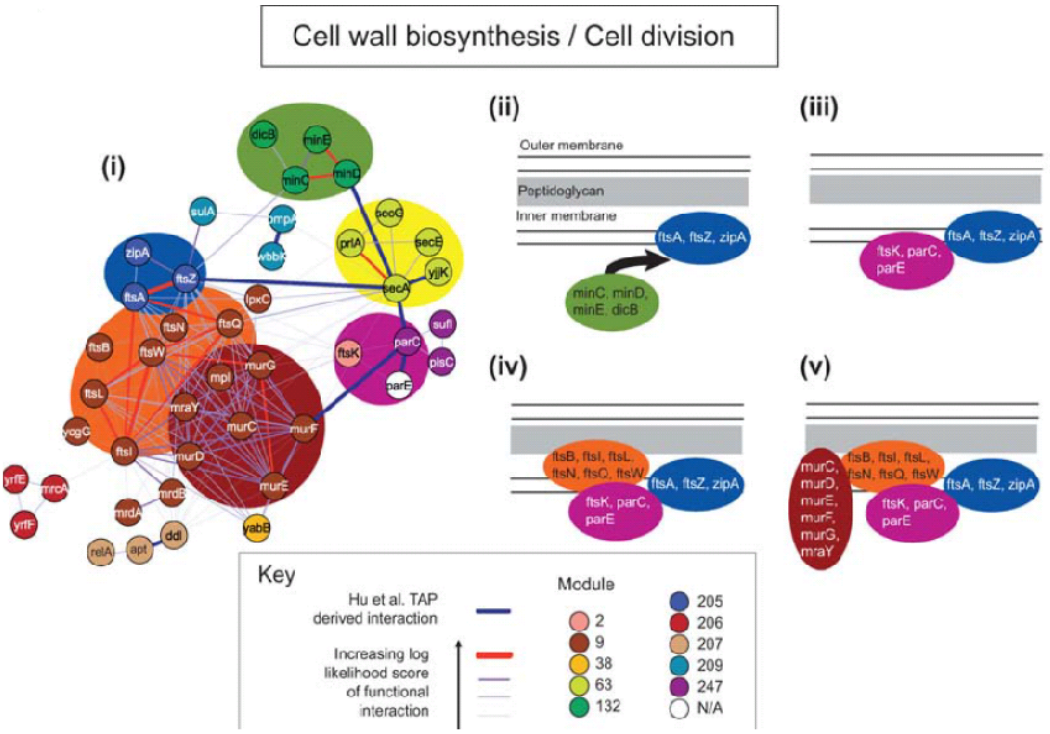

Module 35
sfmD ybgQ ycbS htrE yehB yqiG yraJ yhcD fimD fimH
sfmC ybgP ycbR ecpD yehC yfcS yqiH yraI fimC fimE
fimI fimA
Module 21
fimF fimG
Key to edges for A and B
Hu et al. TAP derived interaction
Increasing log likelihood score of functional interaction
Pili / Pili-like protein
Pili chaperone
Pili chaperone - pseudogene
Outer membrane Usher protein
Outer membrane Usher protein - pseudogene
Pili-regulatory protein
Insertion element
Pseudogene
Pili Assembly
0 nt
4,639,675 nt
yraI
yqiG
yraJ
yhcD
yehC
yfcS
ycbS
yqiH
sfmC
sfmD
fimD
ybgQ
ybgP
fimC
yehB
htrE
ycbR
ecpD
Module 87
fimF
fimH
fimA
fimG
Module 86

Figure S3. Pili Assembly is organized within two distinct modules (21 and 35) in Peregrin-Alvarez results (top), while OSLOM could detect them in one module (module 86-bottom). Two Pili-like proteins were reported in one of the Pili Assembly module (module 21) of Peregrin-Alvarez. In contrast, OSLOM could detect four Pili-like proteins in a separate module (module 78). Peregrin-Alvarez modules visualization (top) are taken from their paper (Peregrin-Alvarez, Xiong et al. 2009).


Cell wall biosynthesis / Cell division
(i)
(ii)
(iii)
(iv)
(v)
Outer membrane
Peptidoglycan
Inner membrane
ftsA, ftsZ, zipA
minC, minD, minE, dicB
ftsK, parC, parE
ftsB, ftsI, ftsL, ftsN, ftsQ, ftsW
murC, murD, murE, murF, murG, mraY
Key
Hu et al. TAP derived interaction
Increasing log likelihood score of functional interaction
Module
2
9
38
63
132
205
206
207
209
247
N/A

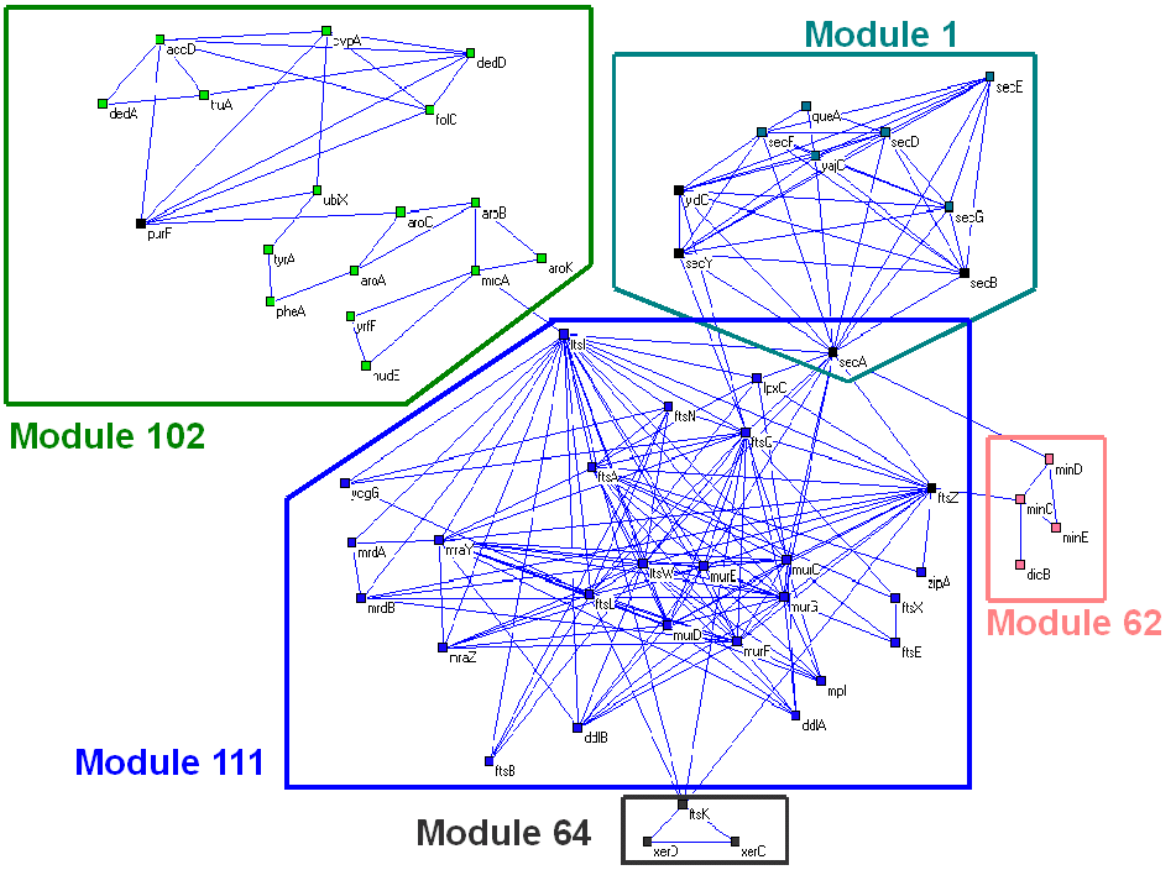

Module 1
Module 102
Module 111
Module 64
Module 62
accD
cypA
dedD
dedA
tuA
folC
purF
ubiX
aroC
aroB
tyrA
aroA
micA
aroK
pheA
yrfF
secE
queA
secF
secD
yajC
ydiC
secG
secY
secB
secA
ftsI
lpxC
ftsN
ftsQ
ftsA
ftsZ
ycgG
mraY
mrdA
ftsW
murE
murC
zipA
mrdB
ftsL
murG
ftsX
muiD
murF
ftsE
mpl
ddlA
ddlB
ftsB
minD
minC
minE
dicB
ftsK
xerD
xerC

Figure A4. Cell wall biosynthesis and cell in Peregrin-Alvarez results (top) and OSLOM (bottom). Cell wall biosynthesis and cell division related genes are organized within 10 distinct modules (2, 9, 38, 63, 132, 205, 206, 207, 209, and 247) in Peregrin-Alvarez results. The related proteins are organized in five distinct modules (1, 62, 64, 102, and 111) in OSLOM results. Peregrin-Alvarez modules visualization (top) are taken from their paper (Peregrin-Alvarez, Xiong et al. 2009).