Understand biological regulatory systems using computational models: Reconstruction, Analysis and Integration

by

Yao Fu

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Bioinformatics and Computational Biology

Program of Study Committee:
Julie A. Dickerson, Co-major Professor
Laura Jarboe, Co-major Professor
Marna D. Nelson
Jacqueline V. Shanks
Reuben J. Peters

Iowa State University
Ames, Iowa
2013

TABLE OF CONTENTS

LIST OF FIGURES

Page

Page

LIST OF TABLES

Page

NOMENCLATURE

TFA	Transcription Factor Activity
TF	Transcription Factor
NCA	Network Component Analysis
PLS	Partial Least Square
EM	Expecation Maximization
GRN	Gene Regulatory Network
ETM	External Transportable Metabolite
ETC	External Two-Component
ISM	Internal Signal Metabolite
IDB	Internal DNA-Binding
ERN	Effective Regulatory Network
GlRN	Global Regulatory Network

ACKNOWLEDGEMENTS

I would like to thank my major professor Dr. Julie Dickerson, my co-major professor Dr. Laura Jarboe and my committee members, Dr. Oliver Eulenstein, Dr. Jacqueline Shanks, and Dr. Reuben Peters, as well as my former committee member Dr. Vasant Honavar, for their guidance and support throughout the course of this research.

In addition, I would also like to thank my friends, especially Fuyuan Jing, Yiming Zhang, Hsien-Chao Chou, Le Zhao, colleagues, Jesse Walsh, Erin Bogeess, Liam Royce, Ping Liu, Ting Wei Tee, Jong Moon Yoon, and all other students, staff and faculties in Thrust 2 of the Center for Biorenewable Chemicals, the BCB graduate program coordinator Trish Stauble, the department faculty and staff for making my time at Iowa State University a wonderful experience.

Finally, thanks to my parents, Qiuan Li and Aimin Fu, my sister Yan Li, and all of my family members for their support and encouragement, and to my girlfriend Qian (Serena) Xu for her hours of patience, respect and love.

ABSTRACT

Biological regulatory system is complex and involves many types of interactions, including transcriptional regulations, protein interactions, metabolic reactions and etc., to ensure the regulations of biological organisms. These regulations forms complex networks and play important roles in living organisms to adapt to the environment, control the rate of growth, and develop different phenotypes accordingly to its life cycle and the surrounding environment. Many of mechanisms and interactions of these networks are still not clear. Although better understanding of the regulatory systems is very important for biological research and engineering, to systematically reconstruct, analyze and integrate the complex regulatory systems is always challenging.

At first, a novel method to reconstruct gene regulatory networks (GRNs) was developed, implemented, tested, and applied to experimental data. This method introduced a hidden transcription factor activity (TFA) layer to the conventional GRN reconstruction methods. The testing results showed significantly improved network reconstruction precision and recall comparing to conventional methods. The Application to *E. coli* transcriptome experimental data demonstrated the potential biological significance of the reconstructed network.

A three level analysis framework to analyze TFAs and GRNs under different experimental conditions was followed up. The first level analyzes TFA patterns of individual transcription factors. The second level uses enrichment test and summarizes TFA behaviors by groups and their properties. The third level identifies key TFs of each

experimental condition using network based analysis approach on effective regulatory network (ERN), a newly proposed differencial regulatory network model between experimental conditions. This analysis framework expands the traditional transcriptome data analysis to TFA and GRN level. The application to *E. coli* data showed the biological meaningfulness and helpfulness of analyzing transcriptome data on TFA and GRN level.

At last, a comprehensive regulatory focused regulatory system model for *E. coli* had been constructed by integrating transcriptional regulatory networks, protein interaction networks, metabolic reaction networks, and all other related regulations. Statistical tests and network property analysis of this constructed network revealed the connection between biological functions and the special network properties of the constructed network. And simulations of the regulatory signal response of this constructed network verified the biological meaningfulness of this network.

CHAPTER I

GENERAL INTRODUCTION

Role of Regulatory Systems

A successful biological organism should be able to regulate itself to adapt to the environment, control the rate of growth, and develop different phenotypes accordingly to its life cycle and the surrounding environment. Biological regulatory system is complex and involves many types of interactions to ensure the successful efficient and robust regulations of biological organisms. Take a simplest example, in unicellular organism, regulatory system receive internal and external stimulate signals through molecular transportations, metabolic reactions, metabolites-protein interactions or protein-protein interactions, change biological functions of metabolic pathways, enzymatic activities and protein activities through these interactions, regulate gene expression through transcription factor proteins, sigma factors and small regulatory RNAs (sRNAs), and regulate the abundance of specific functioning proteins through gene expressions and protein related interactions. All these different types of interactions form complex networks including transcriptional networks, protein interaction networks, metabolic reaction networks and etc. Many of mechanisms and interactions of these networks are still not clear. Although better understanding of the regulatory systems is very important for biological research and engineering, to systematically integrate and model the complex regulatory systems is always challenging.

Gene Regulation

Gene expressions are regulated by several types of interactions, including transcription factor (TF) - gene interactions, Sigma factor - gene interactions, and sRNA - gene interactions. Sigma factors are usually large protein complexes constructed by many gene products and regulate very large amount of genes to globally control the gene expressions under environmental changes or different growth phases [1]. Recently, the importance of sRNA's post-transcriptional regulatory function of genes has been recognized as sRNAs may bind with mRNAs to regulate the expression of genes [2].

In the gene regulation process, an active transcription factor (TF) can bind DNA and control gene expression. However, many TFs are not inherently active. Complex mechanisms, such as forming dimmers, interacting with signal metabolites or binding specific micro-RNAs, are needed in order to control the activities of these TFs [3]. The activities of TFs also differ in different environments or during specific periods of cell development. This activation level is called transcription factor activity (TFA) [3]. Thus, TFA is an essential component of gene regulatory networks. It regulates gene expression in response to internal and external signals to ensure appropriate gene expression and forms relatively much more complex and larger gene regulatory networks than Sigma factor - gene interaction and sRNA - gene interaction networks. Thus, in this study, methodology studies of network reconstruction and analysis are more focused on TF-gene regulations.

Since TFA is governed by various complex molecular interactions, it is difficult to determine directly from experiments, especially if the activation mechanism is unknown. However, it is possible to computationally predict the change of TFAs relative to a reference state using transcriptome data and a known TF-gene network architecture [3, 4]. Network Component Analysis (NCA) developed by Liao et al. defines the problem of calculating TFAs as optimization of a linear least square matrix decomposition. Liao et al. solve the problem using an expectation maximization (EM) approach [5]. Fast Network Component Analysis (FastNCA) uses singular value decomposition (SVD) and a matrix projection technique to approximate the linear least square matrix decomposition problem defined in NCA[6]. ChIP data provides additional information on proteins' DNA binding occupancy. Gao et al. developed an algorithm that combines microarray data for mRNA expression and transcription factor occupancy to define the regulatory network (MA-Networker algorithm) to predict TFAs based on ChIP and transcriptome data using multivariate regression and backward variable selection [7]. With the predicted TFAs, Gao et al. calculate the TF-gene coupling factor using Pearson Correlation [7]. Boulesteix et al. applied statistically inspired modification of the partial least square (SIMPLS) algorithm to find TFAs [8]. Many more complex models are also applied to predict TFAs. For example, Bayesian Network approach [9], state-space model [10], probabilistic dynamical models [11], and Gaussian process model [12]. Besides predicting TFAs from gene expression data and TF network structures from experiments and literature data, DNA sequence motif information is also widely used (e.g. searching for DNA binding site of TFs) in many methods to infer

potential TF-gene links to obtain a more complete TF network structure and improve the prediction of TFAs [4]. However, compared to matrix decomposition and regression approaches, these complex models require more computational power. Thus, these complex models either cannot deal with large scale TFAs or they predict large scale TFAs by converting TFAs into binary.

High-throughput technologies have led to many algorithms for the reconstruction of large scale gene regulatory networks [13]. For example, many sequence analysis approaches which identify potential TF binding sites have been developed [14]. However, many of the predicted potential TF binding sites are not functional (false positive predictions) [13]. From ChIP-chip technology, potential gene regulatory effects can be derived by identifying the portions of a genome that are bound by a particular TF *in vivo* [15]. Transcriptome data (also known as gene expression data) measured by genome-wide DNA microarrays are widely used for gene regulatory network reconstructions. For instance, Stuart et al. use correlation coefficients between mRNA levels of genes as relevance scores to reconstruct correlation networks [16]. The interacting genes are predicted by detecting the correlation score above some set threshold. Other algorithms such as RELNET (RELevance NETworks) [17] and ARACNE (Algorithm for the Reverse engineering of Accurate Cellular NEtworks) [18] use mutual information as the relevance scores. The CLR (Context Likelihood Relatedness) [11] algorithm uses an adaptive background correction method on the relevance scores to improve precisions [19].

Protein Interactions

Protein interactions, also known as protein binding interactions, include protein-protein interactions and protein-ligand interactions. Proteins can interact with other proteins or the same types of proteins to form protein complexes and perform certain biological functions, such as regulate gene expressions, catalyze enzymatic reactions, transport certain metabolic molecules and etc. Similarly, proteins may also bind with other small molecules, known as protein-ligand interactions, to activate or silent its biological function of serving as TFs, enzymes, transporters and etc. Proteins are one of the most important intermediate regulators connecting different cellular components, as well as regulating metabolite transportations, metabolic reactions and gene expressions [20, 21].

Like in transcriptional networks, protein interaction networks can also be constructed directly from experiments or inferred from high-throughput experiment data [22]. Yeast two-hybrid assay is one of the most widely used experiments to directly examine protein-protein interactions [23]. Many protein-protein interaction networks of many species have been constructed using this method, e.g., *helicobacter pylori, Drosophila melanogaster* (fruit fly), *C. elegans, and homo sapiens* (human) [24-28]. However, yeast two-hybrid methods can only detect interactions between two proteins, and not be able to identify interactions among three or more proteins [22]. A large-scale tandem affinity purification method coupled to mass spectrometry (TAP-MS) was developed and able to study multiple interactions of proteins. This method is applied on

E. coli and Yeast to identify protein-protein interactions [29-31]. Recently, many other methods such as protein fragment complementation assay (PCA) [32], matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF-MS) [33] and etc. have been used to construct more comprehensive protein-protein networks. Computational method are also developed to predict protein-protein interactions from shared characteristics of known interactions [34] or phylogenetic evolutionary information [35]. High-throughput proteomics experiment data such as 2D-PAGE and mass spectrometry based proteomics data can also be utilized to infer protein-protein interactions [36, 37]. And many molecular modeling techniques are also used in predicting, evaluating, and studying protein-ligand and protein-protein interactions [38-40].

Metabolic Reactions

Metabolic reactions include chemical reactions, enzyme catalyzed reactions, transport reactions and etc. These reactions connect metabolites into complex and functional networks and metabolic pathways to perform assimilatory and catabolic functions. And this metabolic feature is one of the most important properties of living organisms. These metabolic reactions are regulated by their enzymatic activities, protein activities, reaction co-factors, or even the abundance of their reactants and products. Modeling metabolic systems and pathways can help on better understanding the

biological metabolic process and support metabolic engineering to produce and optimize production of biorenewable chemicals.

Computational models have been developed to model metabolic reactions in the cell. Metabolic network models majorly model chemical reactions, including enzymatic reactions and transport reactions in the cell into an interaction network. By converting the metabolic reaction network into stoichiometric matrix, it is possible to find the solution space of all the feasible metabolic flux pathways using extreme pathways model[41]. Similarly, steady states based elementary mode analysis can identify a unique set of functioning smallest sub-networks that perform metabolic network functions[42]. Unlike extreme pathways and elementary mode, linear programming based flux balance analysis helps to find a optimized solution of the metabolic network model given certain interested objective function[43]. All these analysis are based on well defined metabolic networks, such as iAF1260, which was constructed by Feist et al. in 2007 for E. coli MG1655[44].

Recent comprehensive databases, such as Biocyc [45], KEGG [46] and etc. collect these metabolic reactions as properties of proteins or biological pathways. And genome-scale metabolomics data produced from mass spectrometry based methods and flux analysis brought us possibilities to predict this type of interactions systematically by integrating these data with other types of high-throughput data.

Interacting Regulatory Systems

Components of regulatory systems described above are not only interacting with components within the same system, but also connected with components in each other systems. **Figure 1** summarizes the complex interactions among different types of cellular components and regulatory systems. For example, expression of genes could be regulated by transcription factors, and the activities of transcription factors are controlled by protein-protein or protein-metabolite interaction. Meantime, the abundance of proteins are regulated by gene expressions, and the abundance of metabolites are directly related to metabolic reactions occurs in the cell, most which are regulated by enzymatic activities of enzyme proteins, or transporter activities of transporter proteins.

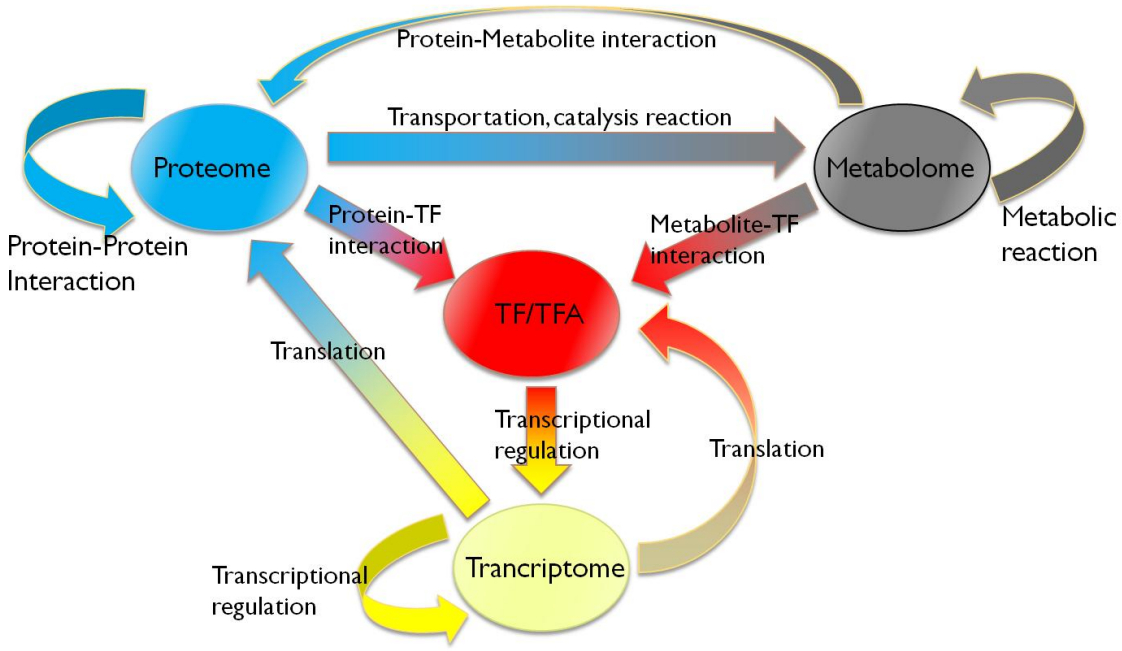


Figure 1-1 *Interactions between regulatory systems and cellular components*

Integrating all these types of systems discussed above and other related interactions is becoming a hot topic in system biology area in recent years. However, large genome-wide integrations of all the systems in the cell are still limited in knowledge based integrations, such as Reactome[47] and Ecocyc [48], or only have several selected systems being integrated for to perform specific analysis and modeling as mentioned above.

Efforts on modeling regulatory signals across different layers of interaction networks have also been taken. For example, the two-component signal transduction systems are being well studied[49]. And RegulonDB recently proposed and constructed 25 Genetic sensory-response units model of E. coli encountering signal, the signal-to-effect reaction end with activation/deactivation of TF, the regulatory swathes, and the consequences to model signals cross multiple interaction layers[50]. However, these regulatory signaling models only cover one specific type of signaling system or small un-connected parts of the whole cell system.

Goal of this work

The ultimate goal of this project is to develop a framework to reconstruct, analyze and model biological regulatory system integrating different layers of regulatory information, including data generated information and related knowledge based information from transcriptomics, proteomics and metabolomics experiments. The project would develop models to help biologist on better understanding of connections

among regulatory systems, identifying target regulatory systems for research and engineering specifically to their targeting environments, and predicting regulatory effects from the environment changes or mutations.

To achieve the ultimate goal, steps and computational experiments are taken towards our modeling organism, *E. coli* MG1655 (K-12). *E. coli* MG1655 is a well characterized organism with fully sequenced genome, well studied transcriptional regulatory networks, protein binding interactions, metabolic reactions and other information stored in public available database such as RegulonDB[50] and Ecocyc[48].

Three major steps have been discussed in this work. Chapter 1 started from reconstruction of gene regulatory networks, followed by Chapter 2, analysis methods of the dynamic of the reconstructed gene regulatory networks, and finally, Chapter 3 proposed a comprehensive model of regulatory system of *E. coli* integrating transcriptional regulatory systems, protein interactions, metabolic reactions and pathways, and all other related regulations.

REFERENCES

1. Loewen, P.C. and R. Hengge-Aronis, *The role of the sigma factor sigma S (KatF) in bacterial global regulation.* Annu Rev Microbiol, 1994. **48**: p. 53-80.

2. Shimoni, Y., et al., *Regulation of gene expression by small non-coding RNAs: a quantitative view.* Mol Syst Biol, 2007. **3**: p. 138.

3. Liao, J.C., et al., *Network component analysis: Reconstruction of regulatory signals in biological systems.* Proceedings of the National Academy of Sciences of the United States of America, 2003. **100**(26): p. 15522-15527.

4. Bussemaker, H.J., B.C. Foat, and L.D. Ward, *Predictive Modeling of Genome-Wide mRNA Expression: From Modules to Molecules.* Annual Review of Biophysics and Biomolecular Structure, 2007. **36**(1): p. 329-347.

5. Tran, L.M., et al., *gNCA: A framework for determining transcription factor activity based on transcriptome: identifiability and numerical implementation.* Metabolic Engineering, 2005. **7**(2): p. 128-141.

6. Chang, C., et al., *Fast network component analysis (FastNCA) for gene regulatory network reconstruction from microarray data.* Bioinformatics, 2008. **24**(11): p. 1349-1358.

7. Gao, F., B. Foat, and H. Bussemaker, *Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data %U http://www.biomedcentral.com/1471-2105/5/31.* BMC Bioinformatics, 2004. **5**(1 %M doi:10.1186/1471-2105-5-31): p. 31.

8. Boulesteix, A.-L. and K. Strimmer, *Predicting transcription factor activities from combined analysis of microarray and ChIP data: a partial least squares approach.* Theoretical Biology and Medical Modelling, 2005. **2**(1 %M doi:10.1186/1742-4682-2-23): p. 23.

9. Nachman, I., A. Regev, and N. Friedman, *Inferring quantitative models of regulatory networks from expression data.* Bioinformatics, 2004. **20**(suppl_1): p. i248-256.

10. Li, Z., et al., *Using a state-space model with hidden variables to infer transcription factor activities.* Bioinformatics, 2006. **22**(6): p. 747-754.

11. Sanguinetti, G., M. Rattray, and N.D. Lawrence, *A probabilistic dynamical model for quantitative inference of the regulatory mechanism of transcription.* Bioinformatics, 2006. **22**(14): p. 1753-1759.

12. Gao, P., et al., *Gaussian process modelling of latent chemical species: applications to inferring transcription factor activities.* Bioinformatics, 2008. **24**(16): p. i70-75.

13. Hecker, M., et al., *Gene regulatory network inference: Data integration in dynamic models--A review.* Biosystems, 2009. **96**(1): p. 86-103.

14. Vlieghe, D., et al., *A new generation of JASPAR, the open-access repository for transcription factor binding site profiles.* Nucleic Acids Research, 2006. **34**(suppl_1): p. D95-97.

15. Ren, B., et al., *Genome-Wide Location and Function of DNA Binding Proteins.* Science, 2000. **290**(5500): p. 2306-2309.

16. Stuart, J.M., et al., *A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules.* Science, 2003. **302**(5643): p. 249-255.

17. Butte, A.J. and I.S. Kohane, *Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements*, 2000.

18. Basso, K., et al., *Reverse engineering of regulatory networks in human B cells.* Nat Genet, 2005. **37**(4): p. 382-390.

19. Faith, J.J., et al., *Large-Scale Mapping and Validation of Escherichia coli Transcriptional Regulation from a Compendium of Expression Profiles.* PLoS Biol, 2007. **5**(1): p. e8.

20. Pawson, T. and P. Nash, *Assembly of Cell Regulatory Systems Through Protein Interaction Domains.* Science, 2003. **300**(5618): p. 445-452.

21. Wilchek, M., E.A. Bayer, and O. Livnah, *Essentials of biorecognition: The (strept)avidin–biotin system as a model for protein–protein and protein–ligand interaction.* Immunology Letters, 2006. **103**(1): p. 27-32.

22. Kim, T.Y., H.U. Kim, and S.Y. Lee, *Data integration and analysis of biological networks.* Curr Opin Biotechnol, 2010. **21**(1): p. 78-84.

23. Uetz, P., et al., *A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae.* Nature, 2000. **403**(6770): p. 623-7.

24. Rain, J.C., et al., *The protein-protein interaction map of Helicobacter pylori.* Nature, 2001. **409**(6817): p. 211-5.

25. Giot, L., et al., *A protein interaction map of Drosophila melanogaster.* Science, 2003. **302**(5651): p. 1727-36.

26. Lamesch, P., et al., *C. elegans ORFeome version 3.1: increasing the coverage of ORFeome resources with improved gene predictions.* Genome Res, 2004. **14**(10B): p. 2064-9.

27. Stelzl, U., et al., *A human protein-protein interaction network: a resource for annotating the proteome.* Cell, 2005. **122**(6): p. 957-68.

28. Rual, J.F., et al., *Towards a proteome-scale map of the human protein-protein interaction network.* Nature, 2005. **437**(7062): p. 1173-8.

29. Butland, G., et al., *Interaction network containing conserved and essential protein complexes in Escherichia coli.* Nature, 2005. **433**(7025): p. 531-7.

30. Krogan, N.J., et al., *Global landscape of protein complexes in the yeast Saccharomyces cerevisiae.* Nature, 2006. **440**(7084): p. 637-43.

31. Gavin, A.C., et al., *Proteome survey reveals modularity of the yeast cell machinery.* Nature, 2006. **440**(7084): p. 631-6.

32. Arifuzzaman, M., et al., *Large-scale identification of protein-protein interaction of Escherichia coli K-12.* Genome Res, 2006. **16**(5): p. 686-91.

33. Tarassov, K., et al., *An in vivo map of the yeast protein interactome.* Science, 2008. **320**(5882): p. 1465-70.

34. Ben-Hur, A. and W.S. Noble, *Kernel methods for predicting protein-protein interactions.* Bioinformatics, 2005. **21 Suppl 1**: p. i38-46.

35. Pazos, F., et al., *Assessing protein co-evolution in the context of the tree of life assists in the prediction of the interactome.* J Mol Biol, 2005. **352**(4): p. 1002-15.

36. Wan, X.Y. and J.Y. Liu, *Comparative proteomics analysis reveals an intimate protein network provoked by hydrogen peroxide stress in rice seedling leaves.* Mol Cell Proteomics, 2008. **7**(8): p. 1469-88.

37. Gstaiger, M. and R. Aebersold, *Applying mass spectrometry-based proteomics to genetics, genomics and network biology.* Nat Rev Genet, 2009. **10**(9): p. 617-27.

38. Cerqueira, N.M., et al., *MADAMM: a multistaged docking with an automated molecular modeling protocol.* Proteins, 2009. **74**(1): p. 192-206.

39. Antony, J., et al., *Protein-ligand interaction energies with dispersion corrected density functional theory and high-level wave function based methods.* J Phys Chem A, 2011. **115**(41): p. 11210-20.

40. Chen, C.Y., *Weighted equation and rules--a novel concept for evaluating protein-ligand interaction.* J Biomol Struct Dyn, 2009. **27**(3): p. 271-82.

41. Price, N.D., et al., *Network-based analysis of metabolic regulation in the human red blood cell.* J Theor Biol, 2003. **225**(2): p. 185-94.

42. Papin, J.A., et al., *Comparison of network-based pathway analysis methods.* Trends Biotechnol, 2004. **22**(8): p. 400-5.

43. Stelling, J., et al., *Metabolic network structure determines key aspects of functionality and regulation.* Nature, 2002. **420**(6912): p. 190-3.

44. Feist, A.M., et al., *A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information.* Mol Syst Biol, 2007. **3**: p. 121.

45. Keseler, I.M., et al., *EcoCyc: a comprehensive database of Escherichia coli biology.* Nucleic Acids Res, 2011. **39**(Database issue): p. D583-90.

46. Kanehisa, M., *The KEGG database.* Novartis Found Symp, 2002. **247**: p. 91-101; discussion 101-3, 119-28, 244-52.

47. Jupe, S., et al., *Reactome - a curated knowledgebase of biological pathways: megakaryocytes and platelets.* J Thromb Haemost, 2012.

48. Keseler, I.M., et al., *EcoCyc: fusing model organism databases with systems biology.* Nucleic Acids Res, 2013. **41**(Database issue): p. D605-12.

49. Capra, E.J. and M.T. Laub, *Evolution of two-component signal transduction systems.* Annu Rev Microbiol, 2012. **66**: p. 325-47.

50. Salgado, H., et al., *RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more.* Nucleic Acids Res, 2013. **41**(Database issue): p. D203-13.

CHAPTER II

RECONSTRUCTING GENOME-WIDE REGULATORY NETWORK OF *E. COLI* USING TRANSCRIPTOME DATA AND PREDICTED TRANSCRIPTION FACTOR ACTIVITIES

A paper published in *BMC Bioinformatics*

Yao Fu[1], Laura R Jarboe[2] and Julie Dickerson[1, 3§*]

Abstract

Background

Gene regulatory networks play essential roles in living organisms to control growth, keep internal metabolism running and respond to external environmental changes. Understanding the connections and the activity levels of regulators is important for the research of gene regulatory networks. While relevance score based algorithms that reconstruct gene regulatory networks from transcriptome data can infer genome-wide

[1]Bioinformatics and Computational Biology Program, Iowa State University, Ames, Iowa, U.S.A

[2]Chemical and Biological Engineering Department, Iowa State University, Ames, Iowa, U.S.A

[3]Electrical and Computer Engineering Department, Iowa State University, Ames, Iowa, U.S.A

[§]Corresponding author

gene regulatory networks, they are unfortunately prone to false positive results. Transcription factor activities (TFAs) quantitatively reflect the ability of the transcription factor to regulate target genes. However, classic relevance score based gene regulatory network reconstruction algorithms use models do not include the TFA layer, thus missing a key regulatory element.

Results

This work integrates TFA prediction algorithms with relevance score based network reconstruction algorithms to reconstruct gene regulatory networks with improved accuracy over classic relevance score based algorithms. This method is called Gene expression and Transcription factor activity based Relevance Network (GTRNetwork). Different combinations of TFA prediction algorithms and relevance score functions have been applied to find the most efficient combination. When the integrated GTRNetwork method was applied to *E. coli* data, the reconstructed genome-wide gene regulatory network predicted 381 new regulatory links. This reconstructed gene regulatory network including the predicted new regulatory links show promising biological significances. Many of the new links are verified by known TF binding site information, and many other links can be verified from the literature and databases such as EcoCyc. The reconstructed gene regulatory network is applied to a recent transcriptome analysis of *E. coli* during isobutanol stress. In addition to the 16 significantly changed TFAs detected in the original paper, another 7 significantly changed TFAs have been detected by using our reconstructed network.

Conclusion

The GTRNetwork algorithm introduces the hidden layer TFA into classic relevance score-based gene regulatory network reconstruction processes. Integrating the TFA biological information with regulatory network reconstruction algorithms significantly improves both detection of new links and reduces that rate of false positives. The application of GTRNetwork on *E. coli* gene transcriptome data gives a set of potential regulatory links with promising biological significance for isobutanol stress and other conditions.

Background

Gene regulatory networks play an essential role in controlling gene expression and ensuring that the right genes are expressed or silenced at the right time in the right place to make the organism function appropriately. Better understanding of gene regulatory structure aids biological researchers and biochemical engineers in obtaining more complete views of the complex gene expression and regulatory mechanisms in organisms.

In the gene regulation process, an active transcription factor (TF) can bind DNA and control gene expression. However, many TFs are not inherently active. Complex mechanisms, such as forming dimers, interacting with signal metabolites or binding specific micro-RNAs, are needed in order to control the activities of these TFs [1]. The activities of TFs also differ in different environments or during specific periods of cell development. This activation level is called transcription factor activity (TFA) [1]

(Figure 2-1). Thus, TFA is an essential component of gene regulatory networks. It regulates gene expression in response to internal and external signals to ensure appropriate gene expression.

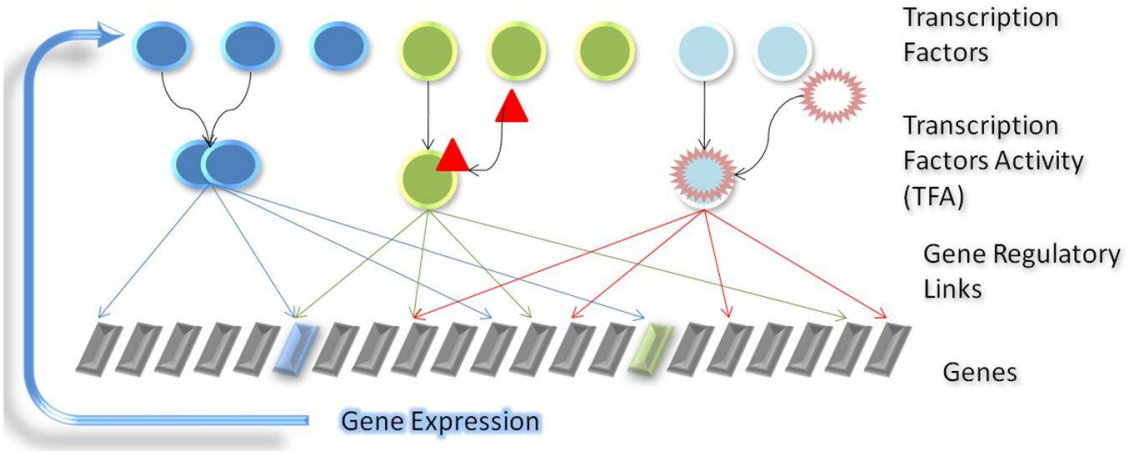


Figure 2-1 *Gene regulatory network model*

In this gene regulatory network model, a layer of Activated Transcription Factors added between the Transcription Factors layer and Gene layer. Only activated transcription factors can regulate the expression of genes through the gene regulatory links, inactivated transcription factors do not have regulatory links to the target genes. And the expression level of genes regulated by activated transcription factors changes by the effect of regulation, and the changed expression levels of genes affect the amount of the translated transcription factors.

Since TFA is governed by various complex molecular interactions, it is difficult to determine directly from experiments, especially if the activation mechanism is unknown. However, it is possible to computationally predict the change of TFAs relative to a reference state using transcriptome data and a known TF-gene network architecture [1,

2]. Network Component Analysis (NCA) developed by Liao et al. defines the problem of calculating TFAs as optimization of a linear least square matrix decomposition. Liao et al. solve the problem using an expectation maximization (EM) approach [3]. Fast Network Component Analysis (FastNCA) uses singular value decomposition (SVD) and a matrix projection technique to approximate the linear least square matrix decomposition problem defined in NCA[4]. Similarly, Alter and Golub use SVD and pseudo-inverse projection, and integrate ChIP and microarray data to calculate the hidden TFA layer between TFs and genes [5]. ChIP data provides additional information on proteins' DNA binding occupancy. Gao et al. developed an algorithm that combines microarray data for mRNA expression and transcription factor occupancy to define the regulatory network (MA-Networker algorithm) to predict TFAs based on ChIP and transcriptome data using multivariate regression and backward variable selection [6]. With the predicted TFAs, Gao et al. calculate the TF-gene coupling factor using Pearson Correlation [6]. Boulesteix et al. applied statistically inspired modification of the partial least square (SIMPLS) algorithm to find TFAs [7]. Many more complex models are also applied to predict TFAs. For example, Nachman et al. apply the Bayesian Network approach to provide a probabilistic model to predict TFAs [8]. The State-space model by Li et al. assumes the TFAs are affected by the TF gene expressions of previous time points [9]. Probabilistic dynamical models by Sanguinetti et al. consider the possibility of the same TF having different activities on different target genes [10]. A Gaussian process model developed by Gao et al. uses the Bayesian marginalization approach to predict TFAs [11]. Besides predicting TFAs from gene expression data and TF network

structures from experiments and literature data, DNA sequence motif information is also widely used (e.g. searching for DNA binding site of TFs) in many methods to infer potential TF-gene links to obtain a more complete TF network structure and improve the prediction of TFAs [2]. However, compared to matrix decomposition and regression approaches, these complex models require more computational power. Thus, these complex models either cannot deal with large scale TFAs or they predict large scale TFAs by converting TFAs into binary.

High-throughput technologies have led to many algorithms for the reconstruction of large scale gene regulatory networks [12]. For example, many sequence analysis approaches which identify potential TF binding sites have been developed [13]. However, many of the predicted potential TF binding sites are not functional (false positive predictions) [12]. From ChIP-chip technology, potential gene regulatory effects can be derived by identifying the portions of a genome that are bound by a particular TF *in vivo* [14]. Transcriptome data (also known as gene expression data) measured by genome-wide DNA microarrays are widely used for gene regulatory network reconstructions. For instance, Stuart et al. use correlation coefficients between mRNA levels of genes as relevance scores to reconstruct correlation networks [15]. The interacting genes are predicted by detecting the correlation score above some set threshold. Other algorithms such as RELNET (RELevance NETworks) [16] and ARACNE (Algorithm for the Reverse engineering of Accurate Cellular NEtworks) [17] use mutual information as the relevance scores. The CLR (Context Likelihood Relatedness) [10] algorithm uses an adaptive background correction method on the

relevance scores to improve precisions [18]. CLR significantly improved the performance of gene regulatory network reconstruction, and is widely adopted in the latest developed gene regulatory network reconstruction algorithms. In the field well known conference on Dialogue for Reverse Engineering Assessments and Methods (DREAM) [19], many winning algorithms are based on CLR. For examples, the best performer algorithm in DREAM2 Challenge 5, synergy augmented CLR (SA-CLR), introduced three way mutual information instead of the pair-wise mutual information in the original CLR [20]. Madar et al. developed a ordinary differential equation (ODE) based dynamic model extension of CLR (mixed-CLR/tl(time-lagged) CLR integrated with Inferelator 1.0) to treat steady-state data and time-series data separately and had an outstanding performance on DREAM3 and DREAM4 100-gene *in silico* network challenge [21, 22]. Huynh-Thu et al. developed a regression and tree based algorithm to reconstruct gene regulatory networks and awarded the best performer in DREAM4 *in silico* Multifactorial challenge [23]. Pinna et al. developed a graph analysis based algorithm to predict directed gene regulatory network from gene knockout experiments [24].

Many gene regulatory network reconstruction algorithms focus only on time series transcriptome data to develop dynamic models [25]. These include network identification by multiple regression [26], microarray network identification [27] and multi-scale time-correlation estimation [28]. time-series network identification [29], directed information-based CLR [30]. Dynamic Bayesian network models use a Bayesian Framework to reconstruct gene regulatory networks [31, 32].

Time-series based algorithms and dynamic Bayesian networks models can provide realistic models to reconstruct gene regulatory networks. However, due to a lack of closely spaced time-series data and computational power, these algorithms are difficult to apply on a genome-wide scale. Relevance score based algorithms are more efficient computationally and can integrate many different types of transcriptome data.

The standard simplified two-layer (TF-gene) model assumes a gene regulatory network model in which expressed TFs affect their target genes directly, despite the fact that TFA plays an important role in gene regulation. This simplification may lead to large false positive detection rates. Recently, the problem that TF gene expression does not necessarily correlate with target gene expression was noted in [33]. This discrepancy was addressed using a knowledge base representation of a TF expression by averaging the expressions of its target genes [33]. In our GTRNetwork model, we introduce a hidden layer of TFAs into relevance score approaches which connects TFs and their target genes. The three layer model (**Figure 2-1**) is more realistic than the two-layer model, and more biologically reasonable than the knowledge base representation model. The GTRNetwork model results in an approach to reconstruct large scale genome-wide gene regulatory networks that is both biologically more meaningful and computationally feasible.

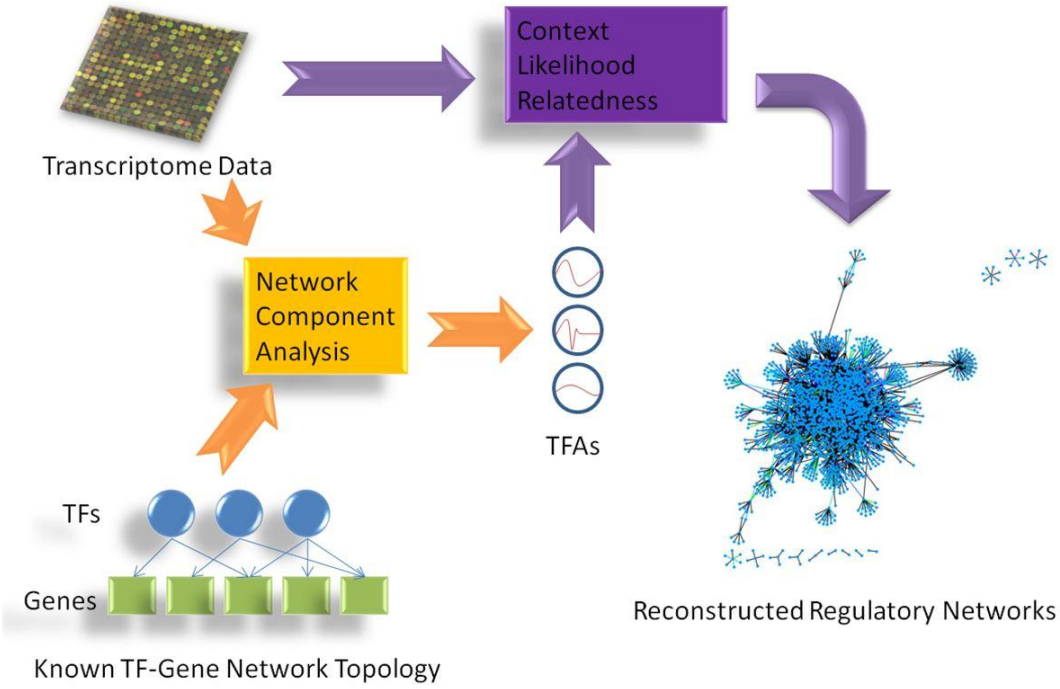


Figure 2-2 *Gene expression and Transcription factor activity based gene Regulatory Network (GTRNetwork) framework*

GTRNetwork algorithm has two steps. Step 1 (Yellow) take input of transctriptome data, predict transcription factor activities (TFAs) of TFs from known TF-Gene Network Topology. Step 2 (Purple) take the input of trancriptome data and introduce the predicted TFAs from step 1 to reconstruct gene regulatory network use score based network reconstruction methods.

The proposed **G**ene expression and **T**ranscription factor activity based **R**elevance **N**etwork (GTRNetwork) is a novel gene regulatory network reconstruction algorithm. It introduces a hidden layer of TFAs into relevance score based network reconstruction algorithms **(Figure 2-2)**. The GTRNetwork combines relevance score based algorithms

and TFA prediction algorithms, and generally follows two major steps. In Step 1, TFA ratios are predicted from transcriptome data and a specified TF-gene network topology. Transcript abundance ratios can be obtained from cDNA microarray or short read sequencing technology data. TF-gene network topologies can be assembled from online databases, such as RegulonDB [34]. However, TFA prediction algorithms are only based on the known TF-gene network topology and not able to predict new regulatory links. In Step 2 of GTRNetwork, gene regulatory networks are reconstructed from the gene expression ratio data and the predicted TFAs. Instead of using gene expression level as the only input to detect relationships between TFs and genes, GTRNetwork uses the relevancies between TFs and genes estimated based on the TFA and gene expression ratios. A check operon step can be used to improve the sensitivity of regulatory link detection. When gene operon information is available, it can be integrated after obtaining the reconstructed gene regulatory networks. By using gene operon information, when one gene in the operon is detected as a TF target, other genes in the same operon are automatically linked to the same TF.

Results

Selection of TFA prediction algorithms and network reconstruction algorithms

Different TFA prediction algorithms and network reconstruction algorithms affect the performance of the GTRNetwork method. In this research, the task is to reconstruct gene regulatory networks of *E. coli* in the whole genome scale, which includes over 4000 genes and 160 TFs. In TFA prediction algorithms, only the algorithms using matrix

decomposition and regression approaches could fit the computational requirements and scale needs of GTRNetwork algorithm for a whole genome. Three major approaches to predict TFAs are: gNCA-r which uses expectation maximization (EM) [3], FastNCA which uses singular value decomposition (SVD) [4], and SIMPLS which uses partial least square (PLS) regression [7].

Similar scale and computational power requirements as the TFA prediction algorithms exist in regulatory network reconstruction algorithms using TFAs and gene expression levels. The relevance scores are calculated by either Pearson correlation coefficients or adaptive partitioning mutual information (APMI) [35]. While using relevance scores approach on microarray experiments, different genes may have different background noise in different patterns and scales. For example, relevance scores may fail to distinguish direct interaction from indirect influences when the experimental conditions are unevenly sampled, or when the microarray normalization fails to remove false background correlations [18]. Research by Faith *et al.* [18] showed that using a background correction in the relevance score based network reconstruction process reduces the false positive detection rate of regulatory links and significantly improves the performance of the network reconstruction. The Context Likelihood Relatedness (CLR) [18] algorithm provides background correction on relevance scores in GTRNetwork.

GTRNetwork Algorithm Testing

GTRNetwork Algorithm Variant	TFA prediction	Relevance score	CLR Background correction
E-A-C	EM	APMI	Yes
E-A-N	EM	APMI	No
E-C-C	EM	Cor	Yes
E-C-N	EM	Cor	No
P-A-C	PLS	APMI	Yes
P-A-N	PLS	APMI	No
P-C-C	PLS	Cor	Yes
P-C-N	PLS	Cor	No
S-A-C	SVD	APMI	Yes
S-A-N	SVD	APMI	No
S-C-C	SVD	Cor	Yes
S-C-N	SVD	Cor	No
N-A-C	None	APMI	Yes
N-A-N	None	APMI	No
N-C-C	None	Cor	Yes
N-C-N	None	Cor	No

Table 2-1 *GTRNetwork Algorithm Combinations*
GTRNetwork algorithms using different combination of TFA prediction algorithm and relevance score based network inference algorithm

The performance of the GTRNetwork algorithm using different combinations of TFA prediction algorithm and relevance score based network inference algorithms have been tested. Three TFA prediction algorithms (EM-based gNCA-r, SVD-based FastNCA, and regression-based SIMPLS) and two relevance score functions (Pearson correlation coefficient and adaptive partitioning mutual information) have been tested with or without using CLR background correction. The GTRNetwork algorithm using the expression level of TFs as TFAs was also tested to demonstrate its performance without

including the TFA layer. Detailed information on the tested algorithms can be found in **Table 2-1**.

To test the performance of the GTRNetwork algorithm using TF-gene network topologies providing different levels of information as inputs, the training datasets of input initial TF-gene network topologies are obtained by randomly knocking out 70%, 50%, 30% or 10% of links from the TF-gene regulatory links dataset of RegulonDB 7.0 [34]. The testing datasets of TF-gene networks are the links that have been removed from the training datasets respectively. Thus, the ability of the algorithm to predict the removed regulatory links is tested. The transcriptome data input for testing the GTRNetwork algorithm is an *E. coli* gene expression data set integrating 466 transcriptome experimental conditions on 4279 gene probes from the M3D database [36]. The operon information was downloaded from the RegulonDB 7.0 database [34] and used in the check operon step to find more regulatory links. GTRNetwork algorithms were applied to the input training datasets to reconstruct gene regulatory networks with different network sizes. The results are compared with the testing datasets described above and the precision and recall (sensitivity) values are calculated for each network:

$$Precision = \frac{Number\ of\ testing\ \ dataset\ verified\ new\ links}{Total\ predicted\ new\ links} \quad (1)$$

$$Recall = \frac{Number\ \ of\ \ testing\ \ dataset\ \ verified\ \ new\ links}{Total\ \ number\ \ of\ \ \ links\ in\ the\ testing\ \ dataset} \quad (2)$$

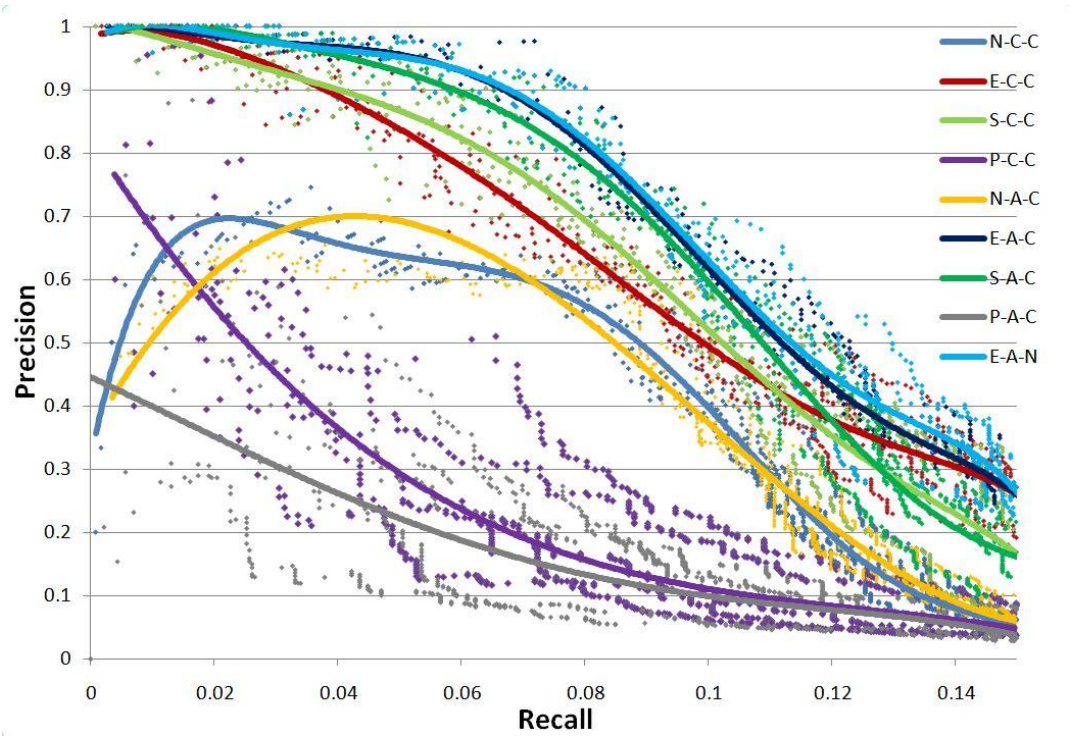


Figure 2-3 *GTRNetwork algorithm combinations on input initial network of 30% regulonDB 7.0 data*

70% of links randomly deleted. Five runs were made for each recall level. The trend lines of data points are fitted by polynomial functions. Under this condition the combination E-A-C (EM-based TFA prediction, APMI relevance score with CLR background correction) and E-A-N (EM-based TFA prediction, APMI relevance score without CLR background correction) give the best performances. All the TFA based algorithms except the SIMPLS based TFA prediction show significantly better performance than the algorithms not using TFA information..

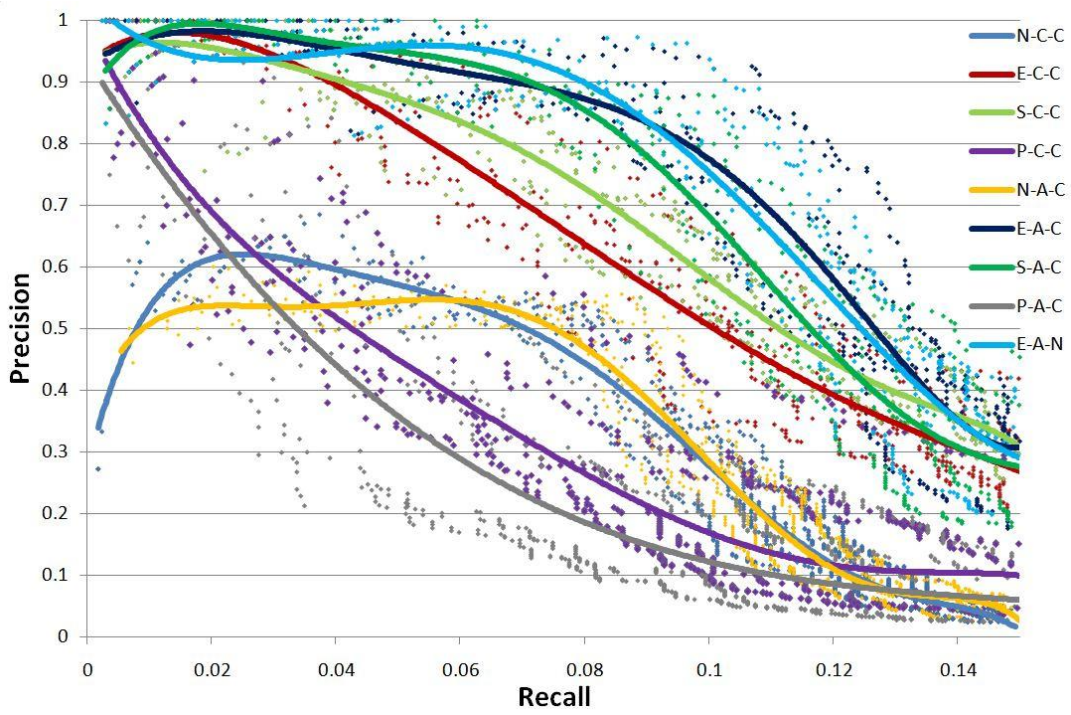


Figure 2-4 *GTRNetwork algorithm combinations on input initial network of 50% regulonDB 7.0 data*

50% of links randomly deleted. Five runs were made for each recall level. The trend lines of data points are fitted by polynomial functions. Under this condition the combination E-A-C (EM-based TFA prediction, APMI relevance score with CLR background correction) and E-A-N (EM-based TFA prediction, APMI relevance score without CLR background correction) give the best performances. All the TFA based algorithms except the SIMPLS based TFA prediction show significantly better performance than the algorithms not using TFA information. At the low recall levels, the regression based TFA prediction algorithms (P-C-C and P-A-C) have better performance than the algorithms not using TFA information while.

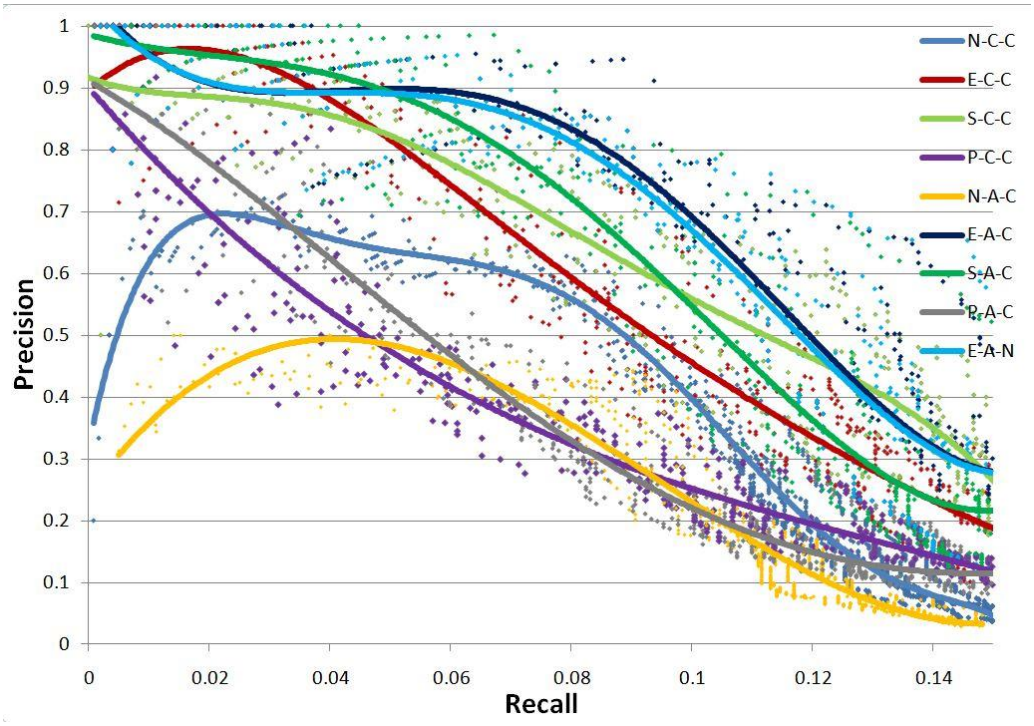


Figure 2-5 *GTRNetwork algorithm combinations on input initial network of 70% regulonDB 7.0 data*

30% of links randomly deleted. Five runs were made for each recall level. The trend lines of data points are fitted by polynomial functions. Under this condition the combination E-A-C (EM-based TFA prediction, APMI relevance score with CLR background correction) and E-A-N (EM-based TFA prediction, APMI relevance score without CLR background correction) give the best performances. All the TFA based algorithms except the SIMPLS based TFA prediction show significantly better performance than the algorithms not using TFA information. At the low recall levels, the regression based TFA prediction algorithms (P-C-C and P-A-C) have better performance than the algorithms not using TFA information.

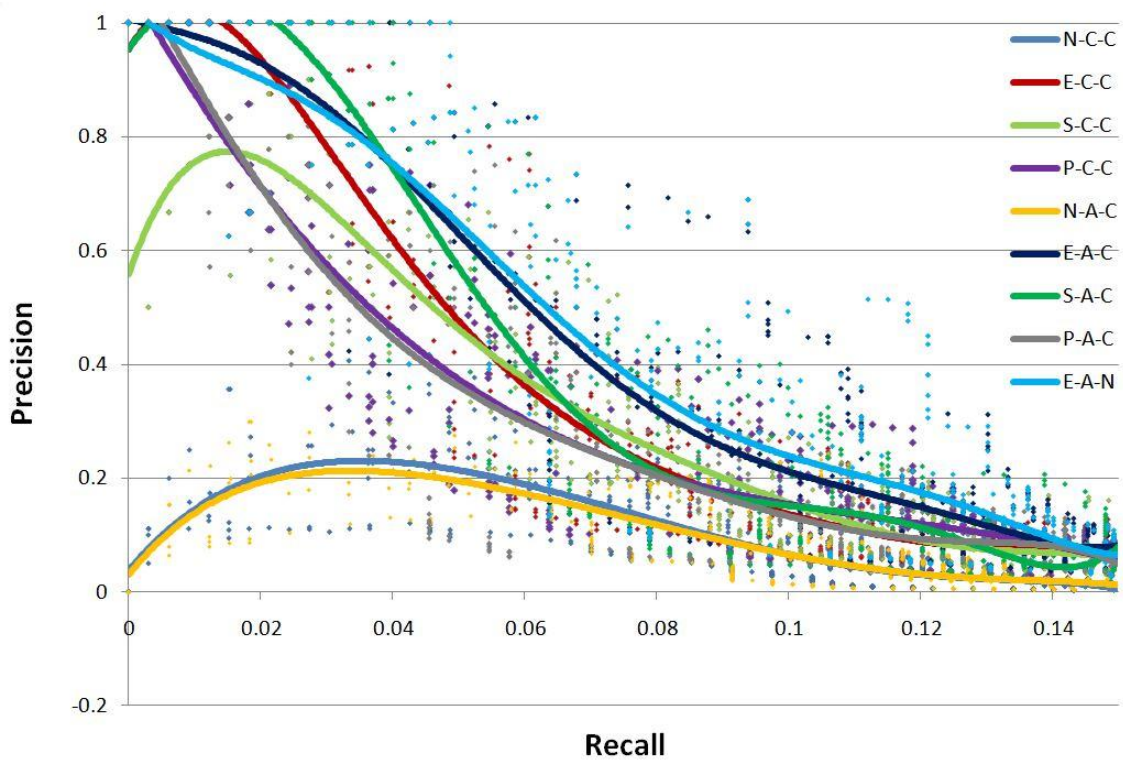


Figure 2-6 *GTRNetwork algorithm combinations on input initial network of 90% regulonDB 7.0 data*

10% of links randomly deleted. Five runs were made for each recall level. The trend lines of data points are fitted by polynomial functions. Under this condition the combination E-A-C (EM-based TFA prediction, APMI relevance score with CLR background correction) and E-A-N (EM-based TFA prediction, APMI relevance score without CLR background correction) give the best performances. All the TFA based algorithms show significantly better performance than the algorithms not using TFA information.

On each percentage level of input training dataset, the test is repeated five times to estimate the stability of GTRNetwork algorithms. In the Precision-Recall plots, all

algorithm combinations show the same trend: as recall value increases, precision decreases. **(Figure 2-3 - 2-6)**. At the same recall level, higher precision suggests better performance of the algorithm; while at the same precision, the larger recall value shows better performance of the algorithm. And the area under precision-recall curve (AUPRC) for each test are calculated **(Figure 2-7)**. The larger AUPRC value tells us the better performance. The test results for all combinations of the GTRNetwork algorithm are shown in **APPENDIX A**.

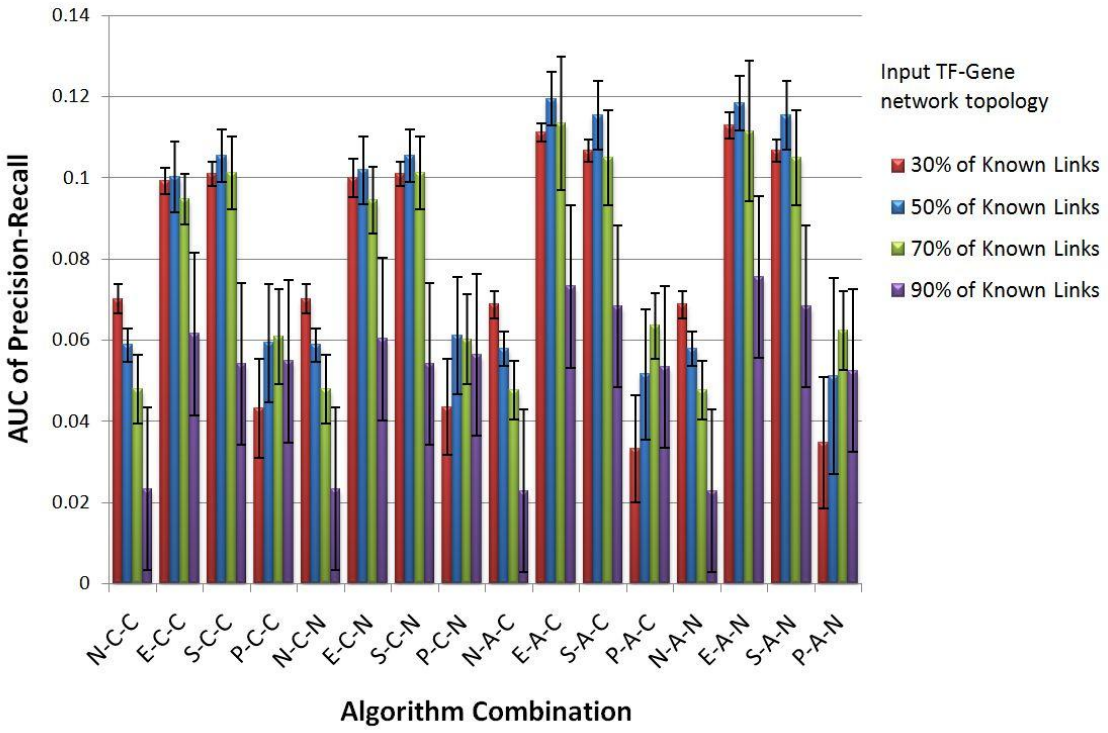


Figure 2-7 *Area under curve of precision-recall (AUCPR) of GTRNetwork algorithm combinations with different input TF-gene network topologies*

The performance of GTRNetwork is relatively consistent while using input TF-gene network topologies containing different percentages of known regulatory links, except

using the 90% of known regulatory links as the input TF-gene network topology. EM-based or SVD-based TFA prediction algorithms (E/S-C-C, E/S-C-N, E/S-A-C, E/S-A-N) give significantly better performance than algorithms without using TFA information (N-X-X) or algorithms using PLS based TFA prediction (P-X-X). The algorithms using APMI relevance score function (the right half of the plot) show slightly better performance than the algorithms using Pearson correlation relevance score function (the left half). And there are no significant differences due to the use of the CLR background correction (X-X-C or X-X-N).

There are four factors which affect the performance of GTRNetwork: the TFA prediction algorithm, the relevance score function, the background correction effect, and the network sizes of initial TF-gene network topology. **Figure 2-7** shows that using predicted TFA information from EM or SVD-based method significantly improved the performance of the gene regulatory network reconstruction. (Two sample t-test p-value < 0.0001). The APMI relevance score function gives slightly better performance than the correlation relevance score function. (Paired two sample t-test p-value < 0.0001). However, there is no clear difference between using or not using the background correction of CLR. (Paired two sample t-test p-value = 0.8342). The performance of most algorithm combinations is relatively consistent while using different level of known knowledge of the initial TF-gene network topologies. However, when using the 90% of known TF-gene links as the initial network topology, the performances drops significantly. This performance drop is expected because as the training data (the portion of known TF-gene links) increases, the testing data is reduced. Many predicted links are

already known, and only few links can be identified as new predicted links. Also many new predicted links might not be included in the testing dataset thus not being verified as a true positive prediction. However, the unverified prediction could still be true since the testing dataset is not a complete dataset; our knowledge of the complete biology of this system is still incomplete. When the portion of the known TF-gene links is increased in the training data, the total number of predicted new links decreases. At the same time, the number of unknown regulatory links in prediction would not change, or even increase because of more complete training information. Thus, the portion of unknown regulatory links in the prediction is increased. In this case, the testing is closer to a prediction. The verification based only on known knowledge cannot reflect the real performance of identifying potential new gene regulatory targets (**Figure 2-8**).

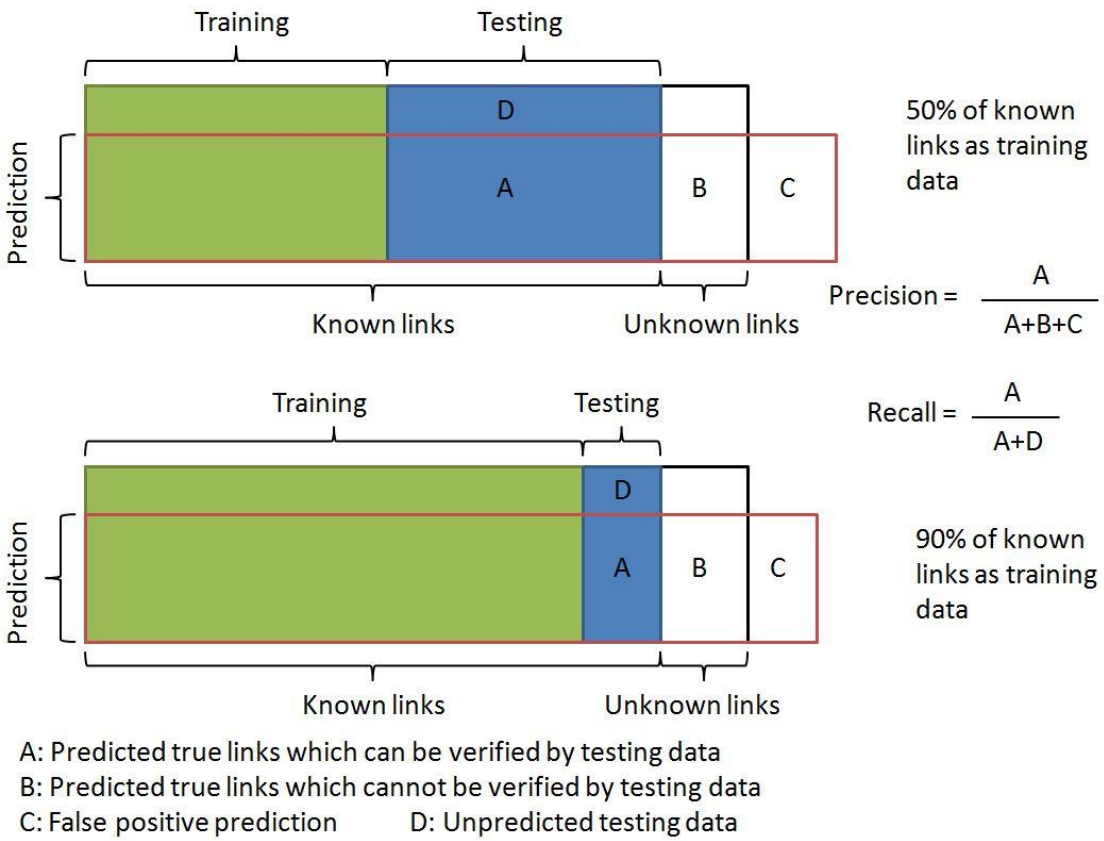


Figure 2-8 *Demonstration of TF-Gene regulatory links data*

The prediction (the area in red line) includes a part of the training data, a part of the testing data, a part of currently unknown links and some false positive predictions. When the percentage of known links as training data increases, since more training data is used, at the same recall level, the false positive decreases, and the precision (portion of area A in the area A+B+C) decreases.

In conclusion, the algorithms using EM-based or SVD-based TFA prediction methods along with the APMI relevance score gave the best performance. In general, using or not using CLR background correction does not give significant differences in performance, but since CLR has low computational requirements (See the discussion session) and has

been shown helpful in gene regulatory reconstruction algorithms [18], we suggest the use of CLR background correction in the GTRNetwork algorithm. Thus, the E-A-C (EM-based TFA prediction, APMI relevance score function and using the CLR background correction) combination is used as the default GTRNetwork algorithm in the testing and application below.

A comparison between the original CLR [18] and GTRNetwork algorithm is also applied on the M3D *E. coli* data **(Figure 2-9)**. Comparisons between CLR algorithm and many other gene regulatory network reconstruction algorithms have been done in the CLR paper [18]. And many DREAM winning algorithms, e.g. SACLR [20] and GENEI3 [23], have compared themselves with CLR on the M3D *E. coli* data and found comparable performance with CLR [20, 23]. GTRNetwork outperforms CLR significantly when we use the full TF-gene regulatory information from RegulonDB 7.0 as the initial TF-gene network topology (**Figure 2-9A**). However, the result is predictable since GTRNetwork uses the additional information of TF-gene links as input, and all other algorithms only use the list of TFs as input. While using a 50% randomly knocked out TF-gene regulatory links from RegulonDB 7.0 as the training initial TF-gene network topology, and the removed regulatory links in the training dataset as the testing data, this situation would be more relative to a real biological application. In most biological cases, only limited TF-gene regulatory information is known, and the task of gene regulatory network reconstruction algorithms is to identify new regulatory links. The result still shows stronger performance of the GTRNetwork algorithm on the task of identifying new regulatory networks based on known knowledge of gene regulatory networks

(**Figure 2-9B**).

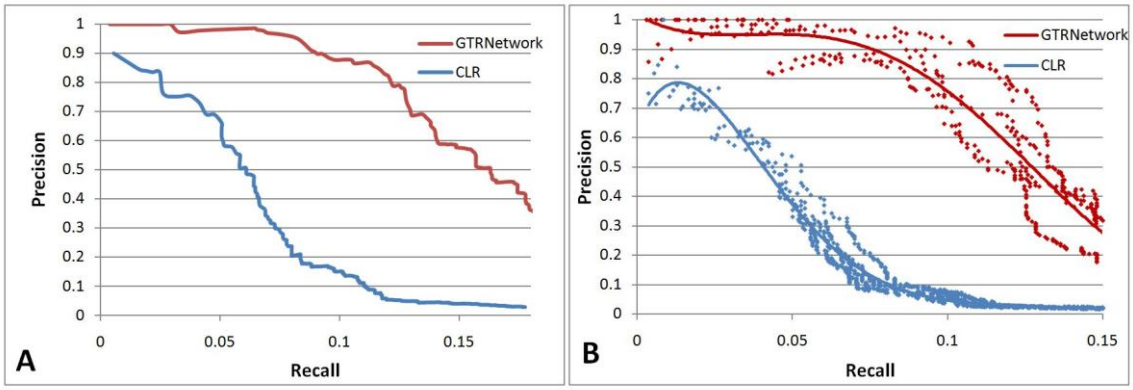


Figure 2-9 *Comparison between GTRNetwork and CLR on E. coli data*

(A) Precision-recall curve of testing results of GTRNetwork and CLR algorithms using transcriptome data from M3D database [36] and the input training TF-Gene topology of the full set of RegulonDB 7.0 [34]. (B) Precision-recall plot of testing results of GTRNetwork and CLR algorithms using transcriptome data from M3D database [36] and the input training TF-Gene topology of 50% links randomly knocked out RegulonDB 7.0 [34] data. Five random replications are applied on the test. The precision and recall are calculated based on the testing data of the knocked out RegulonDB 7.0 [34] on each replication respectively. The trend lines are fitted by polynomial functions.

Application of GTRNetwork Algorithm

According to the test results above, the E-A-C algorithm combination best fits the current known gene regulatory network topology from RegulonDB 7.0. This algorithm combination was applied using the full set of RegulonDB 7.0 TF-gene links as the initial network topology. The gene expression data of *E. coli* integrating 466 transcriptome

experiment conditions on 4279 gene probes from the M3D database was used as the transcriptome data input. Resulting gene regulatory networks with sizes ranging from 100 links to 600 links were reconstructed. Different relevance score thresholds were set to reconstruct gene regulatory networks with different sizes. Higher thresholds result in smaller regulatory networks with fewer false positives. Lower thresholds give more complete networks, but with more false positives. A check operon step using operon information from RegulonDB 7.0 was applied to improve the sensitivity of the reconstructed regulatory networks. The complete detailed predicted results are shown in **APPENDIX B**.

TF	Gene	Supporting Evidence
DicA	*insD*	TF binding site verified (RegulonDB) [34]
DicA	*intQ*	TF binding site verified (RegulonDB) [34]
DicA	*ydfE*	TF binding site verified (RegulonDB) [34]
DcuR	*pepE*	Involve in anaerobic respiration related process (EcoCyc [37])
Fur	*ybdB*	ybdB (entH) is proposed to be regulated by Fur (EcoCyc [37])
Fur	*yncE*	YncE is de-repressed by Fur [41]
IscR	*fdx*	Some evidence that the fdx functions as an intermediate site for Fe-S cluster assembly [42]
IscR	*hscA*	HscA is required for the assembly of iron-sulfur clusters [43, 44]
IscR	*hscB*	HscB is a co-chaperone that stimulates HscA (Hsc66) ATPase activity [44]
IscR	*iscX*	Both involve in Iron-sulfur cluster process [43]
SgrR	*sroA*	TF binding site verified (RegulonDB)[34]

Table 2-2 *Valid search of 12 predicted new links using literature*
New regulatory links of *E. coli* predicted in a reconstructed gene regulatory network of size of 200 links (includes 16 potential new regulatory links).

In the reconstructed 100-link regulatory network, there are three new predicted regulatory links: DicA-*insD*, DicA-*intQ*, DidA-*ydfE*. These new links are biologically verifiable since *insD, intQ* and *ydfE* are in the same operon with a TF binding site of regulator DicA, according to the binding-site information obtained from RegulonDB 7.0

[34]. In the reconstructed 200-link regulatory network, besides the three new links predicted in the 100-links network, another 13 new regulatory links were predicted **(Table 2-2)**. Evidence of biological validity of 8 of these 12 new links can be found in the literature or in databases such as EcoCyc [37]. For example, IscR is an iron-sulfur cluster regulator [38] and *fdx, hscA, hscB* and *iscX* are all involved in the iron-sulfur cluster assembly process.

Gene	Gene function
efeU	ferrous iron permease component of the EfeUOB ferrous iron transporter.
ybdB(entH)	EntH is a thioesterase that is involved in the biosynthesis of enterobactin
bfd	Bacterioferritin-associated ferredoxin; predicted redox component complexing with Bfr in iron storage and mobility [2Fe-2S]
bfr	The bfr gene encodes bacterioferritin, which is an iron storage protein
efeB	Deferrrochelatase, periplasmic; inactive acid inducible low-pH ferrous ion transporter EfeUOB; periplasmic acid peroxidase; heme cofactor
efeO	Inactive acid-inducible low-pH ferrous ion transporter EfeUOB; acid-inducible periplasmic protein
ybaN	Inner membrane protein, DUF454 family, function unknown
ydiE	Function unknown, hemin uptake protein HemP homolog
yncE	Secreted protein, function unknown,suggesting a role in iron acquisition
yqjH	YqjH is an NADPH-dependent ferric reductase containing FAD covalently bound to a cysteine sidechain via a thioether bond.

Table 2-3 *Predicted Fur target genes*

Genes predicted as the targets of regulator Fur in the reconstructed regulatory network of size of 600 links. Gene function information is downloaded from EcoGene database [45]

The 600-link reconstructed gene regulatory network contains 381 new predicted gene regulatory links, including links predicted by checking operon information. These 381 predicted links appear biologically meaningful. For instance, the ferric uptake regulator, Fur, is predicted to have links with many ferrous iron transporters and storage related genes (*efeU, bfd, bfr, efeB, efeO, ybdB (entH), ydiE, yqjH*). Many of these new predicted

targets have unknown biological function, such as inner membrane protein gene, *ybaN*, and secreted protein gene, *yncE*. The fact that these genes may be part of the Fur regulon suggests that their function may be related to iron uptake **(Table 2-3)**.

Despite the fact that *E. coli* is so well-characterized, there are still many genes that have no known regulators. The GTRNetwork predictions help discover the regulators of those genes still have no known regulators. In the 381 predicted links, there are 171 predicted target genes which previously had no known regulators (**APPENDIX B**).

The reconstructed gene regulatory networks with potential new gene regulatory links can be used again in the application of predicting TFAs and identify more significantly changed TFAs in response to the experiment condition changes. For example, Brynildsen *et al.* used the gene regulatory network obtained from RegulonDB and NCA to predict TFAs of *E. coli* under isobutanol stress from transcriptome data and identified 16 significantly changed TFAs in response to the isobutanol condition [39]. We reanalyzed their transcriptome data using our reconstructed gene regulatory network, including the 381 predicted new links. This additional of the new regulatory links resulted in another 7 significantly changed TFAs in response to the isobutanol condition (**Table 2-4**).

TF	Function	Target Genes
ArgR	Arginine catabolism	***argA***, ***gltF***, *argE*, ***argH***, *rimP*, ***rbfA***, *truB*, ***rpsO***, ***pnp***, ***nusA***, ***infB***, ***hisP***, ***gltD***, ***gltB***, ***carB***, *artP*, *artI*, ***artQ***, *artM*, ***artJ***, ***hisJ***, ***hisQ***, *metY*, *astE*, *astB*, *astD*, *astA*, *astC*, *hisM*, *argB*, *argC*, *argD*, *argF*, ***argG***, ***argI***, *argR*, ***carA***
AscG	Arbutin-salicin-cellibiose transport and utilization	***ascB***, *ascF*, ***ascG***, ***htpG***, *prpR*, ***clpB***, ***dnaJ***, ***dnaK***, *tpke11*, *groL*, *groS*, ***grpE***, ***hslU***, ***hslV***, ***ybbN***, ***lipB***, ***ybeD***, ***lnt***, ***ybeX***, *ybeY*, ***ybeZ***
CysB	Novobiocin resistance, sulfur utilization, and sulfonate-sulfur catabolism	*tauA*, *tauB*, *tauC*, *ssuC*, *ssuD*, *ssuA*, *ssuE*, ***hslJ***, ***cbl***, *tauD*, *ssuB*, *cysP*, *cysU*, ***cysW***, ***cysN***, ***cysM***, ***cysK***, ***cysJ***, ***cysI***, ***cysH***, ***cysD***, *cysC*, ***cysB***, ***cysA***, ***gsiA***, ***gsiB***, ***gsiC***, ***gsiD***, ***iaaA***, *yciW*, ***ydjN***, ***yeeD***, ***yeeE***
Lrp	Leucine-responsive regulatory protein	*lhgO*, *alaT*, *alaU*, *alaV*, *gltT*, *gltU*, *gltV*, *gltW*, *ileT*, *ileU*, *ileV*, ***micF***, *rrfA*, *rrfB*, *rrfC*, *rrfD*, *rrfE*, *rrfG*, *rrfH*, *rrlA*, *rrlB*, *rrlC*, *rrlD*, *rrlE*, *rrlG*, *rrlH*, *rrsA*, *rrsB*, *rrsC*, *rrsD*, *rrsE*, *rrsG*, *rrsH*, *rrfF*, *thrV*, *csiD*, *ilvX*, *adhE*, ***aroA***, *fimA*, *fimC*, *fimD*, *fimE*, *fimF*, *fimG*, *fimH*, *gabT*, ***gcvH***, ***gltB***, ***gltD***, *ilvA*, ***ilvD***, *ilvE*, *ilvH*, *ilvH*, *ilvI*, ***ilvM***, *kbl*, ***livF***, ***livG***, ***livH***, *livJ*, ***livK***, ***livM***, *lrp*, *lysU*, ***malT***, *ompC*, ***ompF***, ***oppA***, ***oppB***, ***oppC***, ***oppD***, ***oppF***, ***osmC***, ***sdaA***, *serA*, ***serC***, ***tdh***, *argO*, *ilvL*, *gabD*, *gabP*, ***osmY***, ***hdeA***, ***hdeB***, *yhiD*, *dadA*, ***dadX***, ***gcvT***, ***gltF***, ***stpA***, *gcvP*, ***aidB***, *fimI*, *yeiL*, *yojI*, ***gdhA***, ***ilvG_1***, ***ilvG_2***, ***thrA***, ***thrB***, ***thrC***, *thrL*
MarA	Multiple antibiotic resistance	*pqiB*, *pqiA*, ***ybaO***, *nfsB*, ***micF***, ***slp***, *dctR*, ***acrB***, *acrA*, *marB*, *marR*, ***marA***, ***inaA***, *rfaY*, *rfaZ*, *yhiD*, ***hdeB***, ***hdeA***, ***rob***, ***zwf***, ***fumC***, *fpr*, *nfo*, ***poxB***, *purA*, *putA*, ***sodA***, *tolC*, ***ygiA***, *ygiB*, ***ygiC***, *ltaE*, *ybjT*, ***talA***, ***tktB***, *phr*, ***ybgA***, *yhbW*
MetJ	Methionine biosynthesis and transport	*metF*, ***metK***, *metL*, ***metR***, *yeiB*, ***folE***, *ahpC*, *ahpF*, *metQ*, *metN*, *metI*, ***metA***, *metB*, ***metC***, *metE*
NadR	NAD biosynthesis	***nadA***, *pnuC*, ***pncB***, *nadB*

Table 2-4 *Significantly changed TFAs under isobutanol condition predicted by GTRNetwork reconstructed gene regulatory network*

Table 2-4 *continued*

The reconstructed gene regulatory network includes 381 potential new regulatory links, the 16 significantly changed TFAs predicted by original RegulonDB data from Brynildsen's paper [39] are not included. Bolded genes are expression significantly changed genes according to Brynildson's paper [39]. Underlined genes are predicted new regulatory target genes of the TF from GTRNetwork

Discussion

Algorithm	PLS	EM	SVD	APMI	Correlation
Run time (seconds)	2750	1750	6.2107	1740	1.4086

Table 2-5 *Algorithm run time tests*

The run time of the three TFA prediction algorithms PLS, EM, SVD, the two relevance score functions, APMI and Correlation are tested.
Input: Gene expression data: M3D *E. Coli.* microarray experiments. 466 experiment conditions and 4279 gene probes. TF-gene network topology: RegulonDB 6.7 gene regulatory network. 3989 regulatory links.
Machine: CPU Intel(R) Core(TM) i7 950 @3.07GHz. RAM: 6.00 GB. OS: Windows 7 Professional 64-bit.

Besides precision and recall of predictions, other properties such as the run times of algorithms are important. Among the TFA prediction algorithms, the SVD-based, FastNCA algorithm is the fastest one. FastNCA (SVD) is 280 to 440 times faster than SIMPLS (PLS) and gNCA-r (EM) (**Table 2-5).** APMI takes about 1740 seconds to generate the relevance score matrix, while using correlation as the relevance score gets the score matrix over 1000 times faster (**Table 2-5**). Applying CLR background correction finishes in seconds but can improve the precision of the reconstructed network [18]. Thus, the most time efficient algorithm combination of GTRNetwork is the SVD-Correlation-CLR background correction (S-C-C) combination. Although under some conditions, S-C-C does not perform as well as other combinations, it provides a quick estimation with relatively reliable results. This algorithm combination could be

used to quickly generate a general view of the network.

The algorithm combinations that use regression-based SIMPLS to predict TFAs are not as precise as the other combinations. However, SIMPLS does not have as many restrictions as NCA algorithms have, such as the non-redundancy and full column and row rank of the initial network topology. Thus, SIMPLS does not discard as much information while preprocessing data to fit the input criteria. Studies show that it can predict regulatory links that gNCA-r and FastNCA could not [7]. This property of SIMPLS is especially important when there are some regulators or genes of interest, but other TFA prediction algorithms delete these interesting regulators or genes to fit the NCA criteria (detail in Methods session). There is no optimal combination of algorithms for GTRNetwork; instead, the user needs to choose the appropriate algorithm combination based on their input data and other requirements.

The TFA prediction model does not need any biological knowledge on the detailed mechanisms of the activation of TFs. The model assumes that all of the complex effects that contribute to the change of TFA are included in the predicted TFAs and the control strengths. Thus, the GTRNetwork algorithm is not limited to prokaryotes, but can also be applied to eukaryotes. We plan to apply this method to eukaryotes such as yeast and plants in the near future.

While most relevance score based gene regulatory network reconstruction algorithms are not able to identify the self regulation of TFs, because the gene expression data is directly used as the only input to represent both the regulators and the targets, there are always high relevance scores to connect the TF and its gene. In GTRNetwork, since the

representation of the regulators (TFAs) and the representation of the targets (expression of genes, including TF genes) are well separated, the relevance score between the TF and its gene is meaningful, and the self regulation of TFs can also be identified. The prediction of self regulation of TFs improves interpretation of the cyclic structures of gene regulatory networks. Further analysis of the effect of feedforward and feedback loops is not carried out in this work but will be applied on the reconstructed networks in our future work.

TFA prediction methods are all based on a linear static model of experimental conditions, and treat dynamic time series data as static data of each time point. Thus, although time series transcriptome data can be used as an input of GTRNetwork, the algorithm does not take advantage of dependencies in time series data.

Conclusion

The algorithm GTRNetwork introduces the hidden layer TFA into classic gene regulatory network reconstruction networks. A comparison of the performances of several algorithmic variants of this algorithm showed that the E-A-C variant of the GTRNetwork use EM-based TFA prediction method, adaptive partitioning mutual information as the relevance score function and CLR background correction method. This is the variant best fits the current known TF-gene regulatory networks from RegulonDB. The application on the E-A-C variant on *E. coli* data shows a promising amount of biological significance. It would be interesting and meaningful to verify more predicted result biologically and try other alternative TFA prediction such as the

SIMPLS based methods and network reconstruction algorithms computationally. The application on other organisms such as yeast is also highly recommended to be applied in the future research.

Methods

TFA prediction

TFA prediction is based on the following biological approximation [1]:

$$Er_i = \prod TFAr_j^{CS_{ij}} \quad (3)$$

Er_i is the gene expression ratio between two experiment conditions of the *i*-th gene , $TFAr_j$, $j = 1,\ldots,L$, is a set of TFA ratios of TF *j* , which regulate gene *i*, between the same two conditions, and CS_{ij} represents the control strength of transcription factor *j* on gene *i*. After taking the logarithm of Eq. (3) [1]:

$$log\left([Er]\right) = [CS]\, log\left([TFAr]\right) \quad (4)$$

where $N \times M$ matrix [*Er*] is the relative gene expression level matrix and $L \times M$ matrix [*TFAr*] is the relative transcription factor activities, the elements $Er_{ij}(t) = E_{ij}(t) / E_{ij}(0)$ and $TFAr_{kj}(t) / TFAr_{kj}(0)$, $N \times L$ matrix [*CS*] is the control strength matrix of transcription factors and genes. The gene expression model in Eq. (4) can be decomposed into matrix [*CS*] and matrix *log*([*TFAr*]) using different algorithms.

The relative gene expression level matrix [*Er*] can be obtained from transcriptome experiments such as DNA microarrays or RNAseq, and the control strength information

must be initialized from the literature e.g. RegulonDB [34], Chip-on-chip experiments, and motif information (mNCA [40]). The initial matrix, *CS* is converted from the known database of gene regulatory links between TFs and genes, e.g., RegulonDB [34]. Each row represents a gene and each column represents a TF. When there is a known regulatory link between gene *i* and TF j, CS_{ij}=1, otherwise CSij=0.

With the input of [*Er*] and [*CS*], transcription factor activities *log*([*TFAr*]) can be estimated. There are three major approaches to estimate *log* ([*TFAr*]) expectation maximization (EM) approach (e.g. gNCA-r) [3], singular value decomposition (SVD) approach (e.g. FastNCA) [4] and regression approach (e.g., SIMPLS) [7].

Note: When using gNCA-r or FastNCA to estimate *log* ([*TFAr*]) matrix, $\log([Er])$, $[CS]$ and $\log([TFAr])$ need to fit three criteria given below to ensure the uniqueness of the decomposition [1, 3, 4].

(i) The connectivity matrix [*CS*] must have full-column rank.

(ii) When a node in the regulatory layer is removed along with all of the output nodes Er_i connected to it, the resulting network must be characterized by a connectivity matrix that still has full-column rank. This condition implies that each column of [*CS*] must have at least *L*-1 zeros.

(iii) The matrix, log [*TFAr*], must have full row rank. In other words, each regulatory signal cannot be expressed as a linear combination of the other regulatory signals.

Relevance Scores

Instead of calculating relevance scores between the expression levels of two genes GTRNetwork calculates the relevance score between each TFA and each gene. Pearson correlation coefficient and mutual information are chosen as the relevance score functions:

Pearson Correlation Coefficient:

$$S_{ij} = \frac{\sum_k (X_{ik} - \bar{X}_i)(X_{jk} - \bar{X}_j)}{\sqrt{\sum_k (X_{ik} - \bar{X}_i)^2} \sqrt{\sum_k (X_{jk} - \bar{X}_j)^2}} \quad (5)$$

where X_{ik} is the k-th observation of variable i. and S_{ij} is the Pearson Correlation Coefficient score between variable i and j.

Mutual Information:

$$S_{ij} = \sum_i \sum_j p(i,j) log(\frac{p(i,j)}{p_1(i)\, p_2(j)}) \quad (6)$$

Where $p(i,j)$ is the joint probability of i and j, $p_1(i)$ and $p_2(j)$ are the marginal probabilities of i and j respectively, S_{ij} is the Mutual Information score between variable i and j.

The Pearson Correlation (Eq. 4) performs extremely well in detecting linear relationships between two variables (genes in a set of microarray experiments), and Mutual Information (MI) (Eq. 5) has a relatively balanced performance in detecting both linear and non-linear relationships. However, most MI applications only work for discrete variables, and in this problem, both the gene expression ratio and TFA ratio are continuous variables. Adaptive partitioning [35] adjustments are applied to calculate

mutual information between TFA ratios and gene expression ratios.

Background correction

In the relevance score based network reconstruction approaches; there are tradeoffs between the link detection sensitivities and false positive detection rates [10]. One reason for the false positive detection is the simplification of the two layer gene regulatory network model. Adding the TFA layer to the classic two layer regulatory network model may solve this problem. Another reason for the false positive detections is due to the noise of gene expression data and different relatedness behaviors of TFs and genes. For example, the expression of some genes may be more stable than other genes and not tend to change much in response of different conditions, the relevance score of these genes are tend to lower, and regulatory relationships between these genes and TFs are hard to be detected, the same to TFAs. Thus, a background correction method such as context likelihood relatedness (CLR) [18] is needed.

In the CLR algorithm, along with the relevance score, the statistical likelihood of each relevance score is calculated within each variable by:

$$z_{ij} = \frac{s_{ij} - \bar{s}_i}{\sqrt{\sum_j (s_{ij} - \bar{s}_i)^2}} \qquad (7)$$

where Z_{ij} is the z-score of relevance score between variable i and j within all relevance scores with i, S_{ij} is the relevance score between variable i and j, $\bar{S}i$ is the average of all relevance scores with i. And a joint likelihood between two variables is calculated from

the z-scores from Eq. (6). The methods to calculate the pseudo-z-score Z_{ij} vary and the CLR algorithm use the following method as default [18]:

$$Z_{ij} = \sqrt{z_{ij}^2 + z_{ji}^2} \qquad (8)$$

By putting different thresholds on the matrix [*Z*] with elements Z_{ij} gene regulatory networks with different sensitivities can be reconstructed by searching for gene regulatory links containing TF genes with the Z score larger than the threshold. The information of TF genes (which genes encode TFs) can be found from database such as RegulonDB [34] and EcoCyc [37].

Integration of operon information

In the reconstructed gene regulatory network, when gene *A* is predicted to be regulated by some TFs, the other genes in the same operon as gene *A* are not always predicted to be regulated by the same TFs regulating gene *A*. However, in real gene regulatory networks, all the genes in the same operon tend to have similar behavior. The GTRNetwork algorithm uses an optional check operon step. When the operon information is available, the algorithm searches for genes in the same operon as the target gene and links these genes to the regulators of the target gene. This integration of operon information improves the detection sensitivity of regulatory links.

GTRNetwork algorithm

The GTRNetwork algorithm is implemented using Matlab and the source code is available at:

http://vrac.iastate.edu/~afu/GTRNetwork/GTRNetwork_1.2.1.zip.

Input: a) Log 2 ratio transcriptome data in matrix [*Err*]

b) Initial TF-gene network topology in adjacency matrix [*C*]

c) Desired size of reconstructed regulatory network *S*

d) List of operons and the genes contained in them (Optional)

Output: A list of predicted regulatory links

The GTRNetwork algorithm uses the TFA prediction algorithm to predict TFAs from input a) and b). Then use relevance score functions such as correlation coefficient function or APMI to calculate the relevance score between TFAs of TFs and the expression levels of all genes. A CLR background correction is applied on the relevance score matrix. And then according to the desired size of reconstructed regulatory network (input c), a threshold based on the background corrected relevance score is calculated and the gene regulatory network is reconstructed filtered by the threshold. Finally, an optional check operon step is applied to add missing predicted regulatory links in the same operon of the predicted target genes.

1. Match the genes between the matrix [*Er*] and matrix [*C*]. Remove unmatched genes in [*Er*] and store the reduce matrix as [*Er0*]. Remove unmatched TFs and genes in [*C*] and store the reduced matrix in [*C0*].

2. If the TFA prediction algorithm is gNCA-r or FastNCA, check the three criteria described in TFA prediction section and reduce the matrix [*Er0*] and [*C0*] to fit the criteria.

3. Apply TFA prediction algorithm to predict the $\log_2$ ratio TFA matrix [*TFA*] from matrix [*Er0*] and [*C0*].

4. Calculate the relevance score matrix [*M*] between TFAs and all expression levels of all genes from matrix [*TFA*] and [*Er*].

5. Calculate the joint statistical likelihood matrix [*Z*] of relevance score matrix [*M*]using CLR algorithm.

6. Set a threshold T for matrix [*Z*] so that there are *S* elements in [*Z*] greater than *T*. For all the TF-gene pairs having a *Z* score greater than *T*, construct a regulatory link.

7. If the operon list is available, check and add all genes in the same operon of TF target genes to the regulatory target set of the TF.

Authors' contributions

YF developed and implemented the GTRNetwork algorithm, drafted this manuscript. LRJ and JD developed the initial concept and suggested ways to improve the algorithm and testing methods. All authors read and approved the final version of the manuscript.

Acknowledgment

This material is based upon work supported by the National Science Foundation under Awards EEC-0813570 and IIS-0612240. Any opinions, findings, and conclusions or

recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

1. Liao JC, Boscolo R, Yang Y-L, Tran LM, Sabatti C, Roychowdhury VP: **Network component analysis: Reconstruction of regulatory signals in biological systems**. *Proceedings of the National Academy of Sciences of the United States of America* 2003, **100**(26):15522-15527.

2. Bussemaker HJ, Foat BC, Ward LD: **Predictive Modeling of Genome-Wide mRNA Expression: From Modules to Molecules**. *Annual Review of Biophysics and Biomolecular Structure* 2007, **36**(1):329-347.

3. Tran LM, Brynildsen MP, Kao KC, Suen JK, Liao JC: **gNCA: A framework for determining transcription factor activity based on transcriptome: identifiability and numerical implementation**. *Metabolic Engineering* 2005, **7**(2):128-141.

4. Chang C, Ding Z, Hung YS, Fung PCW: **Fast network component analysis (FastNCA) for gene regulatory network reconstruction from microarray data**. *Bioinformatics* 2008, **24**(11):1349-1358.

5. Alter O, Golub GH: **Integrative analysis of genome-scale data by using pseudoinverse projection predicts novel correlation between DNA replication and RNA transcription**. *Proceedings of the National Academy of Sciences of the United States of America* 2004, **101**(47):16577-16582.

6. Gao F, Foat B, Bussemaker H: **Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data %U http://www.biomedcentral.com/1471-2105/5/31**. *BMC Bioinformatics* 2004, **5**(1 %M doi:10.1186/1471-2105-5-31):31.

7. Boulesteix A-L, Strimmer K: **Predicting transcription factor activities from combined analysis of microarray and ChIP data: a partial least squares approach**. *Theoretical Biology and Medical Modelling* 2005, **2**(1 %M doi:10.1186/1742-4682-2-23):23.

8. Nachman I, Regev A, Friedman N: **Inferring quantitative models of regulatory networks from expression data**. *Bioinformatics* 2004, **20**(suppl_1):i248-256.

9. Li Z, Shaw SM, Yedwabnick MJ, Chan C: **Using a state-space model with hidden variables to infer transcription factor activities**. *Bioinformatics* 2006, **22**(6):747-754.

10. Sanguinetti G, Rattray M, Lawrence ND: **A probabilistic dynamical model for quantitative inference of the regulatory mechanism of transcription**. *Bioinformatics* 2006, **22**(14):1753-1759.

11. Gao P, Honkela A, Rattray M, Lawrence ND: **Gaussian process modelling of latent chemical species: applications to inferring transcription factor activities**. *Bioinformatics* 2008, **24**(16):i70-75.

12. Hecker M, Lambeck S, Toepfer S, van Someren E, Guthke R: **Gene regulatory network inference: Data integration in dynamic models--A review**. *Biosystems* 2009, **96**(1):86-103.

13. Vlieghe D, Sandelin A, De Bleser PJ, Vleminckx K, Wasserman WW, van Roy F, Lenhard B: **A new generation of JASPAR, the open-access repository for transcription factor binding site profiles**. *Nucleic Acids Research* 2006, **34**(suppl_1):D95-97.

14. Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E *et al*: **Genome-Wide Location and Function of DNA Binding Proteins**. *Science* 2000, **290**(5500):2306-2309.

15. Stuart JM, Segal E, Koller D, Kim SK: **A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules**. *Science* 2003, **302**(5643):249-255.

16. Butte AJ, Kohane IS: **Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements**. In.; 2000.

17. Basso K, Margolin AA, Stolovitzky G, Klein U, Dalla-Favera R, Califano A: **Reverse engineering of regulatory networks in human B cells**. *Nat Genet* 2005, **37**(4):382-390.

18. Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, Cottarel G, Kasif S, Collins JJ, Gardner TS: **Large-Scale Mapping and Validation of Escherichia coli Transcriptional Regulation from a Compendium of Expression Profiles**. *PLoS Biol* 2007, **5**(1):e8.

19. Prill RJ, Marbach D, Saez-Rodriguez J, Sorger PK, Alexopoulos LG, Xue X, Clarke ND, Altan-Bonnet G, Stolovitzky G: **Towards a rigorous assessment of systems biology models: the DREAM3 challenges**. *PLoS One* 2010, **5**(2):e9202.

20. Watkinson J, Liang KC, Wang X, Zheng T, Anastassiou D: **Inference of regulatory gene interactions from expression data using three-way mutual information**. *Ann N Y Acad Sci* 2009, **1158**:302-313.

21. Madar A, Greenfield A, Vanden-Eijnden E, Bonneau R: **DREAM3: network inference using dynamic context likelihood of relatedness and the inferelator**. *PLoS One* 2010, **5**(3):e9803.

22. Greenfield A, Madar A, Ostrer H, Bonneau R: **DREAM4: Combining genetic and dynamic information to identify biological networks and dynamical models**. *PLoS One* 2010, **5**(10):e13397.

23. Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P: **Inferring regulatory networks from expression data using tree-based methods**. *PLoS One* 2010, **5**(9).

24. Pinna A, Soranzo N, de la Fuente A: **From knockouts to networks: establishing direct cause-effect relationships through graph analysis**. *PLoS One* 2010, **5**(10):e12912.

25. Guthke R, Moller U, Hoffmann M, Thies F, Topfer S: **Dynamic network reconstruction from gene expression data applied to immune response during bacterial infection**. *Bioinformatics* 2005, **21**(8):1626-1634.

26. Gardner TS, di Bernardo D, Lorenz D, Collins JJ: **Inferring Genetic Networks and Identifying Compound Mode of Action via Expression Profiling**. *Science* 2003, **301**(5629):102-105.

27. di Bernardo D, Thompson MJ, Gardner TS, Chobot SE, Eastwood EL, Wojtovich AP, Elliott SJ, Schaus SE, Collins JJ: **Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks**. *Nat Biotech* 2005, **23**(3):377-383.

28. Du P, Dickerson, J.A.: **Multi-scale genetic network inference based on time series expression profiles**. In: *ICSB (International Conference of Systems Biology)* Boston, MA; 2005.

29. Bansal M, Gatta GD, di Bernardo D: **Inference of gene regulatory networks and compound mode of action from time course gene expression profiles**. *Bioinformatics* 2006, **22**(7):815-822.

30. Kaleta C, Gohler A, Schuster S, Jahreis K, Guthke R, Nikolajewa S: **Integrative inference of gene-regulatory networks in Escherichia coli using information theoretic concepts and sequence analysis**. *BMC Syst Biol* 2010, **4**:116.

31. van Berlo RJP, van Someren EP, Reinders MJT: **Studying the Conditions for Learning Dynamic Bayesian Networks to Discover Genetic Regulatory Networks**. *SIMULATION* 2003, **79**(12):689-702.

32. Perrin B-E, Ralaivola L, Mazurie A, Bottani S, Mallet J, d'Alche-Buc F: **Gene networks inference using dynamic Bayesian networks**. *Bioinformatics* 2003, **19**(suppl_2):ii138-148.

33. Seok J, Kaushal A, Davis RW, Xiao W: **Knowledge-based analysis of microarrays for the discovery of transcriptional regulation relationships**. *BMC Bioinformatics* 2010, **11 Suppl 1**:S8.

34. Gama-Castro S, Salgado H, Peralta-Gil M, Santos-Zavaleta A, Muniz-Rascado L, Solano-Lira H, Jimenez-Jacinto V, Weiss V, Garcia-Sotelo JS, Lopez-Fuentes A *et al*: **RegulonDB version 7.0: transcriptional regulation of Escherichia coli K-12 integrated within genetic sensory response units (Gensor Units)**. *Nucleic Acids Res* 2011, **39**(Database issue):D98-105.

35. Liang K-C: **Gene Regulatory Network Reconstruction Using Conditional Mutual Information**. In. Edited by Wang X, vol. 2008. EURASIP Journal on Bioinformatics and Systems Biology; 2008: 14 pages.

36. Faith JJ, Driscoll ME, Fusaro VA, Cosgrove EJ, Hayete B, Juhn FS, Schneider SJ, Gardner TS: **Many Microbe Microarrays Database: uniformly normalized Affymetrix compendia with structured experimental metadata**. *Nucleic Acids Res* 2008, **36**(Database issue):D866-870.

37. Keseler IM, Bonavides-Martinez C, Collado-Vides J, Gama-Castro S, Gunsalus RP, Johnson DA, Krummenacker M, Nolan LM, Paley S, Paulsen IT *et al*: **EcoCyc: A comprehensive view of Escherichia coli biology**. *Nucleic Acids Research* 2009, **37**(suppl_1):D464-470.

38. Schwartz CJ, Giel JL, Patschkowski T, Luther C, Ruzicka FJ, Beinert H, Kiley PJ: **IscR, an Fe-S cluster-containing transcription factor, represses expression of Escherichia coli genes encoding Fe-S cluster assembly proteins**. *Proceedings of the National Academy of Sciences of the United States of America* 2001, **98**(26):14895-14900.

39. Brynildsen MP, Liao JC: **An integrated network approach identifies the isobutanol response network of Escherichia coli**. *Mol Syst Biol* 2009, **5**:277.

40. Wang C, Xuan J, Chen L, Zhao P, Wang Y, Clarke R, Hoffman E: **Motif-directed network component analysis for regulatory network inference**. *BMC Bioinformatics* 2008, **9**(Suppl 1 %M doi:10.1186/1471-2105-9-S1-S21):S21.

41. Baba-Dikwa A: **Overproduction, purification and preliminary X-ray diffraction analysis of YncE, an iron-regulated Sec-dependent periplasmic protein from Escherichia coli**. In., vol. 64(Pt 10);. Acta Cryst.; 2008: 966-969.

42. Takahashi Y, Nakamura M: **Functional Assignment of the ORF2-iscS-iscU-iscA-hscB-hscA-fdx-0RF3 Gene Cluster Involved in the Assembly of Fe-S Clusters in Escherichia coli**. *Journal of Biochemistry* 1999, **126**(5):917-926.

43. Tokumoto U, Takahashi Y: **Genetic Analysis of the isc Operon in Escherichia coli Involved in the Biogenesis of Cellular Iron-Sulfur Proteins**. *Journal of Biochemistry* 2001, **130**(1):63-71.

44. Vickery LE: **Hsc66 and Hsc20, a new heat shock cognate molecular chaperone system from Escherichia coli**. In. Edited by Jonathan J. Silberg DTT, vol. 6(5). Protein Sci.; 1997: 1047-1056.

45. **The EcoGene Database** [http://www.ecogene.org]. In., 2005-08-05 edn.

CHAPTER III

INTEGRATED APPROACH TO ANALYZE *E. COLI* GENE REGULATORY NETWORK BEHAVIORS BETWEEN EXPERIMENTAL CONDITIONS

A paper to be submitted to *BMC Bioinformatics*

Yao Fu[1], Laura R Jarboe[2] and Julie Dickerson[1, 3§†]

Abstract

Background

Genes differentially express under different experimental conditions through the dynamic control of gene regulatory networks (GRNs). In GRNs, transcription factors (TFs) serve as central links between gene expressions and environmental conditions. Transcription facto activities (TFAs) reflect the dynamics of regulatory effects of transcription factors (TF) to its target genes under different intra and extra cellular environments.

[1]Bioinformatics and Computational Biology Program, Iowa State University, Ames, Iowa, U.S.A

[2]Chemical and Biological Engineering Department, Iowa State University, Ames, Iowa, U.S.A

[3]Electrical and Computer Engineering Department, Iowa State University, Ames, Iowa, U.S.A

[§]Corresponding author

Results

In this work, an integrated analysis framework based on TFAs estimated from transcriptome data under different experimental conditions and with gene regulatory network information is proposed. This analysis framework compares and analyzes gene regulations under different conditions at 3 levels: individual TFA patterns across conditions; groups of TFs share some common features such as regulatory effects, signal sensing mechanisms, and functional pathways; and proposed novel effective regulatory networks model under different conditions to analyze the gene regulations at network and conditions specific level.

Conclusion

These three levels of analysis are carried out on TFAs estimated from transcriptome data of 20 different experimental conditions of MOPS media. Many biological meaningful and useful results have been shown from this analysis including. E.g., 27 activities consistent TFs across all 20 MOPS media conditions and 18 condition specific TFs under 7 different MOPS media conditions have been identified from the individual TF level analysis; enriched conditions of groups of TFs have been found through group level analysis; and key TFs of each MOPS media conditions have been identified from the network level analysis.

Background

Differentially expressed genes play important roles in the response of microorganisms, to environmental stimulations and the resulting control of their

phenotypes [1, 2]. Transcriptome experiments, such as RNA-Seq and DNA microarrays, make it possible to study the expression of genes for different experimental conditions. The Many Microbe Microarrays (M3D) database was built to collect and normalize these transcriptome experiments under different environmental and experimental conditions of *E. coli* and *S. cerevisiae* [3].

However, gene expression information alone is not sufficient to understand the details of how cells respond and adapt to environmental changes. A gene regulatory network (GRN), or transcriptional regulatory network, collects regulatory interaction between transcription factors (TFs) and genes. TFs sense regulatory signals and change their activities, transcription factor activities (TFAs), to regulate the expression of their target genes. Bacterial phylogenetic studies showed that TFs and regulatory networks evolve much faster than their target genes and suggested that TFs and GRNs play more important roles than their target genes to adaptation to environment changes [1, 4]. Many regulatory connections between TFs and their target genes have been identified. These regulatory connections are collected by databases such as RegulonDB for *E. coli* [5].

GRNs are complex networks: according to RegulonDB [5], 909 of the 1,578 transcriptionally-regulated genes in *E. coli* are regulated by more than one TF. An analysis of microarray data from four experimental conditions (minimal medium, heat shock, stationary phase, and anaerobic growth), combined with known regulatory information from RegulonDB, showed that gene expression of regulators varied under different experimental conditions [6]. Thus, the activated GRNs are rewired by activating or de-activating TFs dynamically during environmental changes. Studies of

GRNs in *S. cerevisiae* showed that the topologies of GRNs are significantly changed between different experimental conditions [7]. Recent review concluded that biological networks are commonly regulated and rewired to adapt environmental changes, and suggested that the study on the network differences would become standard for studying the organisms under different environmental conditions [8]. Other researchers classified and analyzed the sensing machinery of TFs and pointed out the importance of TFs' sensing environmental signals to the GRNs and the study of condition-specific network behaviors [2, 9, 10].

To further understand the behavior of GRNs under different experimental conditions, Janga *et al*. performed a systematic analysis of the expression patterns of TFs of *E. coli* across all of the 466 experimental conditions from M3D database [11]. Janga *et al*. clustered the experimental conditions of M3D database to remove bias from redundant experiments, defined activated TFs based on the expression of the TF encoding genes, performed enrichment tests on different groups of TFs, including different regulatory effects groups and different signal sensing groups [2], and finally identified some marker TFs for experimental conditions using network based analysis methods.

The initial analyses of TF behavior assumed the activities of TFs are correlated with the expression of TF-encoding genes. However, the expression of TF-encoding genes does not ensure the activation and successful regulation of target genes. For example, the well-known TF LacI represses the expression of lactose transport and catabolism genes when it is in the active state. However, when allolactose is present and

bound to LacI, even though LacI is expressed, the DNA-binding activity and the repression effect of LacI is inhibited. The expression of genes regulated by LacI is not be affected by the expression of LacI [12, 13]. Thus, analyses that model TFA based solely on TF gene expression cannot fully reflect the behaviors of GRNs.

TFAs indicate the ability of TFs to regulate their target genes. TFA log ratios between experimental conditions can be computationally predicted from transcriptome data with an initial input of known regulatory network topology [14]. Typical TFA prediction algorithms include the expectation maximization based Network Component Analysis (NCA) [14, 15], singular value decomposition based Fast Network Component Analysis (FastNCA) [16] and partial least square based statistically inspired modification of the partial least square (SIMPLS) [17]. But there are several limitations of these TFA prediction algorithms. NCA and FastNCA have strict limitations on the initial regulatory networks that does not allow redundant regulatory patterns of TFs (TFs that always co-regulate the same genes) [14-16]. This limitation results in an eliminated set of TFA predictions that only includes TFs with non-redundant regulatory patterns and one TF from each group of redundant TFs. A previous study of using predicted TFAs to reconstruct GRNs showed that though SIMPLS could model all TFs without eliminating the set of regulatory pattern redundant TFAs, the predictions from SIMPLS is not as precise as predictions from NCA or FastNCA [18]. Another common problem of these TFA prediction algorithms is that changes in TFA direction cannot be predicted. In other words, these TFA prediction algorithms can predict the scale of the TFA log ratio change between experimental conditions, but cannot confidently predict the direction of TFA

change, e.g. more or less TF activation.

Analysis of TFAs under different experimental conditions gives more insight of the behavior of GRNs. However, since the current TFA prediction algorithms are unable to identify the direction of the TFA changes, the meaningfulness and applications of TFA based analysis is limited. In this paper, directed network component analysis (D-NCA), a direction-corrected TFA estimation algorithm developed from the original NCA, is proposed. D-NCA can correct the TFA changing directions from NCA by comparing the predicted TFAs with reference gene regulatory interactions and the original transcriptome data. The D-NCA algorithm also estimates the eliminated TFs from NCA by SIMPLS predictions. Thus, D-NCA predicts a set of comprehensive and biologically meaningful TFAs for further analysis.

To study the behaviors of GRNs of *E. coli* under different experimental conditions, this work uses a systematic three level analysis on TFAs predicted from M3D and RegulonDB *E. coli* data using D-NCA (**Figure 3-1**) instead of analyzing gene expressions of TF encoding genes, as performed by Janga and Contreras-Moreira [11]. The first level of analysis is on individual TFs and analyzes the behavior of each TF across different experimental conditions. The second level of analysis groups TFs by different types of TF properties, e.g. regulatory effects, signal sensing machineries or pathways they regulate. Then, enrichment tests for each group of TFs were performed for different experimental conditions. The final level is the network level. In this level of analysis, a novel Effective Regulatory Network (ERN) model is proposed to capture the GRN differences between experimental conditions. This level of analysis is based on

ERNs and identifies the key TFs for experimental conditions which significantly change network properties and successfully regulate target genes.

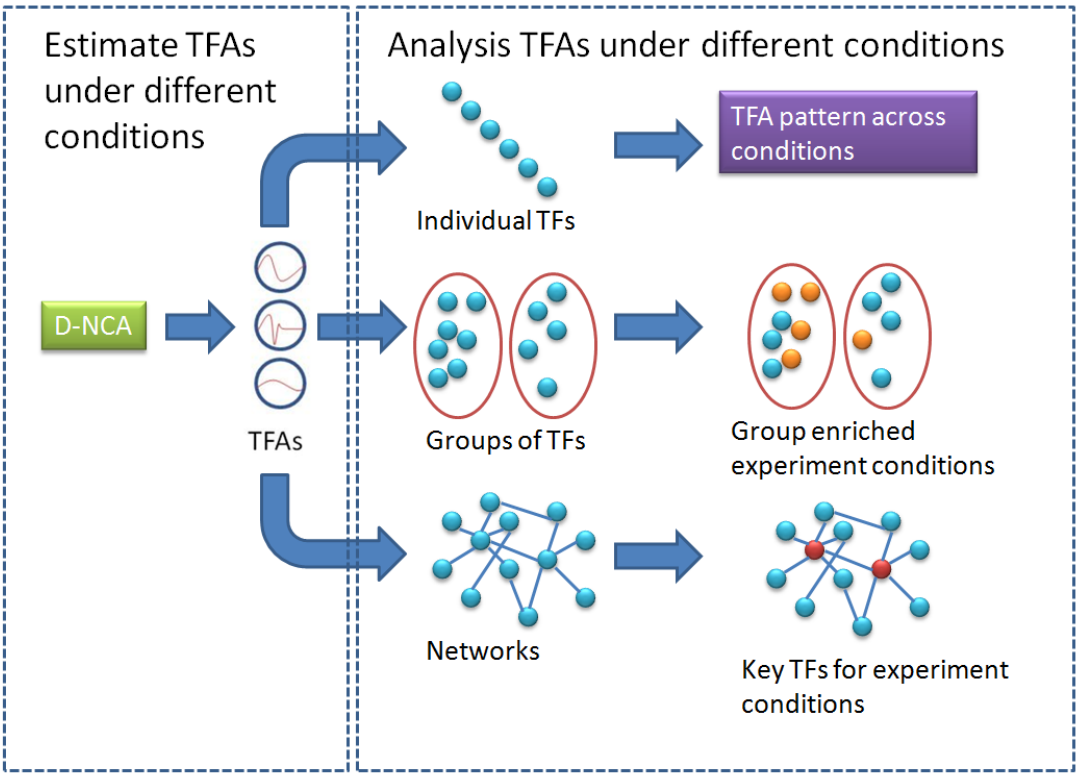


Figure 3-1 *Three levels of analysis on TFAs*

TFAs can be estimated from transcriptome data and known GRN network using D-NCA algorithm. Analysis of TFAs has three levels: analysis of individual TFA patterns to learn the behavior of individual TFs, analysis of the TFAs in groups to find enriched TF groups under certain experimental conditions, and analysis the network effects brought by TFA changes to identify the key TFs that are most responsible to the changes of the whole network under certain condition differences.

Methods

Analysis of activities on each individual TF

Directed Network Component Analysis

The activities of TFs can vary under different experimental conditions. According to previous research, NCA method has the best TFA prediction precisions among all the currently available TFA prediction methods [18]. In this work, NCA is used to initially estimate TFA log ratios between treatment conditions and the control condition [14, 15]. For example, under MOPS media experiments set of M3D database, WT_MOPS_glucose is the control condition, TFA ratios between the control condition and all other MOPS experimental conditions can be estimated. The NCA algorithm predicts TFA log ratios from transcriptome data. However, there are several restrictions on NCA algorithm, e.g., not being able to tell a meaningful direction of TFA changes and not being able to estimate activities of full set of TFs with redundant regulatory relationships among TFs.

To overcome these restrictions of the NCA algorithm, a directed network component analysis (D-NCA) method is developed in this study. D-NCA first predicts TFA log ratios using the NCA algorithm to collect the best prediction available, then fills the TFA log ratios which NCA could not predict because of the restriction described above with TFA predictions from SIMPLS, as SIMPLS can predict a full set of TFA log ratios but with lower accuracy than NCA predictions [17]. For those TFAs which are

predicted by both NCA and SIMPLS, only use NCA prediction in the following analysis. Then, the changing direction of predicted TFA log ratios is determined by correcting the predicted TFA changing directions to match the transcriptome experiment results of gene expressions and the regulatory relationships between TFs and genes to meet the following criteria:

A) For promoting regulation:

$$\frac{\log(TFARatio)}{\log(\text{ExpresionRatio})} > 0$$

B) For repression regulation:

$$\frac{\log(TFARatio)}{\log(\text{ExpresionRatio})} < 0$$

The detailed procedure of D-NCA method is described as following:

1. Run NCA to predict TFA log ratios between treatment and control conditions [15, 18]

2. Run SIMPLS to predict TFA log ratios and fill NCA-removed TFA log ratios with SIMPLS results [17, 18]

3. For each TF *i*, the TFA log ratio for each treatment condition *c* is compared to the control condition(*s*). *TFA(i, c)* is computed from step 1 and 2.

3.1 For each target gene *j* of this TF *i*, Steps 1 and 2 generate a log ratio gene expression level *E(j,c)* for each treatment condition, and a regulatory interaction effecter *D(i,j)* from

the initial TF-Gene topology input of TFA prediction algorithms, where *D(i,j)=1* for up regulating effects, *D(i,j) = -1* for down regulating effects, and *D(i,j)=0* for unknown or dual regulatory effects.

3.1.1 Calculate the correlation coefficient *C(i,j)* between *TFA(i)* and *E(j).*

$$C(i,j) = \frac{\sum_{c}(\mathrm{TFA}(i,c) - \overline{TFA(i)})(\mathrm{E}(j,c) - \overline{\mathrm{E}(j)})}{\sqrt{\sum_{c}(\mathrm{TFA}(i,c) - \overline{TFA(i)})^2 \sum_{c}(\mathrm{E}(j,c) - \overline{\mathrm{E}(j)})^2}}$$

Where $\overline{TFA(i)}$ is the average of *TFA(i,c)* for all the conditions, and $\overline{E(\mathrm{j})}$ is the average of *E(j,c)* for all the conditions.

3.1.2 Calculate the Direction Match Correlation (DMC) coefficient

$$DMC(i,j) = C(i,j) \times D(i,j)$$

3.2 Sum up all the *DMC* of this TF *i,*

$$SDMC(\mathrm{i}) = \sum_{j=\text{all the target genes of TF } i} DMC(i,j)$$

3.3 If the *SDMC(i)* is negative, reverse the sign of *TFA(i,c)* for TF *i* (set *TFA(i, c) = - TFA (i, c)*).

4. Repeat step 3 for all TFs

Analysis on groups of TFs

Define activated TFs under experimental conditions

Although TFA is a continuous property, the activation status of TFs under each experimental condition is categorized into two statuses, activated and not activated. Under a specific experimental condition, an activated TF should actively perform regulation functions and connect regulatory elements in a GRN. The regulatory functions of not activated TFs are silenced and not involved in the GRN under this condition. K-means clustering was used to determine the active status of TFs under each experimental condition by clustering TFAs under each experimental condition into two groups [19]. Within each group, K-means minimizes the sum of squared distances between points of the group. The group of TFs with higher TFA ratio values is considered as activated group of TFs, and the TFs with lower TFA values are considered as not activated group of TFs.

Enrichment tests of activated TFs on groups of TFs

Enrichment tests can be applied to different groups of TFs, e.g., grouped by regulatory effects such as activators, repressors and dual regulator [20], or grouped by the signal type sensed, such as internal signals, external signals and hybrid (sensing signals which can from both inside and outside of the cell) (**Table 3-1**) [10], and pathways that they regulate [21].

Signal Source	Signal Sensing Group	Signal Sensing Mechanism	No. of TFs
External	ETM	Sense transportable metabolites	28
	ETC	Part of two component systems	29
Internal	ISM	Sense/bind metabolites generated by cellular metabolism	30
	IDB	DNA-binding TFs	4
Hybrid	Hybrid	Sense metabolites from both endogenous and exogenous origin	33

Table 3-1 *TF Signal Sensing Groups [9]*

TFs are grouped by signal sensing mechanisms: External signal sensing TFs include TFs that are part of two-component systems (ETC) or that sense transportable metabolites (ETM); Internal signal sensing TFs sense endogenous or intracellular stimuli (I), including TFs binding/sensing metabolites generated by cellular metabolism (ISM), and DNA-binding TFs for nucleoid or chromosome remodeling and compaction(IDB); Hybrid TFs sense metabolites from both endogenous and exogenous origin (Hybrid)

The enrichment test for TFAs tests the null hypothesis that under the specific treatment condition, there is no association between the list of activated TFs and the list of TFs in pre-selected category group which is grouped by different properties of TFs as described above [22]. No significant rejection (p-value greater than 0.05) of the null hypothesis would indicate that the number of activated TFs in the pre-selected group is the same as that under the control condition. Rejection of the null hypothesis suggests significant relationships between pre-selected group and the activated TFs under given treatment condition, thus identifies the associations between TF types/categories and environmental conditions. To test this null hypothesis, hypergeometric probabilities are used. The subject (condition) sampling enrichment test procedures are as follows:

1. For each group of TFs, estimate the expected number of TFs *E* that would be activated under the control condition by counting the activated TFs of each group of TFs

under the control condition.

2. For each group of TFs under each treatment condition, calculate the p-value of activated TFs within this group of TFs using hypergeometric distribution (The probability of at least the number of TFs being activated or the probability of at most the number of TFs being activated, depending on which tail the number of activated TFs is on).

The probability mass function of hypergeometric distribution calculated as following:

$$P(X = A)=\binom{E}{A}\binom{N-E}{N-A}$$, where N is the total number of TFs in the tested TF feature group, A is the number of activated TFs of the tested TF feature group under the tested experimental condition.

3. Bonferroni adjustment multiple test correction for multiple tests on each condition to identify TFA enriched conditions of each group of TFs. Significant enriched groups of TFs under given experimental condition are the groups of TFs with tested *p-value* less than 0.05 divided by the total number of tested enrichment tests.

Target genes weighted enrichment tests for pathway groups of TFs

TFs may regulate many metabolic pathway genes which involve catalysis of reactions or transportation in metabolic pathways. Metabolic pathways are defined in EcoCyc database based on curators' expertise and aspects including historically

definitions, currency metabolites, regulatory units and evolutionarily conserved metabolic units [21, 23].There are usually multiple TFs regulating each pathway, but the number of pathway genes controlled by each TF differs. To test TFA enrichment of metabolic pathways, a target genes weighted enrichment test is developed in this study. This target genes weighted enrichment test procedure is similar to the enrichment test described above, but in step 1, instead of estimating the expected number of TFs would be activated, here the expected number of TFs controlled target genes is estimated by counting the number of all the target genes of activated TFs of the tested group of TFs under the tested experimental condition. In the second step, instead of calculating the probability of activated TFs within the tested group of TFs, the probability of affected genes within all the group of TFs controlled genes is calculated.

To identify how different pathways perform their roles under environment condition changes, TFs were grouped by the pathways they regulate. EcoCyc currently lists 334 pathways [21], but most of these pathways contain too few genes to perform a reasonable enrichment test. Thus, only 23 pathways involving more than 10 genes each were selected to perform the enrichment test. The 23 selected pathways were tested using target gene weighted enrichment test as described above..

Regulatory network based analysis

Effective Regulatory Network (ERN) of treatment conditions

An effective regulatory link is defined as a regulatory link whose source TF shows

effective regulation which leads expression differences of its target genes between two different experimental conditions (**Figure 3-2**). For example, TF *T* and Gene *b*, under two different experimental conditions, if the activity of TF *T* is increased significantly (two-sample t-test *p-value* less than 0.05 with multiple tests correction) and the expression of gene *b* varies significantly (two-sample t-test *p-value* less than 0.05 with multiple tests correction) from one condition to the other, TF *T* is effectively regulating gene *b* under the different conditions. Thus, there is an effective regulatory link between TF *T* and gene *b*. Similarly, an effective expression link is constructed from the TF encoding gene to its respective TF when a significant expression change of TF encoding gene leads to significant activity changes of the respective TF. For example, if gene *t*

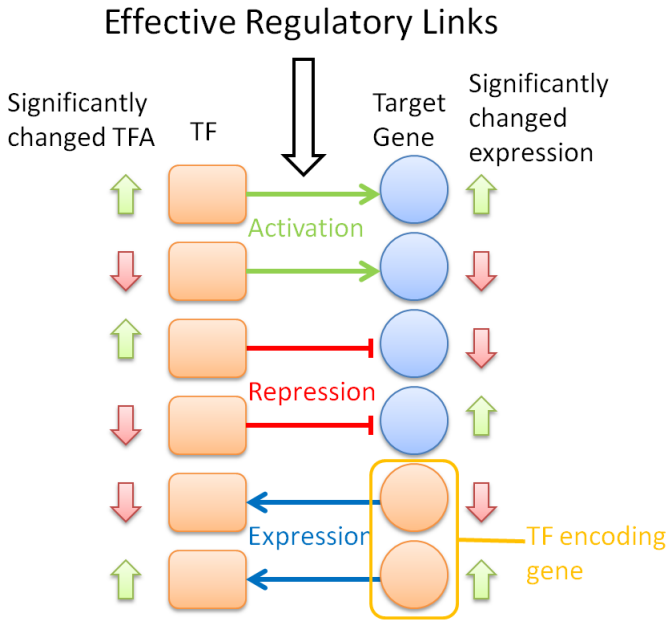


Figure 3-2 Effective Regulatory Links
Regulatory links with matched TFA changes and gene expression changes. E.g. up-regulating links connecting TFAs and gene expression both significantly changed up or down, and down-regulating links connecting more activated TFs with less expressed genes or less activated TFs with more expressed genes.

number of tested genes) genes and the direction of gene expression changes between the two conditions being tested.

3. Identify activity significantly changed TFs (two-sample t-test *p-value* less than 0.05/the total number of tested TFs) and the directions of TFA changes between the two conditions being tested using the methods described above.

4. For each activity significantly changed TF and each of its target genes with significantly differential expressions, if the regulatory effects of the TF under this condition comparison agree with the expression differences of the differentially expressed target gene. Keep the regulatory link between the TF and the differentially expressed gene. Otherwise, remove the link between them from the network generated by step one as the expression and TF changes are not consistent between the conditions.

5. If there is no significant change in activity for a TF, remove the TF and all its regulatory links from the GRN generated by step 1.

6. For each significantly differentially expressed TF encoding gene, if the sign of activity change of the TF being encoded matches the gene expression change of the encoding gene, keep the expression link between them, otherwise, remove the expression link.

The resulting ERN under a specific experimental condition comparison describes a regulatory network within which all the regulatory links are effective in terms of successfully regulating the target genes in response to the significant TFA changes. In

the resulting ERN, all regulatory signals are effectively passing to all of their target nodes and causing gene expression differences between conditions. This ERN masks many regulatory links which do not perform regulation under the specific experimental condition differences from traditional GRN, and results in a more meaningful network to study the differences of gene regulation under different experimental conditions.

Key TFs and Network properties of TFs under experimental conditions

To further identify important regulatory elements of ERNs, Key TFs of ERNs are defined as the TFs in the ERNs which contribute the most to the regulatory network topology changes in response of the experimental condition differences. Janga *et al.* used network properties to indicate and identify the TFs' contribution to the network topology changes and identify Marker TFs in their previous work [11]. In this work, similar method is taken as the criteria for Key TFs.

Three network properties, including degree, network size normalized degree, betweenness centrality, and downstream closeness centrality, are calculated for each TF node in each ERN. These network properties can differ greatly from ERN to ERN and are important criteria for determining the changes of each regulatory node in an ERN under specific condition differences and used as criteria to identify the key TFs underlying each condition difference. The following network properties of ERN nodes have been studied:

Degree: the number of connections of each node. This property shows the number of

direct effective regulatory connections to each node under certain ERN and condition difference.

Network size normalized degree: degree divided by the total number of all other nodes within the network. This property shows the direct coverage of specific TF to the whole ERN.

Betweenness centrality summarizes the number of shortest paths going through a node. This property summarizes the role and importance of a node in term of serving as a signal hub which quickly receives signals from source signal pathways and broadcasts the regulatory signals to downstream signal pathways, and is calculated as:

$$C_B(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

Where $C_B(v)$ is the Betweenness centrality of node v, σ_{st} is the number of shortest paths from node s to node t, and $\sigma_{st}(v)$ is the number of shortest paths from node s to node t passing through node v. The implementation to calculate betweenness centrality for the directed graph like TF-TF regulatory networks follows a algorithm proposed by Ulrik Brandes [24].

Downstream closeness centrality measures the speed of regulatory signal spread out to the network and reach every node. It is calculated as the inverse of the shortest downstream distance to all other nodes.

$$C_C(v) = \frac{n-1}{\sum_{t \in V \setminus v} d_G(v,t)}$$

Where $C_c(v)$ is the downstream closeness centrality of node v, n is the total number of nodes in the graph, $d_G(v,t)$ is the shortest downstream distance between node v and node t. If there is no path between node v and node t, dG(v,t) are defined as n-1.

A base network is generated as described in the Step 1 of the process to generate ERN, which is the whole traditional GRN plus the expression links between TF encoding genes and the TFs being encoded. Degrees, betweenness centralities and closeness centralities are normalized for each TF in the ERN by being divided by degrees, betweenness centralities and closeness centralities of each TF in the base network respectively.

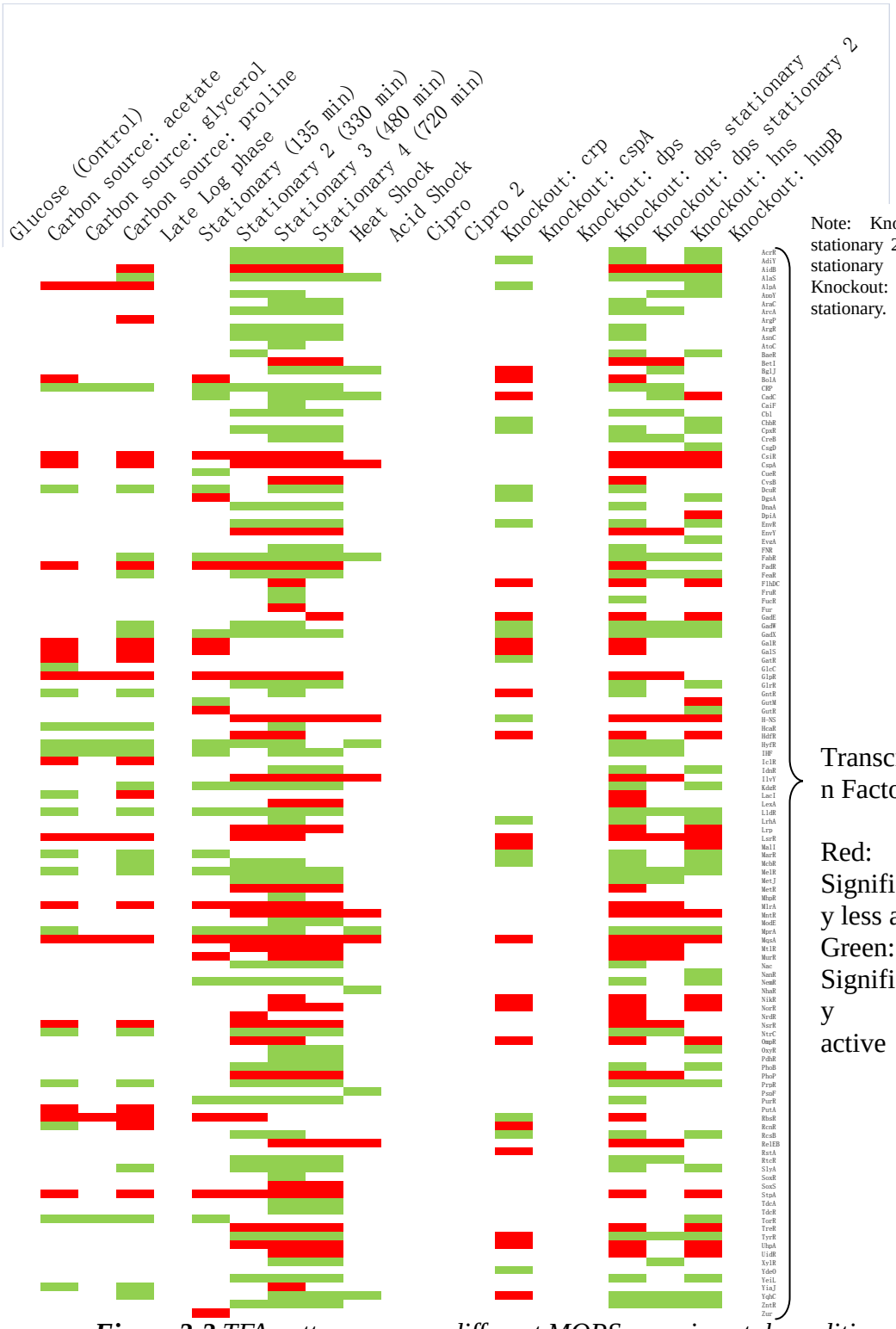
Results and Discussion

Activity patterns of TFs across experimental conditions

MOPS media is a minimal medium, normally used for aerobic growth controlled pH at 7.2; MOPS media composition was defined for the purpose of separating all element sources to facilitate isotopic labeling [25]. The wild type *E. coli* strain MG1655 in MOPS media with 2% glucose (WT_MOPS_glucose) was chosen as the control condition of MOPS media experimental conditions. Twenty microarray experiments in MOPS media were chosen from the M3D database to demonstrate our approach to analyzing the regulatory network behavior of *E. coli* across different environmental

conditions. The detailed compositions of MOPS media and all the 20 experimental conditions were collected from EcoCyc [21] and M3D [3] databases and are shown in **APPENDIX E**. TFA log ratios of 176 TFs in these 20 experiments in MOPS media from M3D database were estimated using the D-NCA algorithm. The initial TF-gene regulatory network of D-NCA input is the regulatory network obtained from RegulonDB 7.2 [20].

Significantly changed TFAs from the control condition to treatment conditions were identified using two sample t-tests between treatment conditions and the control condition. A Bonferroni multiple testing correction was applied to limit the false positive detection rate.



Note: Knockout: dps stationary 2 is the later stationary phase than Knockout: dps stationary.

Figure 3-3 *TFA patterns across different MOPS experimental conditions*

Figure 3-3 *continued*

TFA patterns across experimental conditions: The TFA patterns shown in this figure are all compared to the control experimental condition (wild type MOPS with glucose carbon source). There are 7 experimental conditions have no significant TFA changes from the control conditions. These 7 experimental conditions are wild type late log (Late Log phase), wild type acid shock (Acid Shock), wild type cipro (Cipro), wild type cipro 2 (Cipro2), cspA knockout (Knockout: cspA), dps knockout (Knockout: dps), and hupB knockout (Knockout: hupB). There are 27 TFs have no significant TFA changes under all the 20 MOPS media experimental conditions (Table 3-2). 18 TFs are condition specific, and each only significantly perturbed under one of the MOPS media experiment conditions we tested (Table 3-3). TF MqsA is significantly less activated under all the conditions other than the control and control similar conditions. And TFA patterns under experimental conditions at stationary phases are the most different from the control condition which are sampled at the early exponential phase.

TFA patterns under experimental conditions at stationary phases are the most different from the control condition which are sampled at the early exponential phase **(Figure 3-3)**. This result is consistent with biological knowledge that regulatory system plays an important role in growth phase changes of *E. coli*.

According to the patterns of activities significantly changed TFs (**Figure 3-3**), there are 7 experimental conditions with no significantly changed TFAs relative to the control condition. In terms of TFA changes, these 7 experimental conditions are similar to the control condition (wild type, glucose carbon source). Due to the difference of regulatory mechanisms and the sensing signals of TFs, the activity patterns of TFs across experimental conditions vary.

There are 27 TFs that have no significant TFA change under all 20 experimental conditions (**Table 3-2**). For example, AscG, a repressor of a cryptic operon *ascFB*, has consistent activity across the 20 MOPS media experimental conditions (**Table 3-2**), and literature suggests that this regulator is only de-activated when its gene is interrupted by an insertion sequence [26].

18 TFs are condition specific, and each are only significantly perturbed under one of the experimental conditions analyzed (**Table 3-3**). These condition specific TFs may have special roles in response of respective condition changes. For example, CsgD, which responds to starvation and high cell density [27], is specifically perturbed at the beginning of stationary phase (experiment Stationary), and might serve as an indicator of initiating the stationary phase.

TF	Number of Target Genes	Regulatory effects	Signal Sensing
AgaR	9	Repressor	ETM
AllS	3	Activator	Hybrid
AlsR	6	Repressor	
AscG	5	Repressor	ETM
CdaR	10	Activator	
CynR	4	Dual Regulator	ETM
CytR	12	Repressor	Hybrid
DhaR	4	Dual Regulator	
DicA	2	Repressor	
DsdC	3	Dual Regulator	Hybrid
ExuR	8	Repressor	ETM
Fis	222	Dual Regulator	
HU	9	Dual Regulator	IDB
IscR	27	Dual Regulator	
KdpE	4	Activator	ETC
LeuO	20	Dual Regulator	ISM
LysR	2	Dual Regulator	Hybrid
MarA	37	Dual Regulator	
MngR	3	Repressor	
NadR	4	Repressor	
NarL	113	Dual Regulator	ETC
Rob	24	Activator	
RutR	16	Dual Regulator	
SdiA	5	Dual Regulator	Hybrid
SgrR	7	Dual Regulator	
XapR	2	Activator	ETM
ZraR	3	Activator	ETC

Table 3-2 *Activity consistent TFs*

TFs with no significant change in TFA relative to the wild type control in any of the 20 experimental conditions. The TFs shown in this table have consistent activity values across all the tested conditions under MOPS media, any perturbation of these TFs detected in other MOPS medium experiments could be of great interest.

Estimated TFAs show differing patterns when compared across conditions. Some TFs are extremely stable and had no significant change in activity across all the MOPS

media experimental conditions (**Table 3-2**). Some TFs are very specific to certain conditions (**Table 3-3**) and their activities were only significantly changed under one experimental condition versus the control condition. Also, there are many experimental conditions with very similar TFA patterns to our control condition (wild type glucose), and had no significantly changed TFAs in comparison to the control. These patterns vary across TFs and between experimental conditions. Understanding the pattern of how a TF changes its activities across experimental conditions can help biologists not only design experimental conditions to better control the activity of the TF of interest, but also identify rare activities of the TF of interest. For example, if an experimental study focuses on the cell response to certain metabolite, but also causes an environment change (e.g., intracellular pH change), the regulatory response to the environment change might not be of as much interest as the response to the presence of the metabolite. And knowing the regulatory network response to the environment changes could help eliminate these 'not so interesting' responses. Researchers may be more interested when they identify any significant changes in activity of TFs that are stable across most changes in MOPS media. Also, in future researches, significant TFA changes under the future experimental conditions which had no perturbation in this study should draw interest for further study.

TF	MOPS Experimental Condition	Number of Target Genes	Regulatory Effect	Signal Sensing
ArgP	Carbon source: acetate	7	Activator	
AtoC	Carbon source: proline	4	Activator	ETC
CaiF	Stationary	10	Activator	ETM
CsgD	Stationary	10	Dual Regulator	
CueR	Stationary 3	7	Dual Regulator	ETM
DpiA	Stationary 3	11	Dual Regulator	ETC
EvgA	Stationary 3	14	Activator	ETC
FruR	Stationary 3	67	Dual Regulator	ISM
Fur	Stationary 3	100	Dual Regulator	ETM
GlcC	Stationary 3	7	Dual Regulator	Hybrid
MhpR	Heat Shock	6	Activator	Hybrid
NanR	Heat Shock	8	Dual Regulator	ETM
NhaR	Knockout: crp	7	Activator	ETM
PspF	Knockout: crp	7	Dual Regulator	
RstA	Knockout: hns	10	Dual Regulator	ETC
SoxR	Knockout: hns	3	Dual Regulator	ISM
YdeO	Knockout: hns	5	Dual Regulator	
Zur	Knockout: hns	6	Repressor	

Table 3-3 *Condition Specific TFs*

TFs with TFA that are only significantly changed under one experimental condition. These TFs have specifically perturbed activities under only one of the tested conditions, which may suggest that these TFs perform special and critical roles under their respective conditions

Enrichment of TFAs under experimental conditions

Enrichment tests were used to find significantly changed activities of groups of TFs. The details of the enrichment test are in the methodology section.

To learn how regulatory effects of TFs related to environmental changes, TFs were grouped by their regulatory effects: 47 activators, 71 dual regulators, and 58 repressors. Where activators are defined as TFs that only activate or promote gene expression, repressors are TFs that only repress gene expression, and dual regulators can either activate or repress gene expression.

To understand the relationship between the signaling sensing mechanisms of TFs and the regulatory network behavior under certain condition changes, TFs were grouped by signal sensing mechanisms [10]. 57 of the TFs sense signals from exogenous or environmental stimuli (E), including 29 TFs which are part of two-component systems (ETC), and 28 TFs that sense transportable metabolites (ETM); 30 TFs sensing endogenous or intracellular stimuli (I), including 26 TFs binding/sensing metabolites generated by cellular metabolism (ISM), and 4 DNA-binding TFs for nucleoid or chromosome remodeling and compaction(IDB); and 33 TFs sensing metabolites from both endogenous and exogenous origin (Hybrid) (**Table 3-1**).

Glucose (Control)	Carbon source: acetate	Carbon source: glycerol	Carbon source: proline	Late Log phase	Stationary (135 min)	Stationary 2 (330 min)	Stationary 3 (480 min)	Stationary 4 (720 min)	Heat Shock	Acid Shock	Cipro	Cipro 2	Knockout: crp	Knockout: cspA	Knockout: dps	Knockout: dps stationary	Knockout: dps stationary 2	Knockout: hns	Knockout: hupB	
0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	1	0	0	0	Activator
0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	1	1	0	0	Dual Regulator
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Repressor
0	0	0	0	0	0	1	1	1	1	0	0	0	0	0	0	1	1	0	0	ETC
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	ETM
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	H
0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	1	1	0	0	ISM
0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	1	1	0	0	I
0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	1	1	0	0	E

Table 3-4 TFA Enrichment Tests of Regulatory Effects and Signal Sensing Mechanisms

"1" indicates there are significantly more TFs in the tested group activated under the tested experimental condition; "0" indicates no significant enrichment.

The enrichment test results of regulatory effect groups and signal sensing groups are shown in **Table 3-4**. The signal sensing group I-DB has only 4 TFs, and cannot give a meaningful enrichment test results. Thus, I-DB was not tested. The enrichment tests on the 23 selected pathways are tested using the target gene weighted enrichment test method described in the Method section, and enrichment test results are shown in **Table 3-5**.

It is noticeable in **Table 3-4** that repressors, hybrid signal sensing TFs, and external transportable metabolites sensing TFs have no enrichment across all the MOPS media conditions. It is highly possible that these groups of TFs have fairly stable regulatory mechanism under MOPS media conditions. For example, since the culture media of these experimental conditions are fairly similar, there should be little difference on the

external transportable metabolites among these experiments, thus, TFs sensing these type of signals should have little activity differences among these experiments.

The enrichment of pathway TFs shown in **Table 3-5** are biologically meaningful. For example, tRNA charging and process pathways TFs are negatively enriched in stationary phase. This result is reasonable because during stationary phase, the growth of cells is slowed down and there is less need of tRNAs as compared to the exponential phases [28, 29]. Gluconeogenesis I pathway is positively enriched in stationary phase, acetate carbon source or proline carbon source conditions. In these experimental conditions, glucose is not available, and gluconeogenesis pathways are turned on to produce enough intermediate metabolites of the glycolysis pathways.

The enrichment test of TFAs tests the activities of groups of TFs. The positive or negative enrichment of a group of TFs means that the enriched group of TFs has more active roles at the change of experimental conditions. But positive/negative enrichment does not necessarily lead to increased or decreased pathway activity. For example, a positive enrichment of a pathway TFs means there are significantly more TFs of this pathway activated in the experimental condition than the control condition. However, the majority of these activated TFs could be repressors and thus result in the decreased activity of the pathway.

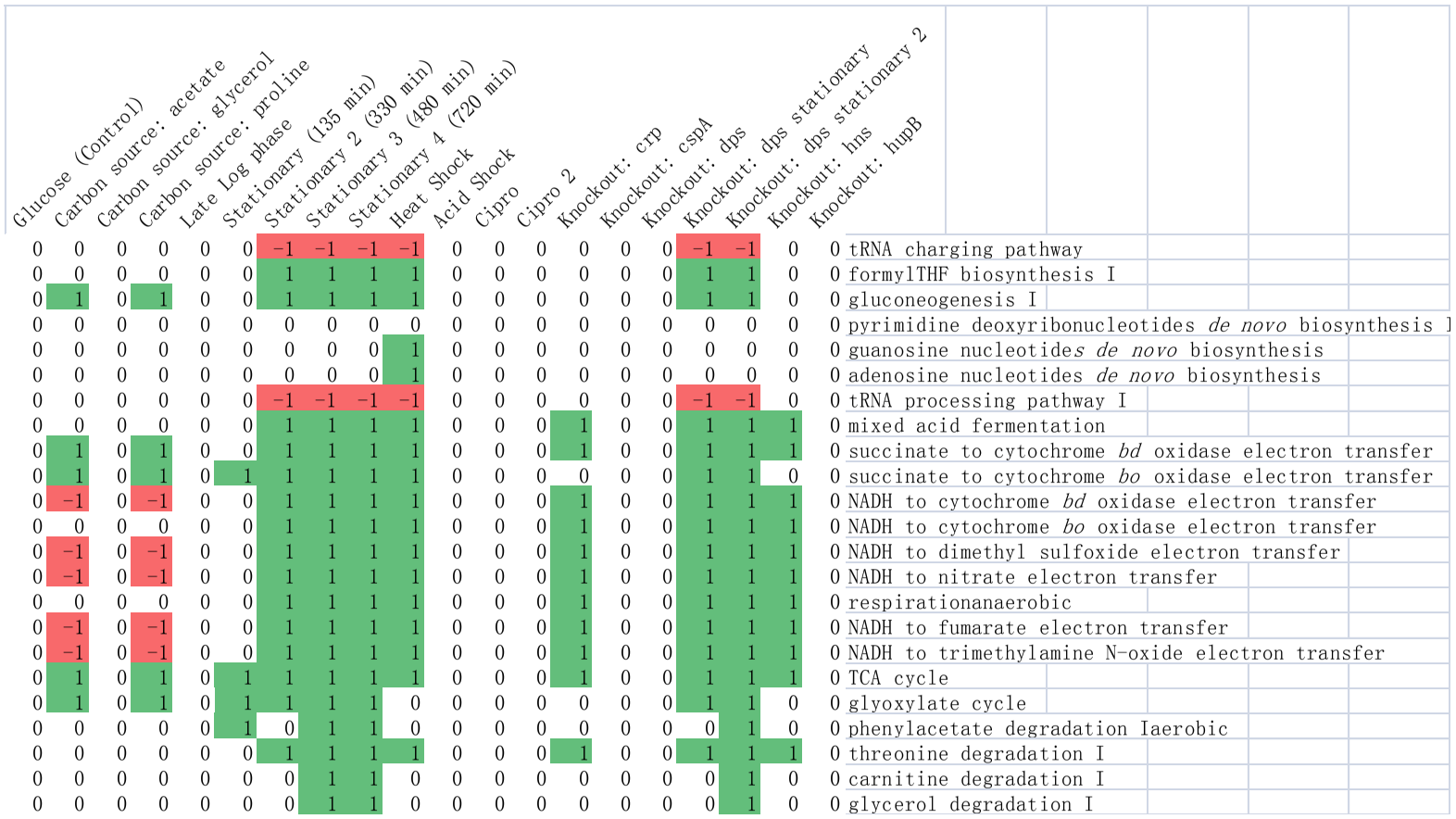


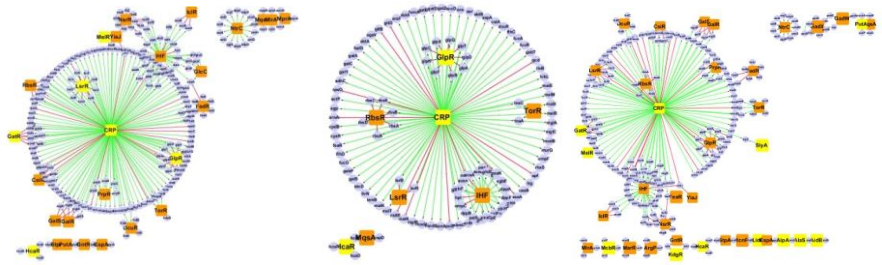
Table 3-5 *TFA Enrichment Tests of Pathway TFs*

"1" indicates there are significantly more TFs in the tested group activated under the tested experimental condition; "-1" indicates there are significantly less TFs in the tested group activated under the tested experimental condition; "0" indicates no significant enrichment.

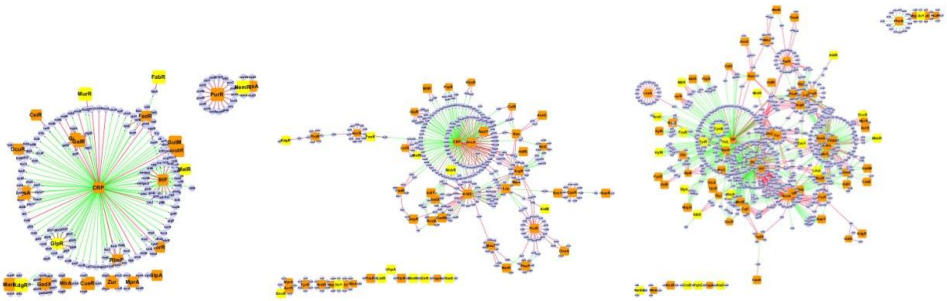
ERN and Key TFs for each experimental condition

An ERN is the most informative part of GRN under certain experimental condition changes, and could be used to compare the regulatory network changes between conditions. ERNs for experimental conditions compared to the control condition were generated (**Figure 3-4, APPENDIX C**). As shown in **Figure 3-4**, the ERNs differ significantly from condition to condition, and are characteristic networks which show regulatory network information specific to the differences between experimental conditions.

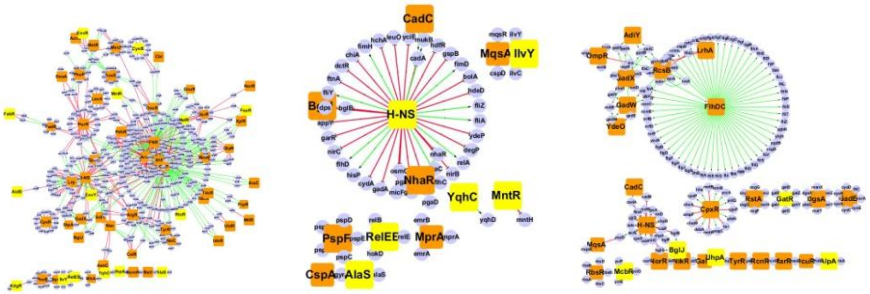
Network properties were calculated for each TF in each ERN to identify key TFs which contribute most to the ERN network topology differences between different conditions. These network properties values compare the number of genes each TF controls (degree out), the number of shortest regulatory paths through each TF, the number of steps needed for a TF to propagate a regulatory signal to the whole network (closeness), and the rate of effective regulations to all the possible regulations of each TF (network normalized degree) under condition specific ERNs



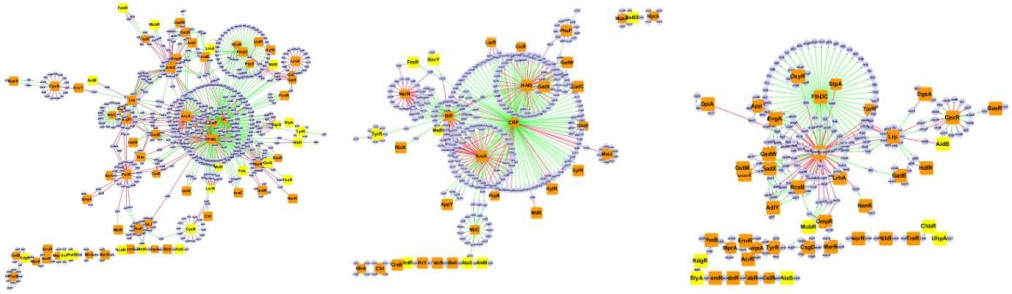
A. Wild Type Acetate **B.** Wild Type Glycerol **C.** Wild Type Proline



D. Wild Type Stationary **E.** Wild Type Stationary2 **F.** Wild Type Stationary3



G. Wild Type Stationary4 **H.** Wild Type Heat Shock **I.** *crp* Knockout



J. *dps* Knockout Stationary **K.** *dps* Knockout Stationary2 **L.** *hns* Knockout

Figure 3-4 *Effective regulatory networks (ERNs) of E. coli at condition change from wild type glucose (MOPS media)*

Figure 3-4 *continued*

Effective regulatory networks (ERNs) of E. coli at condition change from wild type glucose (MOPS media): ERNs are constructed to reflect the change of GRNs between the control experimental conditions and all other experimental conditions. Square nodes indicate TFs, smaller round nodes indicate genes regulated by TFs, green edges reflect the promotion relationship between TFs and their target genes, red edges reflect the repression relationship between TFs and their target genes. TFs in yellow are the key TFs to specific conditions (changing from the control condition). Figures in higher resolution can be found in ***APPENDIX C****.*

Key TFs, which have both significantly changed TFAs and significant regulatory effect on the whole regulatory network respond to a certain condition change, were identified based on network properties. Key TFs of an ERN are those TFs with one or more network properties described above reach certain threshold values. Thresholds of network properties to identify key TFs of an ERN are set to reveal a sensitivity level to identify TFs with top 10% of network properties in the ERN.122 key TFs of 12 experimental conditions are identified (**Table 3-6, Figure 3-4**). Each key TF has at least one of the network properties of ERN significant under a specific experimental condition change (**APPENDIX D**).

Key TFs are the key effective regulatory elements of response to specific condition differences. Under certain environmental difference, key TFs contribute the most to rewire the regulatory network to adapt the new environment at the gene regulation level, and should be the focus of the study of gene regulations and differences of gene regulations between the two conditions.

Experimental Condition	Key TFs
Carbon source: acetate	GatR, GlpR, LsrR, MelR, HcaR, CRP
Carbon source: glycerol	GlpR, HcaR, CRP
Carbon source: proline	AlpA, GatR, KdgR, LldR, McbR, MelR, SlyA, HcaR, AidB, AlaS, PutA, CRP
Stationary	GlpR, KdgR, MelR, FebR, MurR, NemR
Stationary 2	EnvR, FeaR, GlrR, KdgR, LldR, McbR, MelR, MntR, UhpA, AidB, AlaS, IlvY
Stationary 3	CreB, EnvR, EnvY, FeaR, McbR, MntR, RtcR, SlyA, YeiL, YqhC, AraC, CysB, HyfR, LrhA, NikR, TyrR, AidB, AlaS, IlvY, MtlR, RelEB
Stationary 4	EnvR, EnvY, FabR, FeaR, KdgR, MelR, MntR, RtcR, UhpA, YqhC, CysB, AidB, AlaS, IlvY, RelEB
Heat Shock	MntR, YqhC, AlaS, IlvY, RelEB, H-NS
Knockout: crp	AlpA, GatR, McbR, UhpA, BglJ
Knockout: dps stationary	FabR, FeaR, GlrR, KdgR, LldR, McbR, MelR, MntR, SlyA, YeiL, CysB, LrhA, NikR, TyrR, AidB, AlaS, GlpR, IlvY, LsrR, MtlR, RelEB
Knockout: dps stationary2	EnvY, FeaR, MelR, MntR, TyrR, AidB, AlaS, RelEB
Knockout: hns	ChbR, KdgR, McbR, SlyA, UhpA, AidB, AlaS

Table 3-6 *Key TFs*

Key TFs of a condition shown in this table are TFs have significantly changed activities from the control (WT_MOPS_glucose) condition with outstanding network properties of the ERN under the same condition change. Key TFs are the key effective regulatory elements of response to specific condition differences, and should be the focus of the study of gene regulations and differences of gene regulations between the two conditions.

The key TFs of each condition are biologically meaningful. For example, CRP is identified as a key TF under condition changes from control condition to experimental conditions with carbon sources other than glucose (Carbon source: acetate, Carbon source: glycerol, and Carbon source: proline) (**Figure 3-4A-C**). And it is known that CRP regulates many genes involved in secondary carbon source catabolism [12, 13, 30,

31]. According to EcoCyc [21], H-NS plays an important role in adaptation to environmental changes and stresses. And in this work, H-NS is identified as a key TF in response to heat shock (**Figure 3-4H**). EnvY, known for controlling genes encoding cellular envelope proteins at low temperature and during stationary phase, is identified as a key TF in many stationary phase experiments in this study (**Figure 3-4 D-G and J, K**).

Besides effectively regulating target genes to adapt the environmental changes, the more important contribution of key TFs are usually their effective regulation of the whole network. As the two major criteria of key TFs are the regulatory effectiveness to its target genes (based on ERNs), and its network properties to the global network, the key TFs have been found here should have more complex and deeper effective regulatory pathways than only one step depth regulation to its target genes. Some key TFs might not directly regulate genes which can be connected to the experimental conditions, but there should be some genes in the downstream of the regulatory pathway have biological function in response to the experimental conditions. Further study of how these key TFs contribute to the cell response to particular environment is highly recommended.

TFAs are more sensitive to growth phases than environment conditions

According to the results shown in **Table 3-4**, **Table 3-5** and **Figure 3-3, Figure 3-4**, TFAs are more sensitive to differences in growth phase rather than different gene perturbations or carbon sources under MOPS media conditions. It is noticeable in **Figure 3-3** that there are more TFs with significantly changed activities under stationary phase

experimental conditions. ERNs of the stationary phases of *dps* knockout look similar to ERNs of wild type stationary phases (**Figure 3-4 D-G and J, K**), and *dps* knockout at early exponential phase has the same TFA pattern as the control condition which is the wild type at early exponential phase (**Figure 3-3**). In **Table 3-4**, internal signal sensing TFs are only enriched under stationary phases. **Table 3-5** shows that most pathways are significantly enriched under the stationary phases too, comparing to the control condition which is sampled at the start of the exponential phase [3]. Also, it is noticeable in **Table 3-6** that there are more key TFs for conditions with growth phases different from the control condition. At stationary phases of different experimental environments, there are many same key TFs take controls of the regulatory networks.

This results are biological reasonable as at different growth phases, the internal and external environments, as well as the phenotype of cells are different, and the regulatory system has to change according to these differences to adapt the environments and transform the cell to survival.

Some Limitations

The analysis framework proposed in this work is highly based on the estimation of TFAs from gene expression experimental data and previously known GRN topology. Though the NCA based algorithm are widely used and proved with high accuracy of TFA estimation[14, 15, 18], the quality of analysis would be affected by the quality of experimental data and the completeness of the GRN topology. Also, by introducing the

less accurate SIMPLS TFA estimation algorithm to overcome the restrictions of NCA, the accuracy of the analysis could be further affected.

GadE is known as a major acid response regulator [32], but it is not identified significantly changed activity under Acid Shock condition. The reason could be the too strict multiple testing correction eliminated identification of GadE and other potential perturbed regulators. In other words, we cannot say GadE is perturbed in Acid Shock from the trancriptome data from M3D databse. Or the algorithm may have deficiencies on identifying such type of regulator. Not only GadE, but also many other acid resistance related regulators are not being identified significant change in activities. Experiment data with higher qualities might help to identify more significantly changed TFAs under Acid Shock condition. Also MOPS media has pH buffer, which might also be some reason of the insignificant response of acid response regulators under Acid Shock condition.

RpoH, also known as Sigma 32, is a major heat shock regulator [33]. However, sigma factors are not included in the analysis. Because sigma factors usually regulate a very large number of genes, and would affect the accuracy of TFA predictions of other TFs.

Conclusions

In this study, an improved TFA prediction method D-NCA is developed based on previous NCA and SIMPLS algorithms to make predictions of TFA log ratios

between two experimental conditions with more biological meaningful TFA changing directions. A regulatory network model focusing on capturing the network differences between experimental conditions, effective regulatory network (ERN), is defined to analyze GRNs. And a three level analysis framework to analyze TFAs and GRNs under different experimental conditions was developed. The first level, analyzing TFA patterns of individual TFs is shown to be helpful for biological research. The second level of analysis uses enrichment test and summarizes TFA behaviors by groups and their properties. The third level of analysis identifies key TFs of each experimental condition using network based analysis approach on ERNs. This analysis framework expands the traditional transcriptome data analysis to TFA and GRN level. The application to *E. coli* data showed the biological meaningfulness and helpfulness of analyzing transcriptome data on TFA and GRN level.

Authors' contributions

YF developed and implemented the analysis framework, drafted this manuscript. LRJ and JD developed the initial concept and suggested ways to improve the algorithm and testing methods. All authors read and approved the final version of the manuscript.

Acknowledgment

This material is based upon work supported by the National Science Foundation under Awards EEC-0813570 and IIS-0612240. Any opinions, findings, and conclusions or

recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

1. Lozada-Chavez, I., S.C. Janga, and J. Collado-Vides, *Bacterial regulatory networks are extremely flexible in evolution.* Nucleic Acids Res, 2006. **34**(12): p. 3434-45.

2. Balazsi, G. and Z.N. Oltvai, *Sensing your surroundings: how transcription-regulatory networks of the cell discern environmental signals.* Sci STKE, 2005. **2005**(282): p. pe20.

3. Faith, J.J., et al., *Many Microbe Microarrays Database: uniformly normalized Affymetrix compendia with structured experimental metadata.* Nucleic Acids Res, 2008. **36**(Database issue): p. D866-70.

4. Price, M.N., P.S. Dehal, and A.P. Arkin, *Orthologous transcription factors in bacteria have different functions and regulate different genes.* PLoS Comput Biol, 2007. **3**(9): p. 1739-50.

5. Salgado, H., et al., *RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more.* Nucleic Acids Res, 2013. **41**(Database issue): p. D203-13.

6. Gutierrez-Rios, R.M., et al., *Regulatory network of Escherichia coli: consistency between literature knowledge and microarray profiles.* Genome Res, 2003. **13**(11): p. 2435-43.

7. Luscombe, N.M., et al., *Genomic analysis of regulatory network dynamics reveals large topological changes.* Nature, 2004. **431**(7006): p. 308-12.

8. Ideker, T. and N.J. Krogan, *Differential network biology.* Mol Syst Biol, 2012. **8**: p. 565.

9. Janga, S.C., et al., *Coordination logic of the sensing machinery in the transcriptional regulatory network of Escherichia coli.* Nucleic Acids Res, 2007. **35**(20): p. 6963-72.

10. Martinez-Antonio, A., et al., *Internal-sensing machinery directs the activity of the regulatory network in Escherichia coli.* Trends Microbiol, 2006. **14**(1): p. 22-7.

11. Janga, S.C. and B. Contreras-Moreira, *Dissecting the expression patterns of transcription factors across conditions using an integrated network-based approach.* Nucleic Acids Res, 2010. **38**(20): p. 6841-56.

12. Gorke, B. and J. Stulke, *Carbon catabolite repression in bacteria: many ways to make the most out of nutrients.* Nat Rev Microbiol, 2008. **6**(8): p. 613-24.

13. Deutscher, J., *The mechanisms of carbon catabolite repression in bacteria.* Curr Opin Microbiol, 2008. **11**(2): p. 87-93.

14. Liao, J.C., et al., *Network component analysis: reconstruction of regulatory signals in biological systems.* Proc Natl Acad Sci U S A, 2003. **100**(26): p. 15522-7.

15. Boscolo, R., et al., *A generalized framework for network component analysis.* IEEE/ACM Trans Comput Biol Bioinform, 2005. **2**(4): p. 289-301.

16. Chang, C., et al., *Fast network component analysis (FastNCA) for gene regulatory network reconstruction from microarray data.* Bioinformatics, 2008. **24**(11): p. 1349-58.

17. Boulesteix, A.L. and K. Strimmer, *Predicting transcription factor activities from combined analysis of microarray and ChIP data: a partial least squares approach.* Theor Biol Med Model, 2005. **2**: p. 23.

18. Fu, Y., L.R. Jarboe, and J.A. Dickerson, *Reconstructing genome-wide regulatory network of E. coli using transcriptome data and predicted transcription factor activities.* BMC Bioinformatics, 2011. **12**: p. 233.

19. Jain, A.K., M.N. Murty, and P.J. Flynn, *Data clustering: a review.* ACM Comput. Surv., 1999. **31**(3): p. 264-323.

20. Gama-Castro, S., et al., *RegulonDB version 7.0: transcriptional regulation of Escherichia coli K-12 integrated within genetic sensory response units (Gensor Units).* Nucleic Acids Res, 2011. **39**(Database issue): p. D98-105.

21. Keseler, I.M., et al., *EcoCyc: a comprehensive database of Escherichia coli biology.* Nucleic Acids Res, 2011. **39**(Database issue): p. D583-90.

22. Goeman, J.J. and P. Buhlmann, *Analyzing gene expression data in terms of gene sets: methodological issues.* Bioinformatics, 2007. **23**(8): p. 980-7.

23. Green, M.L. and P.D. Karp, *Genome annotation errors in pathway databases due to semantic ambiguity in partial EC numbers.* Nucleic Acids Res, 2005. **33**(13): p. 4035-9.

24. Brandes, U., *A faster algorithm for betweenness centrality.* Journal of Mathematical Sociology, 2001. **25**(2): p. 163-177.

25. Neidhardt, F.C., P.L. Bloch, and D.F. Smith, *Culture medium for enterobacteria.* J Bacteriol, 1974. **119**(3): p. 736-47.

26. Hall, B.G. and L. Xu, *Nucleotide sequence, function, activation, and evolution of the cryptic asc operon of Escherichia coli K12.* Mol Biol Evol, 1992. **9**(4): p. 688-706.

27. Brombacher, E., et al., *The curli biosynthesis regulator CsgD co-ordinates the expression of both positive and negative determinants for biofilm formation in Escherichia coli.* Microbiology, 2003. **149**(Pt 10): p. 2847-57.

28. Dong, H., L. Nilsson, and C.G. Kurland, *Co-variation of tRNA abundance and codon usage in Escherichia coli at different growth rates.* J Mol Biol, 1996. **260**(5): p. 649-63.

29. Emilsson, V. and C.G. Kurland, *Growth rate dependence of transfer RNA abundance in Escherichia coli.* EMBO J, 1990. **9**(13): p. 4359-66.

30. Fic, E., et al., *cAMP receptor protein from escherichia coli as a model of signal transduction in proteins--a review.* J Mol Microbiol Biotechnol, 2009. **17**(1): p. 1-11.

31. Kolb, A., et al., *Transcriptional regulation by cAMP and its receptor protein.* Annu Rev Biochem, 1993. **62**: p. 749-95.

32. Tramonti, A., et al., *Stability and oligomerization of recombinant GadX, a transcriptional activator of the Escherichia coli glutamate decarboxylase system.* Biochim Biophys Acta, 2003. **1647**(1-2): p. 376-80.

33. Zhao, K., M. Liu, and R.R. Burgess, *The global transcriptional response of Escherichia coli to induced sigma 32 protein involves sigma 32 regulon activation followed by inactivation and degradation of sigma 32 in vivo.* J Biol Chem, 2005. **280**(18): p. 17758-68.

CHAPTER IV

WHOLE-GENOME AND WHOLE-CELL SCALE CONSTRUCTION OF GLOBAL REGULATORY NETWORK MODEL OF E. COL

A paper to be submitted to *BMC Systems Biology*

Yao Fu[1], Laura R Jarboe[2] and Julie Dickerson[1, 3§3]

Abstract

Background

Regulation occurs in most types of biological systems, including metabolic networks, protein interaction networks, transcriptional regulatory networks and etc. Regulatory signals virtually travel around the cell through a global regulatory network to all the cellular components where regulations are needed to keep cell adaptive to external environments changes and robust in internal environment. Though lots of work on

[1]Bioinformatics and Computational Biology Program, Iowa State University, Ames, Iowa, U.S.A

[2]Chemical and Biological Engineering Department, Iowa State University, Ames, Iowa, U.S.A

[3]Electrical and Computer Engineering Department, Iowa State University, Ames, Iowa, U.S.A

[§]Corresponding author

integrating multiple cellular systems and modeling regulatory signals have been done in recent years, still no step had been taken on a whole-genome and whole-cell scale modeling of regulation focused global system.

Results

In this work, an exploration on modeling a regulation focused whole-genome and whole-cell scale Global Regulatory Network of *E. coli* has been taken. Major interactions such as protein binding reactions, chemical reactions, enzymatic reactions, transport reactions, gene regulations and etc. have been converted into interactions reflecting regulatory relationships, and integrated into Global Regulatory Network for *E. coli*. The resulting network contains 10424 network elements, including gene products, protein complexes, enzymatic/transport reactions, and metabolites, all connected by 37411 regulatory interactions. Several network properties and number of feedback loops of resulting Global Regulatory Network has been compared with those of randomly connected networks statistically to test the significance of constructed network. Simulations of the regulatory signal response of Global Regulatory Network of *E. coli* to lactose stimulates have been performed to further verify the resulting regulatory system and simulation model.

Conclusions

Statistical tests and analysis shows that the Global Regulatory Network of *E. coli* constructed in this work has significantly special network properties comparing to random connected network. And these special properties are closely associated with the biological behavior of regulatory systems such stability and adaptation to environments.

Also the biological meaningfulness of the simulation method and Global Regulatory Network Model has been verified by a nearly perfect match between the simulation results and theoretical expectations. The results of this work suggest the feasibility and meaningfulness of modeling regulatory focused whole-genome and whole-cell signaling systems, and encourage further investigation on this direction.

Introduction

Regulations can be found in most types of biological systems and under various types of mechanisms such as chemical catalysis, chemical reactions, protein-protein interactions, protein-ligand interactions, molecule transpirations, gene regulations and etc. All these types of regulations together form a whole regulatory system transferring regulatory signals through all over the cell, and between all types of cellular components. Regulatory system as a global network controls the behavior of the cell as a central commander, and keeps cell internally stable enough to carry out all types of precise cellular functions, as well as flexible enough to adapt to all different kinds of external environmental changes and stimulates.

Better understanding and modeling of regulatory systems can help biologists to better learn and explain the complex behaviors and outcomes of cells, and understand the connections between all types of cellular components better. Also, biochemical engineers can be benefited from the knowledge and modeling of regulatory systems on predicting strain engineering outcomes and identify potential engineer targets.

Lots of efforts have been taken to study, summarize and model different types of regulatory systems, inducing metabolic reaction and flux networks, protein-protein and protein-ligand interaction systems and, gene regulatory systems, and etc.

Metabolic network models majorly model chemical reactions, including enzymatic reactions and transport reactions in the cell into a interaction network. By converting the metabolic reaction network into stoichiometric matrix, it is possible to find the solution space of all the feasible metabolic flux pathways using extreme pathways model[1]. Similarly, steady states based elementary mode analysis can identify a unique set of functioning smallest sub-networks that perform metabolic network functions[2]. Unlike extreme pathways and elementary mode, linear programming based flux balance analysis helps to find a optimized solution of the metabolic network model given certain interested objective function[3]. All these analysis are based on well defined metabolic networks, such as iAF1260, which was constructed by Feist *et al.* in 2007 for *E. coli* MG1655[4].

Protein can interact with proteins or other metabolites to form protein complexes in order to perform biological functions or just for inhibiting the protein its self. These binding interactions are usually identified by experiments or computational simulations and predictions[5]. Various databases collects these protein interactions to be used for knowledge references and construction of protein interactions networks[6-9]. The protein interaction networks are analyzed in various ways to learn the details of regulatory motifs[10] and construct regulatory systems[11]. Yeger-Lotem *et al.*

integrated protein-protein interaction network with transcription-regulation network and analyzed the possible network motifs with less than four nodes [12].

Gene regulatory network, also known as transcriptional regulation network, contains regulatory interactions between transcription factor and genes, sigma-factor and genes, as well as sRNA and genes, though many researches only focus on the most complex subset, transcription factor - gene regulatory networks. These gene regulatory interactions can be identified experimentally [13] or computationally predicted computationally from DNA sequence information or trancriptome data [14, 15]. And all regulatory information of *E. coli* are collected in RegulonDB database[16]. Coupled with transcriptome experiment data, gene regulatory networks are used to predict the activities of transcription factors [17, 18] and analysis gene regulatory network behaviors[19].

Integrating all these types of systems discussed above and other related interactions is becoming a hot topic in system biology area in recent years. However, large genome-wide integrations of all the systems in the cell are still limited in knowledge based integrations, such as Reactome[20] and Ecocyc [21], or only have several selected systems being integrated for to perform specific analysis and modeling as mentioned above.

Efforts on modeling regulatory signals across different layers of interaction networks have also been taken. For example, the two-component signal transduction systems are being well studied[22]. And RegulonDB recently proposed and constructed 25 Genetic sensory-response units model of *E. coli* encountering signal, the signal-to-effect reaction

end with activation/deactivation of TF, the regulatory swathes, and the consequences to model signals cross multiple interaction layers[16]. However, these regulatory signaling models only cover one specific type of signaling system or small un-connected parts of the whole cell system.

So far, there are still no large scale integration of the three major types of regulatory systems (metabolic network, protein interaction network, and gene regulatory network) and other related interactions into a wholly connected network model. The major challenge is all different types of interactions have their own biological mechanisms behind, and are presented differently in their own network models, which are hard to be integrated directly.

In this work, an exploratory step have been taken to integrate metabolic network, protein interaction network, gene regulatory network and other related interactions of *E. coli* and build a fully connected Global Regulatory Network (GlRN) with focusing on regulatory signaling across all the components within the cell. In GlRN, all the interactions and reactions are converted into regulatory interactions only reflecting regulations between network elements. And four major types of network elements have been defined:

Protein Complex: Proteins which are binding with other proteins or ligands to form protein complex. Could be regulators to regulate other protein, genes, as well as enzymes or transporters to catalyze reactions or molecule transportations.

Metabolite: Small Molecules which are the major compounds of chemical reactions in the cell, may also perform regulate functions to other reactions or proteins.

Gene Product: The direct gene product from gene transcriptions and translation. Could be protein monomers, single proteins or small RNAs. Some protein monomers can directly

performing biological functions such as regulation, catalysis, and transportation. And small RNAs are majorly regulators to other genes.

Enzymatic/Transport Reaction: Reactions depends enzymes and catalyzing enzymatic reactions, or transport reactions rely on transporter proteins to transport molecule into or out of cell, can be regulated by enzyme/transporter proteins and other cofactors.

The four types of network elements interact with each other and form complex regulatory interaction system (**Figure 3-1**) through following types of interactions.

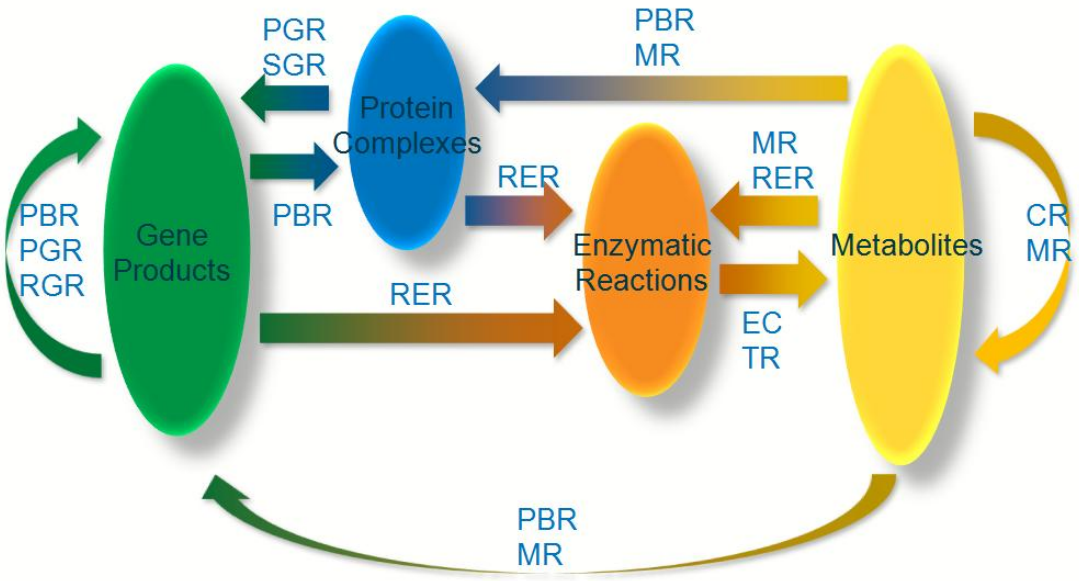


Figure 4-1 *Interactions between Elements of Global Regulatory Networks*

Protein Binding Regulation(PBR): Regulatory interactions from protein binding reactions, could be protein-protein binding or protein-ligand binding. Regulations in the form of components positively regulate the protein complex, and repress other components.

Chemical Reaction (CR): Nature chemical reactions without need or enzyme catalysis. Regulations in the form of reactants promote products but repress other reactants.

Transport Regulation (TR): Reaction to transport molecules in or out of the cell. Usually need transport proteins, and regulation in the form of protein promoting molecules if in-taking and repressing molecules if transporting out.

Catalysis of Enzymatic Reaction (CER): Enzyme catalyzed reactions. Regulation in the form of enzyme reaction promoting products and repressing source

Regulation of Enzymatic Reaction (RER): Regulatory interaction to regulate enzymatic reactions. Regulators are enzyme proteins or co-factors, and targets are enzymatic reactions

Molecule Regulation (MR): Small molecule may also regulate enzymatic reactions or chemical reactions.

Protein Gene Regulation (PGR): also known as TF-Gene regulation

Sigma-factor Gene Regulation (SGR): Gene regulatory interactions regulated by Sigma-factors

sRNA Gene Regulation(RGR): sRNA-Gene regulatory interactions.

Analysis on several network properties and feedback loops of the resulting GlRN of *E. coli* have been performed to learn the properties of GlRN and better understands relationships between biological behaviors and GlRN of *E. coli*. Also, simple simulations on regulatory signal transduction through the GlRN of *E. coli* are

implemented and carried out to demonstrate biological meaningfulness and applications of GlRN.

Methods

Integration of multiple networks and interactions

Chemical Reactions

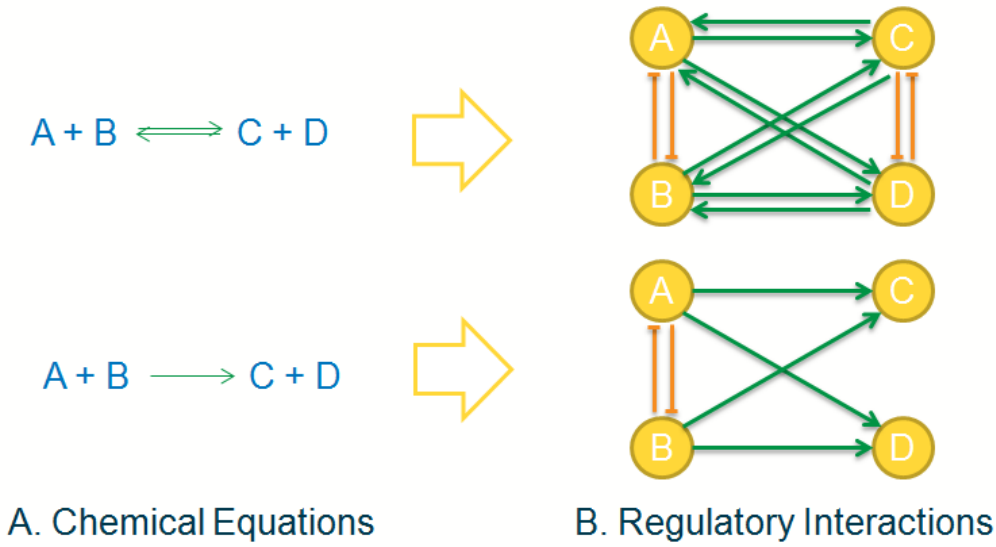


Figure 4-2 *Convert Chemical Equations into Regulatory Interactions*

Chemical reactions are usually represented in the forms of chemical equation as shown in **Figure 4-2A**, where reactants are at one side of the equation and products are at the other side of equation, and the two sides of equation are connected by one or two one-directional arrow to reflect the reaction directions. However, this form of representation

is not sufficient to reflect all the regulatory relationship between compounds of a chemical reaction. For example with increase of the amount of one reactants, the reaction would be promoted towards producing more products and consuming more reactants, in this case, from a regulatory point of view, the increased reactant would promote all the products and repress all other reactants. To reflect these regulatory relationships of a reaction, a regulatory focused representation of chemical reactions is proposed as shown in **Figure 4-2B**.

In the regulatory focused representation of chemical reactions, all the reactants are repressing each other and promoting all the products. And for a reversible reaction, products would also be able to repress each other and promote reactants.

All the chemical reactions are extracted, converted and integrated from Ecocyc database v16.5 [21].

Enzymatic Reactions

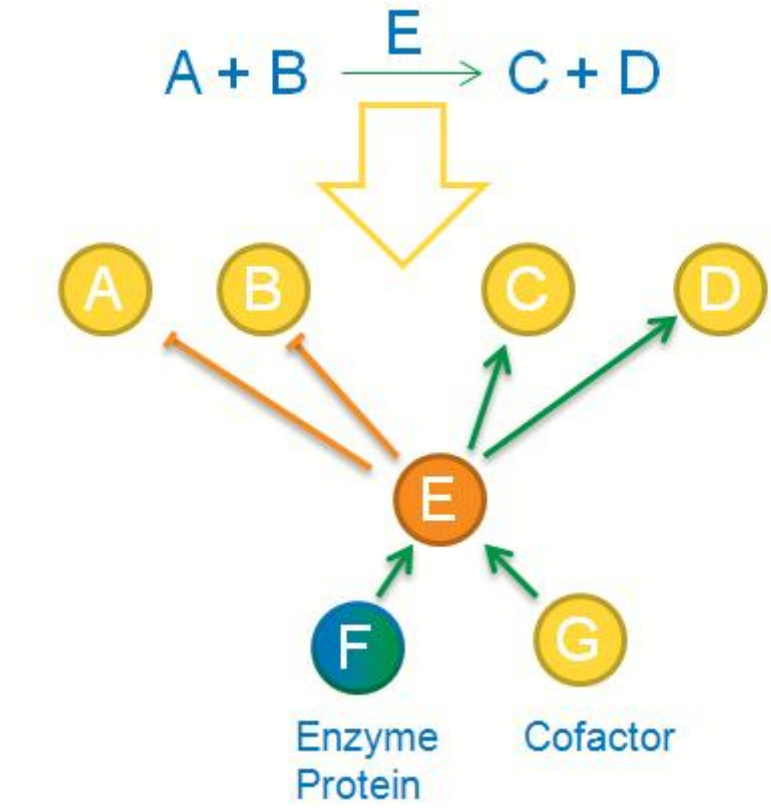


Figure 4-3 *Convert Enzymatic Catalysis into Regulatory Interactions*

Similar to chemical reactions, enzymatic reactions are also converted into regulatory focused representations from chemical equation representation to reflect the regulatory relationships between compounds. Additionally, a virtual element type Enzymatic Regulator is also defined in the integrated network to reflect the catalysis of Enzymatic Reactions to their compounds **(Figure 4-3)**.

Enzymatic Regulator element in the integrated network promotes products of the enzymatic reaction and represses the reactants of the enzymatic reaction. Also,

enzymatic reactions are positively regulated by its respective enzyme protein and co-factors.

Enzymatic reactions are extracted, converted and integrated from Ecocyc database v16.5 [21].

Transport Reactions

Transport reactions transporting molecules into and out of the cell, in a regulatory point of view, regulate the intracellular concentration/amount of molecules. This type of regulatory interactions are integrated into Global Regulatory Network as virtual element Transport Regulator up-regulates its target molecules if it is an in-taking transportation and down-regulates its target molecules if it is a secretion transportation. Similar to Enzymatic Reactions, Transport Reactions are regulated by transporter proteins and some co-factors.

Transport Reactions are extracted, converted and integrated from Ecocyc database v16.5 [21].

Metabolite Regulations

Many metabolites in the cell are performing regulatory functions to some chemical reactions and enzymatic reactions. This regulatory relationships are collected in Ecocyc database v16.5 [21] and is used to in this work to construct Global Regulatory Network of *E. coli*.

For metabolite directly regulating enzymatic reactions or transport reactions, a regulatory connection between the metabolite and its target enzymatic reaction or transport reaction is constructed to reflect this regulation. For metabolites regulating chemical reactions, the regulatory relationships are constructed as following: For promotional regulation, regulator metabolites repress reactants and promote products of target reactions; for repression regulation, regulator metabolites promote reactants and repress products of target reactions.

Protein Interactions

Protein interactions include protein-protein interactions and protein-ligand interactions. The interactions are in the form of binding reaction and usually involve two or more proteins, or protein and metabolites combinations binding into a larger protein complex. To convert these binding reactions in a regulatory focused format, component proteins and ligands repress each other as they are consuming each other to form protein complexes. And at the same time, component proteins and ligands also promote the protein complexes they are forming (**Figure 4-4**).

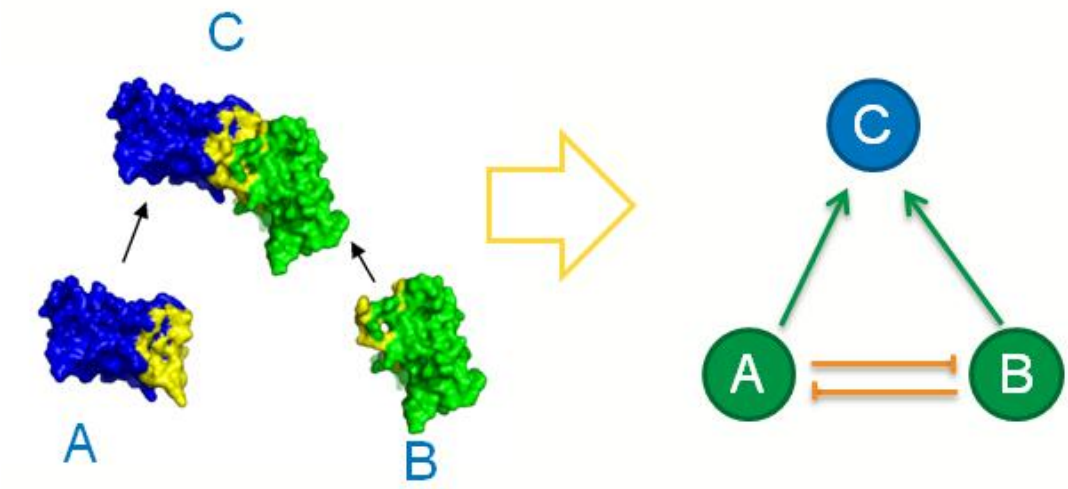


Figure 4-4 *Convert Protein Binding Interactions into Regulatory Interactions*

Protein interactions are collected, converted, and integrated from Protein Complexes collection and Protein-ligand complexes collection of Ecocyc v16.5 [21].

Gene Regulations

Gene regulations including transcription factor gene interactions, sigma-factor gene interactions and sRNA gene interactions are integrated into the Global Regulatory Network. All the gene regulation information is obtained from RegulonDB 8.1[16].

In RegulonDB 8.1, many transcription factors has multiple confirmations, and only certain confirmation perform regulatory function to genes. Only these regulatory functioning confirmations have been constructed regulatory links to their target genes, and integrated into Global Regulatory Network.

Analysis of Global Regulatory Network

Network Properties

Following basic network properties are calculated for understanding some of the basic properties of the Global Regulatory Network.

Degree:

The number of connections each element of the network has. The average degree reflects how dense the Global Regulatory Network is connected, and the degrees of each individual element can help on indentifying highly connected hubs of the network.

Betweenness:

The number of shortest regulatory paths between two elements passing through a certain elements. This property further indicate the importance of a element in term of being a common path of multiple regulatory signaling pathways.

Closeness:

The inverse of the sum of shortest distances to all other elements in the network (the shortest distance to an unreachable element is assumed to be the number of all the elements minus one). This property reflects the speed of a regulatory signal reaching from or spreading to all the elements of the network. For Global Regulatory Network, which is a directed graph, two types of closeness can be calculated: Upstream closeness measures the speed of any regulatory signal from the network reaches to the measured element; downstream closeness measures the speed of regulatory signal spreading to all

over the network. High closeness elements could be either a high efficiency global signal sensor or regulator or both.

Feedback Loops

Loops of a directed graph provide feedback features to a network. In Global Regulatory Network, feedback loops help to build a stable and robust regulatory system. Positive feedbacks which gives positive regulatory signal back to the element itself can keep the regulatory signal continuously at a certain level after an one time stimulate. Negative feedback loop can make sure the regulatory signal back to a normal level after an one time signal stimulate or keep the regulatory signal level stable for a continuous signal stimulate. The number of positive and negative feedback loops with certain length (containing certain number of elements in the loop) passing each element is calculated as following.

Algorithm:

Input: Directed graph of network, specified loop length *L*.

Output: number of loops at length *L* each element involves in

1. Construct an adjacency matrix *M* of the Global Regulatory Network, where if there is a regulation from element *i* to element *j*, let *M*(*i*,*j)* be 1 if it is a positive regulation and *M(i,i)* be -1 if it is a negative regulation.

2. The number of self feedback loops are reflected in the diagonal of matrix *M*, 1 indicate positive feedback and -1 indicate negative feedback respectively to the elements.

3. Construct an non-directional adjacency matrix *N* of the Global Regulatory Network, where if there is a regulation from element *i* to element *j*, let *N*(*i*,*j*) be 1.

4. Copy matrix *M* and *N* to matrix *M1* and *N1*, set diagonals of *M1* and *N1* 0s.

5. Let $p = 2$

5.1 $O = N1 \times N(p-1)$

5.2 $O+ = [N1 \times N(p-1) + M1 \times M(p-1)]/2$

5.3 $O- = [N1 \times N(p-1) - M1 \times M(p-1)]/2$

5.4 Let $q = 2$

5.4.1 $O = \min(O, Nq \times N(p-q))$

5.4.2 $O+ = \min(O+, [Nq \times N(p-q) + Mq \times M(p-q)]/2)$

5.4.3 $O- = \min(O-, [Nq \times N(p-q) - Mq \times M(p-q)]/2)$

Where min(*A*,*B*) for matrix *A* and *B* is to construct a matrix with all the elements are the smaller respective elements chosen from *A* or *B*.

5.5 Repeat step 5.4 till $q = p$

5.6 $Np = O$

5.7 $Mp = O+ - O-$

6. Repeat step 5 till $p = L$

7. The diagonal of *O*+ from the last recursion of step 5 indicates the number of positive feedback loops at length *L* of each element respectively; the diagonal of *O*- from the last recursion of step 5 indicates the number of negative feedback loops at length *L* of each element respectively.

Random Networks

Randomly generated networks with the same network elements and total degrees as Global Regulatory Network but connected randomly can be used to test the significance of the regulatory functions by comparing network properties and feedback loops with Global Regulatory Network. The randomly connected networks are generated by randomly shuffling the entries of the adjacency matrix of Global Regulatory Network.

Simulate Regulatory Signals across the Network

Predict Regulatory Effects

Regulatory signal transductions across the Global Regulatory Network can be simulated step wisely. Given a regulatory signal input or multiple regulatory signals stimulates to elements of the network, the responses of perturbed elements' target elements could be predicted and form a second step regulatory signal perturbations. Similarly, these signal perturbations at second step can be transferred to further steps (**Figure 4-5**).

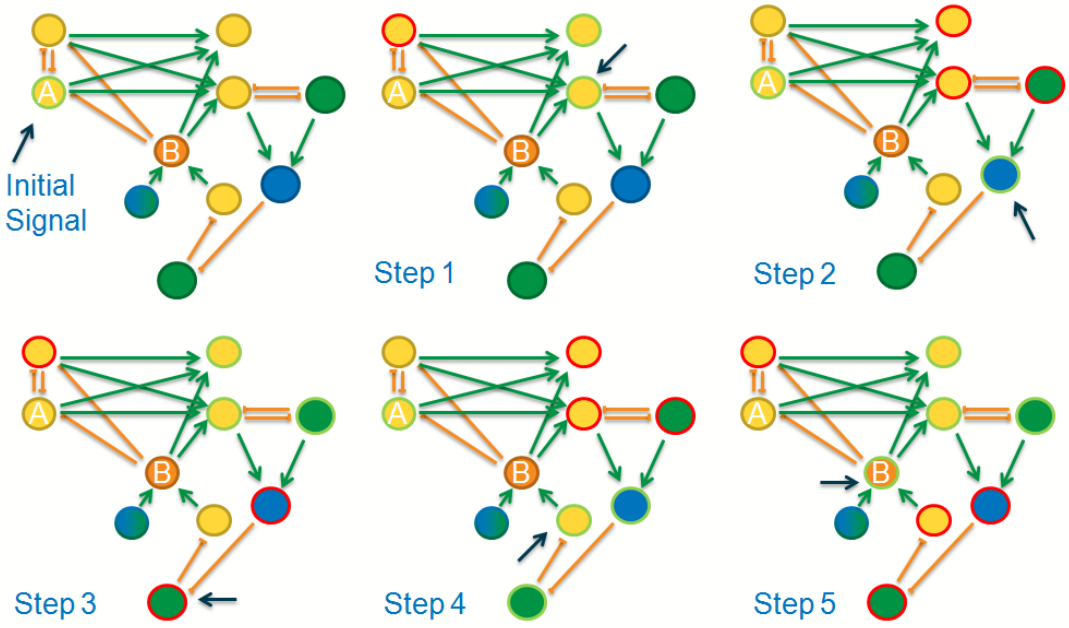


Figure 4-5 *Step-wise Signal Transduction Example*

Signal Transduction Rate

The signal transduction rate for regulations from direct molecular interactions such as chemical reactions, protein binding reactions, metabolite regulations, and regulation of enzymatic or transport reactions is considered linear and transferred without amplifying the total signal strength, thus assumed to be the inverse of the total number of regulatory targets of its source elements. The signal transduction rate for enzymatic or metabolite catalysis and gene regulations are considered able to transfer signals without immediate decrement of signal strengths and defined as two times of the inverse of the total number of regulatory targets of its source elements.

Signal Transduction Model Assumption

The signal perturbation of an element at each step is estimated as the signal perturbation of its regulator elements at last step multiply the signal transduction rates of respective

regulatory interactions. If an element responds to multiple signals from last step, the signal perturbation is assumed linear combined as the sum of responses from all its source signals.

Signal Transduction Model

While signal strength is used to quantify signal perturbations of each step, signal strength of each element *e S(e)* is estimated as following:

S(e) = sum(SL(i) x STR(i,e))

Where *i* is the regulator elements of *e, SL(i)* is the signal strength of element *i* at last step, and *STR(i,e)* is the signal transduction rate from *i* to *e.*

Mask Environmental Noise

While simulating regulatory signal transduction of cells under certain environments, e.g. certain growth media for microbes, the high abundant environment molecules such as water and many ions might affect the regulatory signal transductions. For example, while regulatory signals affects water molecules, the regulatory signal of water should be changed. But this signal of water will actually be disappear in the background as the amount of water molecules in the environment is much higher than the signal change, and thus, signal of water molecules would not be able to further transport to following signaling pathways. Taking this into consideration, while simulating regulatory signals under certain environments, molecules present in the environment with high abundance should be masked by removing all the regulatory connections from these molecule elements.

Results

Global Regulatory Network (GlRN) of *E. coli*

Global Regulatory Network of *E. coli* (**Figure 4-6**) is constructed by integrating multiple types of interactions, including regulatory relationships converted from chemical reactions, enzymatic reactions, transport reactions, protein binding reactions from Ecocyc 16.5[21], and gene regulatory networks including transcription factor gene regulation, sigma-factor gene regulation and sRNA gene regulation from RegulonDB 8.1 [16]. The result Global Regulatory Network *E. coli* contains total of 10424 elements and 37411 interactions. More detailed statistics of Global Regulatory Network *E. coli* is shown in **Table 4-1**.

Element	Number	Interaction	Number	Interaction	Number
Gene Product	3692	Protein Binding Regulation	8205	Metabolite Regulation	5858
Protein Complex	1220	Chemical Reaction Regulation	12185	Protein-Gene Regulation	4175
Metabolite	3484	Transport Regulation	789	Sigma-factor-Gene Regulation	4062
Enzymatic/Transport Reaction	2028	Catalysis of Enzymatic Reaction	5936	sRNA-Gene Regulation	223
Total	**10424**	Regulation of Enzymatic Reaction	2756	**Total**	**37411**

Table 4-1 *Summary of Global Regulatory Network of E. coli*

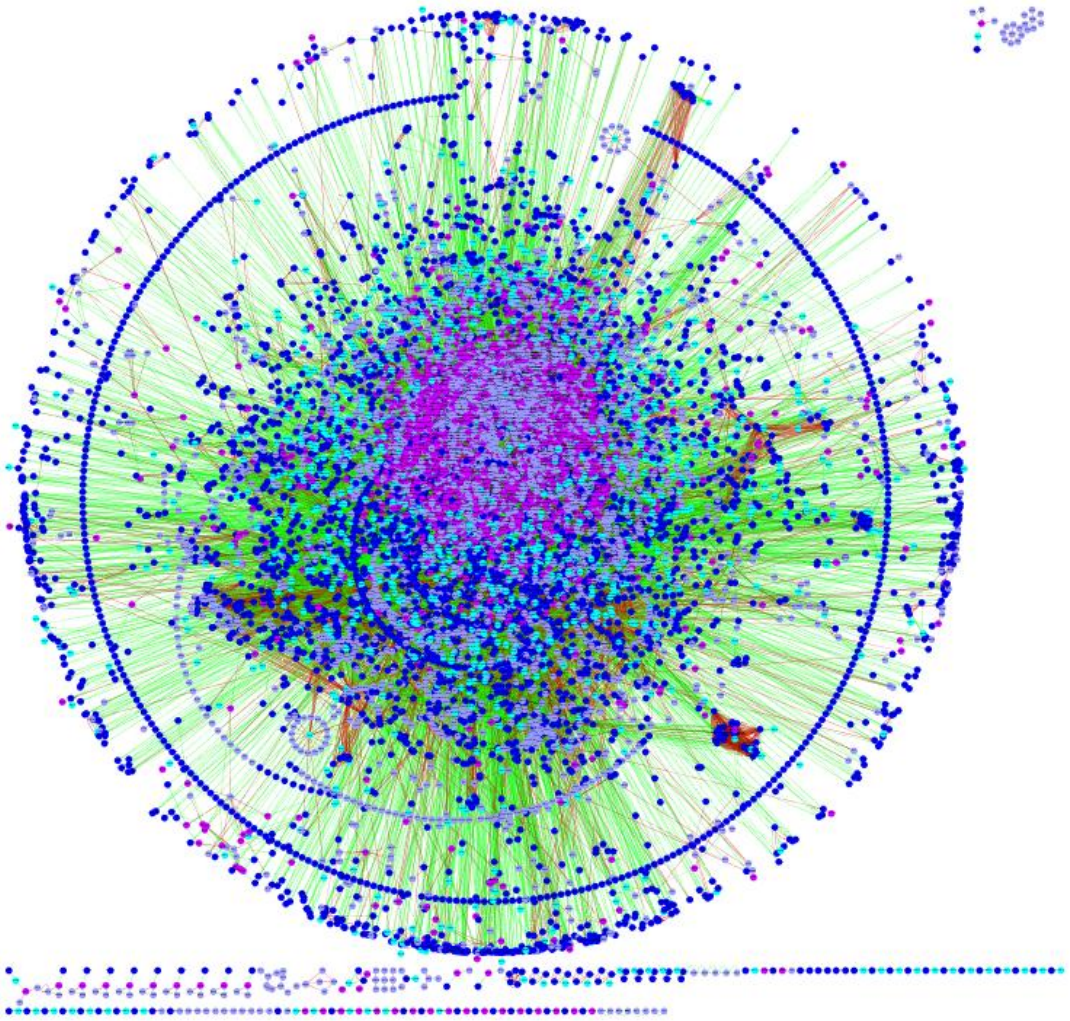


Figure 4-6 *Global Regulatory Network of E. coli*

Properties of Global Regulatory Network of *E. coli*

Network properties and number of feedback loops of Global Regulatory Network elements are calculated. Also, these properties and feedback loops for 5 randomly generated networks with the same number of elements and total number of connections as Global Regulatory Network of *E. coli* are calculated to be compared with these properties of Global Regulatory Network of *E. coli.* Two sample t-tests between these

		Mean	Max	Median	p-value*
Degree	GlRN of *E. coli*	7.17	2273	3	0.99
	Random Networks	7.18	1453	6	
Betweenness	GlRN of *E. coli*	22,739	20,143,726	9	<0.0001
	Random Networks	61,787	11,046,639	24,549	
Downstream Closeness	GlRN of *E. coli*	0.000350	0.000566	0.000558	<0.0001
	Random Networks	0.000529	0.000572	0.000544	
Upstream Closeness	GlRN of *E. coli*	0.000193	0.000214	0.000213	
	Random Networks	0.00239	0.00324	0.00287	<0.0001

Table 4-2 *Network Properties of Global Regulatory Network of E. coli*
*p-value is testing the significance of GlRN different from random connected networks

prosperities of Global Regulatory Network and random network are used to test the network significance of Global Regulatory Network. Summary of network properties and feedback loops are collected **Table 4-2** and **Table 4-3**, detailed properties and feedback loops information for each element is collected in **Supplemental File 1**.

As shown in **Table 4-2**, the degree of GlRN of *E. coli* is not significantly different from the 5 randomly generated networks. This result is expectable because the random networks are generated by randomly connect the elements of GlRN of *E. coli* without changing the total number of connections, and result in not significantly changed average degrees. However, other network properties such as betweenness, and upstream/downstream closeness of GlRN of *E. coli* are significantly different from those of random connected networks (**Table 4-2**). The numbers of both positive feedback loops and negative feedback loops of GlRN of *E. coli* are also significantly different from those of random networks (**Table 4-3**). All these significantly differed properties

and numbers confirms that GlRN of *E. coli* is a specially organized network significantly different from randomly connected networks and could perform special biological functions.

Number of Elements Loop Contains	Positive Feedback Loops			Negative Feedback Loops		
	GlRN of *E. coli*	Random Networks	p-value*	GlRN of *E. coli*	Random Networks	p-value*
1	77	1	<0.0001	47	4	<0.0001
2	9604	6	<0.0001	710	36	<0.0001
3	18881	21	<0.0001	53447	121	<0.0001
4	2341668	87	<0.0001	893460	408	<0.0001
5	30214477	261	<0.0001	45322283	1244	<0.0001
6	1929503838	841	<0.0001	1313894738	5106	<0.0001
7	45541607209	3070	<0.0001	547164631337	18512	<0.0001
8	2.15118E+12	11182	<0.0001	1.82788E+12	68584	<0.0001
9	6.52124E+13	39950	<0.0001	7.11568E+13	247171	<0.0001
10	2.69949E+15	144904	<0.0001	2.51258E+15	875176	<0.0001
11	9.13796E+16	516852	<0.0001	9.53023E+16	3193755	<0.0001
12	3.56606E+18	1866081	<0.0001	3.4517E+18	11492808	<0.0001

Table 4-3 *Sum of Feedback Loops of Global Regulatory Network of E. coli*
*p-value is testing the significance of GlRN different from random connected networks

Degree

Degree of an element measures the total number of connections an element has. The more degrees an element has, the more regulations this element involves in.

It is noticeable in **Table 4-4** that some protein complexes and metabolites could have very high degrees, and involves in a very high number of regulations. This is because protein complexes contains some global gene regulators regulates thousands of genes, such like Sigma-factors and transcription factor, and metabolites could include some common essential molecules such as water and protons (**Table 4-5**).

		Degree	Betweenness	Downstream Closeness	Upstream Closeness
Enzymatic Reactions	Mean	5.779093	11906.05	0.000451	0.000196
	Max	38	346592.7	0.000559	0.000214
	Median	5	8637	0.000558	0.000213
Gene Products	Mean	5.158722	14112.81	0.000305	0.0002
	Max	268	11193843	0.000562	0.000214
	Median	3	0	9.6E-05	0.000213
Protein Complexes	Mean	9.845902	53132.23	0.000375	0.000192
	Max	2273	10265724	0.000566	0.000214
	Median	2	8637	0.000558	0.000213
Metabolites	Mean	9.183123	27542.87	0.00033	0.000184
	Max	2033	20143726	0.00056	0.000214
	Median	3	0	0.000558	0.000213

Table 4-4 *Network Properties for Different Types of Elements*

Element	Name	Degree
ProteinComplex	RNA polymerase sigma 70	2273
Metabolite	H+	2033
Metabolite	H2O	1700
Metabolite	ATP	1062
ProteinComplex	CRP-cAMP DNA-binding transcriptional dual regulator	725
ProteinComplex	RNA polymerase sigma 24	552
Metabolite	|Pi|	485
ProteinComplex	RNA polymerase sigma 32	453
Metabolite	ADP	443
ProteinComplex	FNR DNA-binding transcriptional dual regulator	409
Metabolite	diphosphate	358
Metabolite	NAD+	356
Metabolite	NADPH	340
Metabolite	NADH	328
Metabolite	*S*-adenosyl-L-methionine	323
ProteinComplex	RNA polymerase sigma 38	321
ProteinComplex	Fis DNA-binding transcriptional dual regulator	320
ProteinComplex	IHF DNA-binding transcriptional dual regulator	300

Table 4-5 *High Degree Elements of Global Regulatory Network of E. coli*

Betweenness

Betweenness summarizes the total number of shortest regulatory pathways between two elements paths through a measured element. Different from degrees, higher betweenness indicates more involvement in regulatory signaling pathways of the regulatory system, and more importance in bridging the network.

Most Protein Complexes and Enzymatic Reactions are at least bridging several thousands of regulatory signaling pathways (non-zero medians in **Table 4-4**) as their major functions in regulatory signaling network is to bridge between gene products and

metabolites (**Figure 4-1**). And the elements at the two "ends' of **Figure 4-1**, gene products and metabolites, though have some connections other elements, usually have zero betweennesses (**Table 4-4**).

However, the elements with highest betweenness are some essential metabolites such as ATP, water and protons, and gene regulator proteins or components of gene regulator proteins such as transcription factors, Sigma-factors, and RNA polymerase components (**Table 4-6**). And it is noticeable that ArcAB Two-Component Signal transduction System and RcsCDB Two-Component Signal Transduction System play important role on connecting regulatory systems (**Table 4-6**).

Element	Name	PWY-Name	Bewteenness
Metabolite	ATP	143 pathways	20143725.57
Metabolite	H2O	170 pathways	16696092.99
Metabolite	H+	238 pathways	12703497.79
GeneProduct	rpoH	N/A	11193842.94
ProteinComplex	DnaA-ATP transcriptional dual regulator	N/A	10265723.86
ProteinComplex	RNA polymerase sigma 70	N/A	7358462.268
ProteinComplex	CRP-cAMP DNA-binding transcriptional dual regulator	N/A	6089442.276
Metabolite	cyclic-AMP	N/A	6086386.047
GeneProduct	rpoA	N/A	3473100.009
GeneProduct	rpoC	N/A	3472659.87
GeneProduct	rpoB	N/A	3472659.87
ProteinComplex	ArcA-Phosphorylated DNA-binding transcriptional dual regulator	(1);ArcAB Two-Component Signal Transduction System, quinone dependent	3132258.562
ProteinComplex	RNA polymerase sigma 32	N/A	2994030.488
Metabolite	|Pi|	69 pathways	2989062.48
ProteinComplex	RNA polymerase sigma 24	N/A	2245125.888
GeneProduct	rpoS	N/A	2224010.107
Metabolite	nitric oxide	N/A	1926414.851
ProteinComplex	NsrR DNA-binding transcriptional repressor	N/A	1785867.882
GeneProduct	arcB	(1);ArcAB Two-Component Signal Transduction System, quinone dependent	1575707.569
ProteinComplex	RcsB-phosphorylated DNA-binding transcriptional activator	(1);RcsCDB Two-Component Signal Transduction System	1312834.685
ProteinComplex	ArcB sensory histidine kinase - his717 phosphorylated	(1);ArcAB Two-Component Signal Transduction System, quinone dependent	1199268.723

Table 4-6 *High Betweenness Elements of Global Regulatory Network of E. coli*

Downstream Closeness

Downstream closeness measures the ability of signal from the measured element reach to the whole network. The higher closeness, the faster a signal from the element spreads to the whole network.

Similar to betweenness, most enzymatic reactions and protein complexes have relatively higher downstream closeness (**Table 4-4**). The reason should still be that the main role of these two types of elements is to serve as connections between the other two types of elements, gene products and metabolites.

But according to **Table 4-7**, most of the highest downstream closeness elements, whose signals tends to affect the whole regulatory network very quickly are gene products performing various functions in different pathways, though transcription factor CpxR is the fastest one sending signal to the whole network. It is also noticeable that regulatory systems of *E. coli* gives very quick responses to many gene products in fatty acid related pathways and glutathione ABC transporters, and etc (**Table 4-7**).

Element	Name	Comment	Closeness
ProteinComplex	CpxR-Phosphorylated	CpxR-Phosphorylated	0.000566
Gene Product	fabI	superpathway of unsaturated fatty acids biosynthesis and other 7 pathways	0.000562
Gene Product	cdaR	CdaR DNA-binding transcriptional activator	0.000561
Gene Product	nadR	NadR DNA-binding transcriptional repressor and NMN adenylyltransferase	0.000561
Protein Complex	NadR DNA-binding transcriptional repressor and NMN adenylyltransferase	NadR DNA-binding transcriptional repressor and NMN adenylyltransferase	0.000561
Gene Product	fabG	superpathway of unsaturated fatty acids biosynthesis and other 9 pathways	0.00056
Gene Product	zupT	heavy metal divalent cation transporter ZupT	0.00056
Protein Complex	RcsB-Pasp56	RcsB-P^{asp56}	0.00056
Gene Product	rihC	ribonucleoside hydrolase 3	0.00056
Gene Product	thiG	thiazole synthase	0.00056
Gene Product	thiH	tyrosine lyase	0.00056
Gene Product	gsiA	glutathione ABC transporter - ATP binding subunit	0.00056
Gene Product	gsiB	glutathione ABC transporter - periplasmic binding protein	0.00056
Gene Product	gsiD	glutathione ABC transporter - membrane subunit	0.00056
Gene Product	gsiC	glutathione ABC transporter - membrane subunit	0.00056
Gene Product	tesA	superpathway of unsaturated fatty acids biosynthesis and other 2 pathways	0.00056
Protein Complex	multifunctional acyl-CoA thioesterase I and protease I and lysophospholipase L1	multifunctional acyl-CoA thioesterase I and protease I and lysophospholipase L1	0.00056
Metabolite	Ca2+		0.00056

Table 4-7 *High Downstream Closeness Elements of Global Regulatory Network of E. coli*

Upstream Closeness

Upstream Closeness measures the ability of an element to receive regulatory signals from all over the network. The higher upstream closeness indicates higher sensitivity to network signals.

According to **Table 4-4**, the upstream closeness are relatively identical across all types of elements , which means most of elements in the GlRN of *E. coli* have the similar sensitivities to network stimulates.

Feedback Loops

Feedback loops bring special self regulating features of elements, these feedbacks are important for elements to control and balance the regulatory signal of its self and to keep the network stable and robust to most signal stimulates.

Table 4-4 shows that Protein Complexes and Enzymatic Reactions are averagely more stable in terms of number of feedback loops. This is also matching their biological roles, as most of them are severing as important regulators of the system; they have to be stable and robust to minor changes in order to keep a stable internal environment of the cell.

Many of ion, important molecule related protein complexes, transporters, enzymatic reactions and even metabolites have very high numbers of feedback loops to keep these internal environment ions and key metabolites balanced and stable (**Table 4-8**).

Element	Name	Comment
ProteinComplex	SufBC2D Fe-S cluster scaffold complex	SufBC2D Fe-S cluster scaffold complex
GeneProduct	csgE	curli transport specificity factor
EnzymaticReaction	glycogen phosphorylase	GLYCOPHOSPHORYL-CPLX
GeneProduct	osmF	YehW/YehX/YehY/YehZ ABC transporter
Metabolite	|N-formyl-L-methionyl-tRNAfmet|	
GeneProduct	yehX	YehW/YehX/YehY/YehZ ABC transporter
Metabolite	|tRNA-pseudouridine55|	
ProteinComplex	glutamate dehydrogenase	glutamate dehydrogenase
Metabolite	ferrichrome	

Table 4-8 *Elements with high number of feedback loops*

Simulate Regulatory Signal across the Network

As motioned in Methods part, many high abundant environment molecules such as water and oxygen in aerobic environments could block regulatory signal transductions by absorbing signals of these molecules into background. Thus, the regulatory effects of these environment molecules should be removed from GlRN accordingly to simulate regulatory signal of GlRN of *E. coli* under realistic environment condition. Regulatory masks for a commonly used aerobic experiment medium for *E. coli*, M9 is constructed to remove regulatory effects of molecules in these medium. Compounds being removed for M9 medium are listed below:

Water, Oxygen, Na^+, phosphate, chloride, K^+, ammonium, sulfate and Mg^{2+}

Simulations of regulatory signals across the GlRN of *E. coli* under following scenarios are performed to verify the biological meaningfulness of GlRN and demonstrate applications for GlRN of *E. coli*.

Lactose operon system

Regulatory system of lactose and *lac* operon are well studied[23]. As shown in **Figure 4-7**, lactose regulatory system involves all types of elements defined in a GlRN, including metabolites, protein complexes, gene products, and enzymatic/transport reactions. Also, most kinds of regulation relationships including PBR, CR, ECR, ER, TR, MR and PGR are included in the lactose system. Thus, lactose system is an ideal testing system for GlRN of *E. coli*.

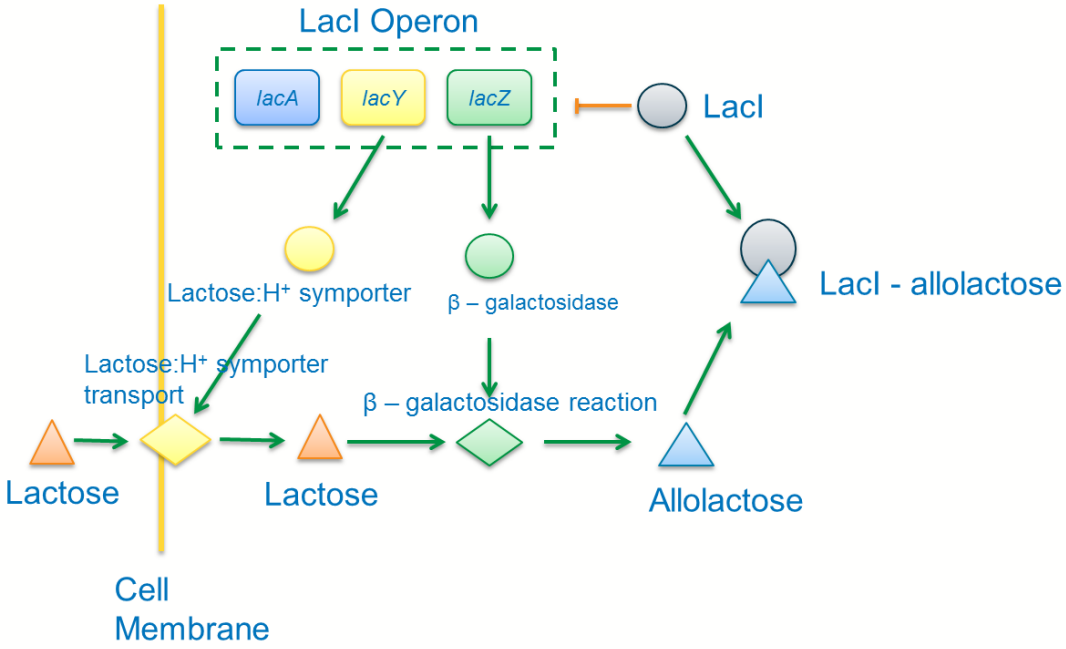


Figure 4- 7 *Lactose Operon Regulation System*

The GlRN response with present of lactose (continuously increased lactose signal) under M9 base medium is simulated across the GlRN of *E. coli*. The regulatory signal of allolactose, the product of beta-galactosidase, is shown in **Figure 4-8** as an example of the signal responses to continuous signal stimulate of lactose. The signal of allolactose is initially tossing around 0.394 and finally stabilized at 0.394 (**Figure 4-8**). This pattern is also explainable and reasonable in biology. As while regulatory signal is spreading to the whole network at the beginning, the regulatory signals are not stable as more and more regulations staring to affect the signals step after steps, while the signals getting to all its reachable elements, and all the effective regulations are contributing their regulations to signals stably, the regulatory signals get stabilized.

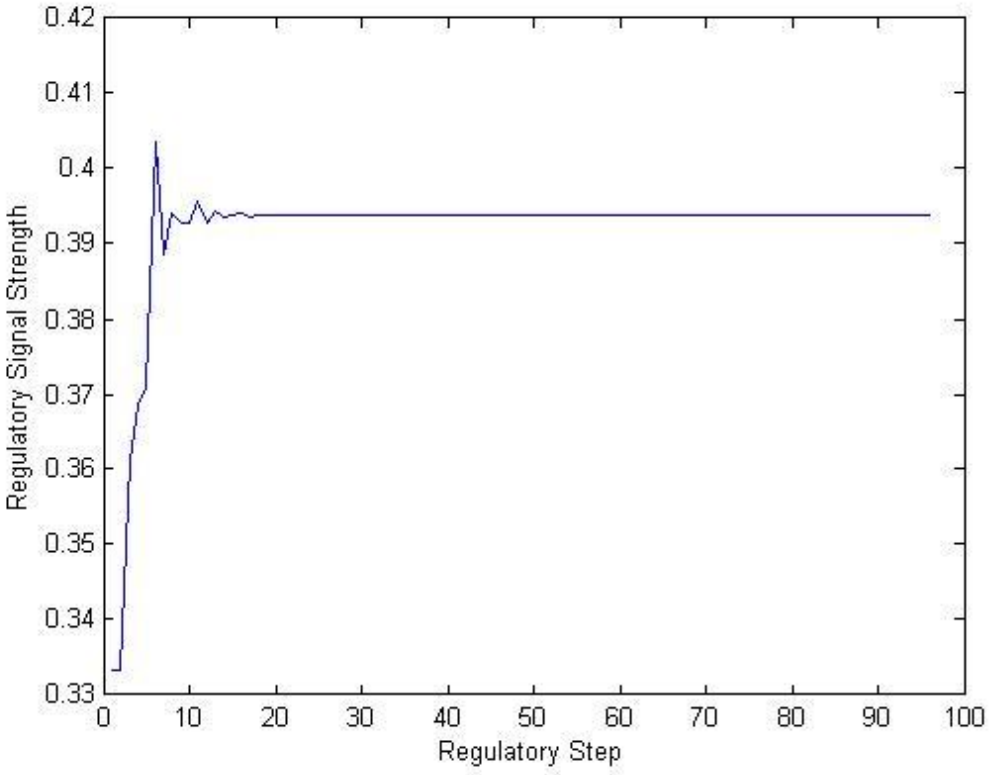


Figure 4-8 *Regulatory Signal of Allolactose in response of Lactose Stimulate*

Table 4-9 collects elements with high stable signal values with presents of both lactose and glucose under M9 medium. Significance p-value of the stabilized signal strength of an element is calculated by comparing the its value of this experiment with 4000 simulated values of the same element under different randomly generated signal inputs using t-test and Bonferroni multiple testing correction. According to **Figure 4-8**, 11 elements theoretically should be up or down regulated. And 9 of them are matched in the top 12 signal significantly perturbed elements in **Table 4-9**.

p-value	Stabled Signal	Respond in Step	Element	Common Name	PWY-Name
<0.0001	0.3430	1	Metabolite	beta-D-galactose	(2);galactose degradation I (Leloir pathway);lactose degradation III
<0.0001	0.3938	1	Metabolite	allolactose	N/A
<0.0001	0.0446	1	Enzymatic Reaction	glycogen phosphorylase	(1);glycogen degradation I
<0.0001	-0.1969	2	Gene Product	lacI	N/A
<0.0001	0.1641	2	Protein Complex	LacI-allolactose	N/A
<0.0001	0.0656	3	Gene Product	lacY	N/A
<0.0001	0.0656	3	Gene Product	lacA	N/A
<0.0001	0.0656	3	Gene Product	lacZ	(1);lactose degradation III
<0.0001	0.0328	4	Enzymatic Reaction	melibiose:H+ symporter	N/A
<0.0001	0.0328	4	Enzymatic Reaction	lactose:H + symporter LacY	N/A
<0.0001	0.0656	4	Protein Complex	beta-galactosidase	N/A
<0.0001	0.0328	5	Enzymatic Reaction	beta-galactosidase	N/A

Table 4-9 *Lactose stimulate response elements with highest significances*

Discussion

Life Essential Molecules

By comparing network properties of GlRN of *E. coli* with those of random connected networks, it is not hard to find that GlRN is specially organized to reveal biological features. For example, a very low median and mean of betweenness of GlRN but extremely high (**Table 4-3**) maximum infers very strong bridges molecules exists in the network, which should be essential to GlRN and hence essential to life. Water is one of these high betweenness and life essential molecules.

Extremely High Number of Feedbacks

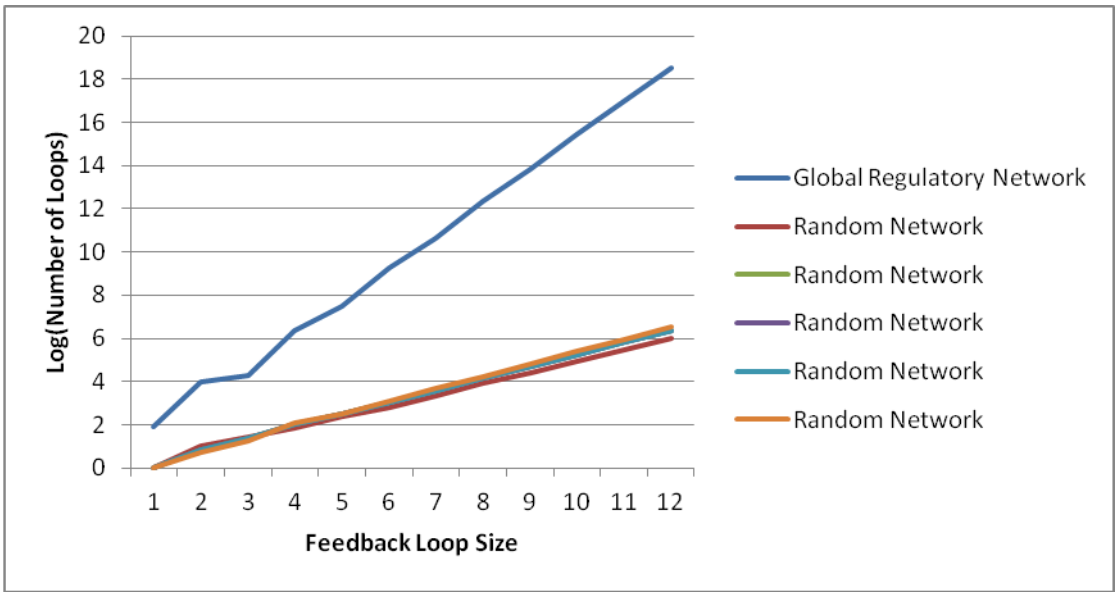


Figure 4-9 *Comparison of Feedback Loops between Global Regulatory Network and Random Network*

By take logarithm of the number of feedback loops involve different numbers of elements, **Figure 4-9** shows linear relationships between log feedback loop numbers and

the number of elements involves in a feedback loop, which means the number of feedback loops of elements growths exponentially along the loop size. And compare with the log feedback loop numbers of random connected networks, the log feedback loop numbers not only significantly greater but also grow much faster than that of random connected networks **Figure 4-9**. Besides the betweenness of life essential molecules, the most different feature between GlRN and random connected networks is the much higher in magnitude number of feedback loops. As feedback loops could bring lots of different functionalities such as amplification, stabilization, and adaptation. This extremely high number of feedback loops of GlRN should be the key of regulatory system's stability and flexibility, e.g. driving a quickly responding and adapting signal response of stimulate as shown in **Figure 4-8**.

Biological Meaningfulness

The nearly perfect matches between theoretical expectations and simulation results of the test on lactose operon system proved the biological meaningfulness of GlRN model and simulation methods proposed in this paper, and suggest further and more precise instigation on this network modeling and integration direction.

Reverse the Simulation

By reversing the simulation direction, regulators of elements can be identified from the Global Regulatory Network. This method could be used to find potential engineering target to achieve certain biological outcome. Regulator efficiency value for each element is proposed to quantitatively indicate the potential efficiency of manipulating this element to achieve certain regulatory outcome. The regulatory efficiency of each

element e for a given outcome *O*, which is a collection of elements with expected perturbations, is calculated as following steps:

1. Similar to predicting regulatory effects, reverse signal transduction rate (rSTR) between two connected elements are defined as:
    a. for a direct molecular interaction for rSTR is the inverse of the total number of direct molecular interactions to its target element
    b. for an indirect catalysis, transportation, or gene regulation, rSTR is either 1 for positive regulation or -1 for negative regulation
2. For each element *r* in *O*, the regulatory efficiency of its regulator

    RE(r) = sum(O(o) x rSTR(r,o))

    Where *o* is the element in *O* and *O(o)* is a predefined outcome regulatory signal level, *rSTR(r,o)* is the reverse signal transduction rate of regulatory interaction from r to o.
3. For each element *e*， *RE(e) = sum(RE(t) x rSTR(e,t))*

Conclusions

A whole-genome and whole-scale comprehensive regulatory system model for *E. coli* have been built in this work by integrating regulatory systems including transcriptional regulatory networks, protein interaction networks, metabolic reaction networks and other related regulations. Statistical tests on network properties revealed statistical significance of this network. And these special network properties of the constructed GlRN have been shown connections with biological properties. Regulatory signaling model of the constructed GlRN was defined to simulate regulatory signals in response of the changes

in the environment and perturbations within the cell. This model was tested by simulating regulatory signal response of presenting lactose as the environmental carbon source, the simulation results matches the theoretical outcomes, and verified the biological meaningfulness of this GlRN model.

Authors' contributions

YF developed and implemented the modeling and simulation methods and algorithms, drafted this manuscript. LRJ and JD suggested ways to improve the algorithm and testing methods. All authors read and approved the final version of the manuscript.

Acknowledgment

This material is based upon work supported by the National Science Foundation under Awards EEC-0813570 and IIS-0612240. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

1. Price ND, Reed JL, Papin JA, Wiback SJ, Palsson BO: **Network-based analysis of metabolic regulation in the human red blood cell**. *Journal of theoretical biology* 2003, **225**(2):185-194.

2. Papin JA, Stelling J, Price ND, Klamt S, Schuster S, Palsson BO: **Comparison of network-based pathway analysis methods**. *Trends in biotechnology* 2004, **22**(8):400-405.

3. Stelling J, Klamt S, Bettenbrock K, Schuster S, Gilles ED: **Metabolic network structure determines key aspects of functionality and regulation**. *Nature* 2002, **420**(6912):190-193.

4. Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, Karp PD, Broadbelt LJ, Hatzimanikatis V, Palsson BO: **A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information**. *Molecular systems biology* 2007, **3**:121.

5. Xia JF, Wang SL, Lei YK: **Computational methods for the prediction of protein-protein interactions**. *Protein and peptide letters* 2010, **17**(9):1069-1078.

6. Licata L, Briganti L, Peluso D, Perfetto L, Iannuccelli M, Galeota E, Sacco F, Palma A, Nardozza AP, Santonico E *et al*: **MINT, the molecular interaction database: 2012 update**. *Nucleic acids research* 2012, **40**(Database issue):D857-861.

7. Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A, Lin J, Minguez P, Bork P, von Mering C *et al*: **STRING v9.1: protein-protein interaction networks, with increased coverage and integration**. *Nucleic acids research* 2013, **41**(Database issue):D808-815.

8. Butland G, Peregrin-Alvarez JM, Li J, Yang W, Yang X, Canadien V, Starostine A, Richards D, Beattie B, Krogan N *et al*: **Interaction network containing conserved and essential protein complexes in Escherichia coli**. *Nature* 2005, **433**(7025):531-537.

9. Xenarios I, Salwinski L, Duan XJ, Higney P, Kim SM, Eisenberg D: **DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions**. *Nucleic acids research* 2002, **30**(1):303-305.

10. Ciriello G, Guerra C: **A review on models and algorithms for motif discovery in protein-protein interaction networks**. *Briefings in functional genomics & proteomics* 2008, **7**(2):147-156.

11. Pawson T, Nash P: **Assembly of cell regulatory systems through protein interaction domains**. *Science* 2003, **300**(5618):445-452.

12. Yeger-Lotem E, Sattath S, Kashtan N, Itzkovitz S, Milo R, Pinter RY, Alon U, Margalit H: **Network motifs in integrated cellular networks of transcription-regulation and protein-protein interaction**. *Proceedings of the National Academy of Sciences of the United States of America* 2004, **101**(16):5934-5939.

13. Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E *et al*: **Genome-wide location and function of DNA binding proteins**. *Science* 2000, **290**(5500):2306-2309.

14. Hecker M, Lambeck S, Toepfer S, van Someren E, Guthke R: **Gene regulatory network inference: data integration in dynamic models-a review**. *Bio Systems* 2009, **96**(1):86-103.

15. Fu Y, Jarboe LR, Dickerson JA: **Reconstructing genome-wide regulatory network of E. coli using transcriptome data and predicted transcription factor activities**. *BMC bioinformatics* 2011, **12**:233.

16. Salgado H, Peralta-Gil M, Gama-Castro S, Santos-Zavaleta A, Muniz-Rascado L, Garcia-Sotelo JS, Weiss V, Solano-Lira H, Martinez-Flores I, Medina-Rivera A *et al*: **RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more**. *Nucleic acids research* 2013, **41**(Database issue):D203-213.

17. Liao JC, Boscolo R, Yang YL, Tran LM, Sabatti C, Roychowdhury VP: **Network component analysis: reconstruction of regulatory signals in biological systems**. *Proceedings of the National Academy of Sciences of the United States of America* 2003, **100**(26):15522-15527.

18. Chang C, Ding Z, Hung YS, Fung PC: **Fast network component analysis (FastNCA) for gene regulatory network reconstruction from microarray data**. *Bioinformatics* 2008, **24**(11):1349-1358.

19. Janga SC, Contreras-Moreira B: **Dissecting the expression patterns of transcription factors across conditions using an integrated network-based approach**. *Nucleic acids research* 2010, **38**(20):6841-6856.

20. Jupe S, Akkerman JW, Soranzo N, Ouwehand WH: **Reactome - a curated knowledgebase of biological pathways: megakaryocytes and platelets**. *Journal of thrombosis and haemostasis : JTH* 2012.

21. Keseler IM, Mackie A, Peralta-Gil M, Santos-Zavaleta A, Gama-Castro S, Bonavides-Martinez C, Fulcher C, Huerta AM, Kothari A, Krummenacker M *et al*: **EcoCyc: fusing model organism databases with systems biology**. *Nucleic acids research* 2013, **41**(Database issue):D605-612.

22. Capra EJ, Laub MT: **Evolution of two-component signal transduction systems**. *Annual review of microbiology* 2012, **66**:325-347.

23. Griffiths AJF GW, Miller JH, et al.: **Regulation of the Lactose System**. In: *Modern Genetic Analysis.* New York: W. H. Freeman; 1999.

CHAPTER V

GENTERAL CONCLUSIONS

General Conclusions

In general, this thesis introduced a series of methods, frameworks and models to computationally reconstruct, analyze and model biological regulatory systems. These proposed work have been shown both statistically significant and biologically meaningful.

Reconstruction

Transcriptional networks are essential for regulatory systems. Though many works had been done to reconstruct transcriptional networks as described above, there are always space for improvement to obtain better and more comprehensive transcriptional networks. Chapter 1 focused on developing methods to refine and reconstruct transcriptional networks using currently known network knowledge and transcriptomics data.

More specifically, an algorithm to reconstruct gene regulatory (transcriptional) networks using transcriptome data and predicted TFAs have been developed and applied on *E. coli* data to reconstruct a genome-wide transcriptional network.

The proposed **G**ene expression and **T**ranscription factor activity based **R**elevance **N**etwork (GTRNetwork) [51] is a novel gene regulatory network reconstruction algorithm. It introduces a hidden layer of TFAs into relevance score based network reconstruction algorithms. Instead of using gene expression level as the only input to detect relationships between TFs and genes, GTRNetwork uses the relevancies between TFs and genes estimated based on the TFA and gene expression ratios. Different combinations of TFA prediction algorithms and relevance score functions have been applied to find the most efficient combination. A comparison between GTRNetwork and CLR, a standard network reconstruction algorithm shows significant improvement in precision and recall using GTRNetwork. When the integrated GTRNetwork method was applied to *E. coli* data, the reconstructed genome-wide gene regulatory network predicted 381 new regulatory links. This reconstructed transcriptional network including the predicted new regulatory links show promising biological significances. Many of the new links are verified by known TF binding site information, and many other links can be verified from the literature and databases such as EcoCyc. The reconstructed gene regulatory network is applied to a recent transcriptome analysis of E. coli during isobutanol stress. In addition to the 16 significantly changed TFAs detected in the original paper, another 7 significantly changed TFAs have been detected by using our reconstructed network.

Analysis

Regulatory networks are dynamically changing according to the surrounding and internal environment of the cell. Thus, in order to study and understand how regulatory systems response and adapt to environmental changes, it is important to analysis these networks in dynamics under different internal and external environment factors. For example, Janga *et al*. performed a systematic analysis of the expression patterns of TFs of *E. coli* across all of the 466 experimental conditions from M3D database [52]. Janga *et al*. clustered the experimental conditions of M3D database to remove bias from redundant experiments, defined activated TFs based on the expression of the TF encoding genes, performed enrichment tests on different groups of TFs, including different regulatory effects groups and different signal sensing groups [53], and finally identified some marker TFs for experimental conditions using network based analysis methods. However, as discussed above the expression of TF-encoding genes does not ensure the activation and successful regulation of target genes. Analyses that model TFA solely based on TF gene expression will not fully reflect the behaviors of GRNs.

Therefore, analysis of TFAs using computational predicted TFAs by many TFA prediction methods such as NCA [54] under different experimental conditions could give more direct insight of the behavior of GRNs. But current TFA prediction methods reach their limitation that they are not able to predict biological meaningful changing directions of TFA quantity between experimental conditions, such as NCA and PLS base algorithms. Although NCA package allow researchers to determine the TFA changing directions for

each TFA manually, an automated algorithm to determine the changing directions of TFAs are needed for further investigation and analysis on the dynamics of TFAs under different environmental conditions. In this work, directed network component analysis (D-NCA), a direction-corrected TFA prediction algorithm developed from the original NCA, is proposed. D-NCA can correct the TFA changing directions from NCA by comparing the predicted TFAs with reference gene regulatory interactions and the original transcriptome data. D-NCA algorithm also fills the eliminated TFs from NCA by SIMPLS predictions. Thus, D-NCA could predict a set of comprehensive and biologically meaningful TFAs for further analysis.

To study the behaviors of GNRs of *E. coli* under different experimental conditions, instead of analyzing gene expressions of TF encoding genes, as performed by Janga and Contreras-Moreira [52], in this thesis a systematic three level analysis is demonstrated on TFAs predicted from M3D and RegulonDB *E. coli* data using D-NCA. The first level of analysis is on individual TFs and analyzes the behavior of each TF across different experimental conditions. The second level of analysis groups TFs by different types of TF properties, e.g. regulatory effects, signal sensing machineries or pathways they regulate. Then, enrichment tests for each group of TFs were performed for different experimental conditions. The final level is the network level. In this level of analysis, a novel Effective Regulatory Network (ERN) model is proposed to capture the dynamic of GRN changes between experimental conditions. This level of analysis is based on ERNs and identifies the key TFs for experimental conditions which significantly change network properties, efficiently rewire GRN, and successfully regulate target genes. The analysis results are explainable by biological knowledge and biological meaningful. Some of the results also suggest further biological study targets on regulations of *E. coli* in response of environmental conditions changes.

Integration

So far, there are still no large scale integration of the three major types of regulatory systems (metabolic network, protein interaction network, and gene regulatory network) and other related interactions into a wholly connected network model. The major challenge is all different types of interactions have their own biological mechanisms behind, and are presented differently in their own network models, which are hard to be integrated directly.

In this thesis, an exploratory step have been taken to integrate metabolic network, protein interaction network, gene regulatory network and other related interactions of E. *coli* and build a fully connected Global Regulatory Network (GlRN) with focusing on regulatory signaling across all the components within the cell. In GlRN, all the interactions and reactions are converted into regulatory interactions only reflecting regulations between network elements. The resulting network contains 10424 network elements, including gene products, protein complexes, enzymatic/transport reactions, and metabolites, all connected by 37411 regulatory interactions. Several network properties and number of feedback loops of resulting Global Regulatory Network has been compared with those of randomly connected networks statistically to test the significance of constructed network. Simulations of the regulatory signal response of Global Regulatory Network of *E. coli* to lactose stimulates have been performed to further verify the resulting regulatory system and simulation model.

Statistical tests and analysis shows that the GlRN of *E. coli* constructed in this work has significantly special network properties comparing to random connected network. And these special properties are closely associated with the biological behavior of regulatory systems such stability and adaptation to environments. Also the biological meaningfulness of the simulation method and GlRN Model has been verified by a nearly perfect match between the simulation results and theoretical expectations. The results of this work suggest the feasibility and meaningfulness of modeling regulatory focused whole-genome and whole-cell signaling systems, and encourage further investigation on this direction.

REFERENCES

1. Fu, Y., L.R. Jarboe, and J.A. Dickerson, *Reconstructing genome-wide regulatory network of E. coli using transcriptome data and predicted transcription factor activities.* BMC Bioinformatics, 2011. **12**: p. 233.

2. Janga, S.C. and B. Contreras-Moreira, *Dissecting the expression patterns of transcription factors across conditions using an integrated network-based approach.* Nucleic Acids Res, 2010. **38**(20): p. 6841-56.

3. Balazsi, G. and Z.N. Oltvai, *Sensing your surroundings: how transcription-regulatory networks of the cell discern environmental signals.* Sci STKE, 2005. **2005**(282): p. pe20.

4. Liao, J.C., et al., *Network component analysis: reconstruction of regulatory signals in biological systems.* Proc Natl Acad Sci U S A, 2003. **100**(26): p. 15522-7.

APPENDIX A

Test results of GTRNetwork Algorithm combinations

GTRNetwork algorithm using TF-gene network topologies providing different level of information as input, the input initial TF-gene network topologies are obtained by randomly deleting 70%, 50%, 30% or 10% links of the TF-gene links data from RegulonDB 7.0

Area under precision-recall curve of GTRNetwork algorithm combinations								
30% of known regulonDB 7.0 links as training input TF-gene network topology								
Algorithm Combination	AUC rep. 1	AUC rep. 2	AUC rep. 3	AUC rep. 4	AUC rep. 5	Average	Standard Deviation	95% Confidence level (Average +/-)
N-C-C	0.076671	0.066471	0.071091	0.069857	0.066552	0.070128	0.0041824	0.003666
E-C-C	0.101369	0.102941	0.094092	0.100902	0.096386	0.099138	0.0037279	0.00326759
S-C-C	0.101193	0.104256	0.101721	0.10235	0.095194	0.100943	0.0034159	0.00299411
P-C-C	0.050712	0.04397	0.030715	0.028649	0.061647	0.043138	0.0138284	0.01212091
N-C-N	0.076671	0.066471	0.071091	0.069857	0.066552	0.070128	0.0041824	0.003666
E-C-N	0.101868	0.102759	0.093834	0.094656	0.106214	0.099866	0.0053898	0.00472427
S-C-N	0.101193	0.104256	0.101721	0.10235	0.095194	0.100943	0.0034159	0.00299411
P-C-N	0.050712	0.04397	0.032224	0.028649	0.061647	0.04344	0.013502	0.01183482
N-A-C	0.074901	0.065236	0.069692	0.067282	0.066595	0.068741	0.0038038	0.00333413
E-A-C	0.11018	0.1147	0.108586	0.109286	0.112942	0.111139	0.0025885	0.00226884
S-A-C	0.106712	0.109543	0.103712	0.110286	0.1036	0.10677	0.0031406	0.00275282
P-A-C	0.038509	0.039149	0.016859	0.018917	0.052538	0.033195	0.0150714	0.01321041
N-A-N	0.074901	0.065236	0.069692	0.067282	0.066595	0.068741	0.0038038	0.00333413
E-A-N	0.115271	0.115089	0.108586	0.109286	0.116227	0.112892	0.0036453	0.00319518
S-A-N	0.106712	0.109543	0.103712	0.110286	0.1036	0.10677	0.0031406	0.00275282
P-A-N	0.035509	0.039149	0.016859	0.018917	0.062677	0.034622	0.0185104	0.01622478

50% of known regulonDB 7.0 links as training input TF-gene network topology								
Algorithm Combination	AUC rep. 1	AUC rep. 2	AUC rep. 3	AUC rep. 4	AUC rep. 5	Average	Standard Deviation	95% Confidence level (Average +/-)
N-C-C	0.063152	0.055708	0.062173	0.060711	0.052044	0.058758	0.0047213	0.00413831
E-C-C	0.106166	0.098208	0.087387	0.095777	0.113617	0.100231	0.0100413	0.00880142
S-C-C	0.110801	0.110348	0.099047	0.111407	0.096001	0.105521	0.0073885	0.0064762
P-C-C	0.060522	0.059998	0.044009	0.085775	0.046371	0.059335	0.0166125	0.0145612
N-C-N	0.063152	0.055708	0.062173	0.060711	0.052044	0.058758	0.0047213	0.00413831
E-C-N	0.10493	0.102257	0.087978	0.099542	0.11446	0.101833	0.009571	0.00838921
S-C-N	0.110801	0.110348	0.099047	0.111407	0.096001	0.105521	0.0073885	0.0064762
P-C-N	0.073677	0.059551	0.044009	0.081608	0.046371	0.061043	0.0165092	0.01447065
N-A-C	0.062511	0.054143	0.060899	0.060152	0.05156	0.057853	0.0047335	0.004149
E-A-C	0.117634	0.115871	0.109702	0.126044	0.128093	0.119469	0.0075717	0.00663679
S-A-C	0.121154	0.106542	0.104112	0.126202	0.119306	0.115463	0.0096298	0.00844072
P-A-C	0.058192	0.056528	0.032829	0.076458	0.033996	0.051601	0.0183569	0.01609023
N-A-N	0.062511	0.054143	0.060899	0.060152	0.05156	0.057853	0.0047335	0.004149
E-A-N	0.117309	0.114527	0.108546	0.123207	0.128432	0.118404	0.0076982	0.00674761
S-A-N	0.121154	0.106542	0.104112	0.126202	0.119306	0.115463	0.0096298	0.00844072
P-A-N	0.063795	0.043432	0.022258	0.092355	0.033996	0.051167	0.0275969	0.02418927

70% of known regulonDB 7.0 links as training input TF-gene network topology								
Algorithm Combination	AUC rep. 1	AUC rep. 2	AUC rep. 3	AUC rep. 4	AUC rep. 5	Average	Standard Deviation	95% Confidence level (Average +/-)
N-C-C	0.044919	0.053894	0.049285	0.042504	0.048879	0.047896	0.0043819	0.00384082
E-C-C	0.109379	0.083959	0.085507	0.100116	0.094757	0.094744	0.010544	0.00924204
S-C-C	0.10853	0.087111	0.108278	0.109674	0.092511	0.101221	0.0106021	0.00929294
P-C-C	0.059501	0.058564	0.06932	0.060274	0.056321	0.060796	0.0049902	0.00437402
N-C-N	0.044919	0.053894	0.049285	0.042504	0.048879	0.047896	0.0043819	0.00384082
E-C-N	0.108787	0.083445	0.088529	0.102396	0.089058	0.094443	0.0106529	0.00933751
S-C-N	0.10853	0.087111	0.108278	0.109674	0.092511	0.101221	0.0106021	0.00929294
P-C-N	0.059501	0.058141	0.068208	0.058826	0.056321	0.0602	0.0046311	0.00405922
N-A-C	0.044315	0.052739	0.047307	0.044978	0.048817	0.047631	0.0033777	0.0029606
E-A-C	0.112172	0.10093	0.111217	0.128538	0.114497	0.113471	0.0099006	0.00867806
S-A-C	0.101989	0.09329	0.108712	0.128104	0.092565	0.104932	0.0145632	0.01276497
P-A-C	0.056507	0.059217	0.069528	0.061777	0.070379	0.063482	0.0062022	0.00543639
N-A-N	0.044315	0.052739	0.047307	0.044978	0.048817	0.047631	0.0033777	0.0029606
E-A-N	0.111292	0.10093	0.107942	0.124722	0.112773	0.111532	0.0086725	0.00760162
S-A-N	0.101989	0.09329	0.108712	0.128104	0.092565	0.104932	0.0145632	0.01276497
P-A-N	0.049403	0.059353	0.070992	0.067034	0.064819	0.06232	0.0083565	0.00732464

90% of known regulonDB 7.0 links as training input TF-gene network topology								
Algorithm Combination	AUC rep. 1	AUC rep. 2	AUC rep. 3	AUC rep. 4	AUC rep. 5	Average	Standard Deviation	95% Confidence level (Average +/-)
N-C-C	0.026407	0.009021	0.024838	0.035376	0.021095	0.023347	0.0095735	0.00839138
E-C-C	0.066524	0.058837	0.068202	0.050779	0.063091	0.061487	0.006977	0.00611548
S-C-C	0.044262	0.053516	0.054446	0.047655	0.070595	0.054095	0.010135	0.00888355
P-C-C	0.037763	0.068964	0.063395	0.043625	0.060197	0.054789	0.0134048	0.0117496
N-C-N	0.026407	0.009021	0.024838	0.035376	0.021095	0.023347	0.0095735	0.00839138
E-C-N	0.065276	0.05913	0.068954	0.044775	0.063221	0.060271	0.009362	0.00820602
S-C-N	0.044262	0.053516	0.054446	0.047655	0.070595	0.054095	0.010135	0.00888355
P-C-N	0.037428	0.069916	0.063823	0.050846	0.059801	0.056363	0.0126502	0.01108821
N-A-C	0.02495	0.009439	0.025112	0.031515	0.023216	0.022847	0.008133	0.00712876
E-A-C	0.043666	0.076072	0.081618	0.070644	0.094129	0.073226	0.018681	0.01637431
S-A-C	0.047646	0.067687	0.079409	0.065627	0.081007	0.068275	0.0134068	0.01175135
P-A-C	0.037052	0.060022	0.056224	0.054717	0.058839	0.053371	0.0093592	0.00820357
N-A-N	0.02495	0.009439	0.025112	0.031515	0.023216	0.022847	0.008133	0.00712876
E-A-N	0.042637	0.081795	0.080575	0.076983	0.095794	0.075557	0.0197445	0.01730647
S-A-N	0.047703	0.067687	0.079409	0.065627	0.081007	0.068287	0.013385	0.01173226
P-A-N	0.033518	0.060829	0.058804	0.052364	0.056811	0.052465	0.0110457	0.00968178

APPENDIX B

Potential new regulatory links of *E. coli* predicted using GTRNetwork

Gene expression data is obtained from M3D database and contains 466 transcriptome experiment conditions on 4279 gene probes. TF-gene regulatory network from RegulonDB 7.0 is used as the initial known TF-gene regulatory topology input. 381 potential new gene regulatory links are predicted. The reconstructed network size can be used as a reference of confidence of the predicted links. Smaller reconstructed network sizes indicate more confidential predictions. Gene functions information is downloaded from EcoGene database

*** Target genes previously had no known regulators**

TF	Gene	Reconstructed network size	Gene Function
DicA	insD*	100	IS2 transposase B
DicA	intQ*	100	Function Unknown
DicA	ydfE*	100	Function Unknown
AscG	groL*	200	Chaperonin Cpn60; phage morphogenesis; GroESL large subunit GroEL, weak ATPase; binds Ap4A
AscG	groS*	200	Chaperonin Cpn10; GroESL small subunit GroES; phage morphogenesis
AsnC	rsmG*	200	16S rRNA m(7)G527 methyltransferase, SAM-dependent; mutant has low level streptomycin resistance
CysB	yeeD*	200	Function unknown
CysB	yeeE*	200	Inner membrane protein, UPF0394 family, function unknown
DcuR	pepE*	200	alpha-Aspartyl dipeptidase
Fur	ybdB*	200	Function Unknown
Fur	yncE*	200	Secreted protein, function unknown
IscR	fdx*	200	Ferredoxin, an iron-sulfur protein; involved in assembly of other Fe-S clusters
IscR	hscA*	200	Hsc66, DnaK-like chaperone, specific for IscU; involved in FtsZ-ring formation; HscB is the J-like co-chaperone for HscA

TF	Gene	Reconstructed network size	Gene Function
IscR	hscB*	200	Hsc20, DnaJ-like co-chaperone for HscA; specific for IscU
IscR	iscX*	200	Function unknown; downstream of fdx, isc genes; binds IscS; possibly involved in Fe-S cluster assembly
SgrR	sroA*	200	Function Unknown
AscG	dnaJ*	300	DnaK co-chaperone; DNA-binding protein; stress-related DNA biosynthesis, responsive to heat shock; binds Zn(II)
AscG	dnaK*	300	Hsp70 molecular chaperone, heat-inducible; bichaperone with ClpB for protein disaggregation
AscG	tpke11*	300	Function Unknown
CadC	yjdL*	300	Probable dipeptide and tripeptide permease; membrane protein
CpxR	cheB	300	Chemotaxis MCP protein-glutamate methylesterase; reverses CheR methylation at specific MCP glutamates
CpxR	cheR	300	Chemotaxis MCP protein methyltransferase, SAM-dependent; binds C-terminus of chemoreceptors; makes glutamate methyl esters
CpxR	cheY	300	Response regulator for chemotactic signal transduction; CheA is the cognate sensor protein
CpxR	cheZ	300	CheY-P phosphatase
CpxR	tap	300	Dipeptide chemoreceptor, methyl-accepting; MCP IV; flagellar regulon
CpxR	tar	300	Aspartate, maltose chemoreceptor, methyl-accepting; MCP II; also senses repellents cobalt and nickel; flagellar regulon
CpxR	flgK*	300	Flagellar synthesis, hook-associated protein
CpxR	flgL*	300	Flagellar synthesis, hook-associated protein
CpxR	fliC	300	Flagellin, structural gene, H-antigen
FadR	sroD*	300	Function Unknown
Fis	cysT*	300	Cysteine tRNA(GCA)
Fis	glyW*	300	Glycine tRNA(GCC) 3
Fis	leuZ*	300	Leucine tRNA(GAG) 4
Fis	argQ*	300	Arginine tRNA(ACG) 2; tandem quadruplicate genes
Fis	argV*	300	Arginine tRNA(ACG) 2; tandem quadruplicate genes
Fis	argY*	300	Arginine tRNA(ACG) 2; tandem quadruplicate genes
Fis	argZ*	300	Arginine tRNA(ACG) 2; tandem quadruplicate genes
Fis	serV*	300	Serine tRNA(GCU) 3
FruR	pyrG*	300	CTP synthase
Fur	efeU*	300	Function Unknown
Fur	ydiE*	300	Function unknown, hemin uptake protein HemP homolog
Fur	yqjH*	300	Function unknown
LeuO	ilvH	300	Acetohydroxy acid synthase III (AHAS-III); acetolactate synthase III (ALS-III); valine sensitive; small subunit
LeuO	ilvI	300	Acetohydroxy acid synthase III (AHAS-III); acetolactate synthase III (ALS-III); valine sensitive; large subunit
MarA	ltaE*	300	L-allo-threonine aldolase
MarA	ybjT*	300	Function unknown

TF	Gene	Reconstructed network size	Gene Function
PepA	pyrB*	300	Aspartate carbamoyltransferase, catalytic subunit; ATCase; aspartate transcarbamylase; aspartate transcarbamoylase
PepA	pyrI*	300	Aspartate carbamoyltransferase, regulatory subunit; aspartate transcarbamylase; ATCase; aspartate transcarbamoylase
PepA	pyrL*	300	pyrBI operon regulatory leader peptide
SdiA	ddlB	300	D-alanine:D-alanine ligase B, ADP-forming
SdiA	ftsI	300	Transpetidase, PBP3; penicillin-binding protein 3 involved in septal peptidoglycan synthesis
SdiA	ftsL	300	Cell division and growth, membrane protein
SdiA	ftsW	300	Stabilizes FtsZ ring, membrane protein; faciltates septal petidoglycan synthesis by recruiting the cognate FtsI transpeptidase; SEDS protein
SdiA	lpxC	300	Lipid A synthesis, UDP-3-O-(R-3-hydroxymyristoyl)-N-acetylglucosamine deacetylase; zinc metalloamidase; cell envelope and cell separation
SdiA	mraY	300	UDP-N-acetylmuramoyl-pentapeptide:undecaprenyl-PO4 phosphatase
SdiA	mraZ*	300	Function unknown, MraZ family; expressed gene in dcw (division, cell wall) gene cluster
SdiA	murC	300	UDP-N-acetylmuramate:L-alanine ligase; L-alanine adding enzyme
SdiA	murD	300	D-glutamic acid adding enzyme; UDP-N-acetylmuramoyl-L-alanine:D-glutamate ligase
SdiA	murE	300	meso-diaminopimelate adding enzyme; UDP-N-acetylmuramoyl-L-alanyl-D-glutamate:meso-diaminopimelate ligase
SdiA	murF	300	D-alanyl:D-alanine adding enzyme; UDP-N-acetylmuramoyl-tripeptide:D-alanyl-D-alanine ligase
SdiA	murG	300	N-acetylglucosaminyl transferase; UDP-N-acetylglucosamine:N-acetylmuramyl-(pentapeptide) pyrophosphoryl-undecaprenol N-acetylglucosamine transferase; murein synthesis peripheral membrane protein interacting with cardiolipin
SdiA	rsmH*	300	16S rRNA m(4)C1402 methyltranserfase, SAM-dependent; membrane-associated. expressed gene in dcw gene cluster; non-essential
YoeB-YefM	yeeZ*	300	Function unknown; predicted enzyme with a nucleoside diphosphate sugar substrate and an NAD(P) cofactor
Zur	yebA*	300	Predicted metalloprotease, function unknown; M37 family
AgaR	bcsA*	400	Celullose synthase, catalytic subunit; inner membrane protein
AgaR	bcsB	400	Cellulose synthase, regulatory subunit; may bind cyclic-di-GMP; probably periplasmic
AgaR	bcsC*	400	Oxidase involved in cellulose synthesis
AgaR	bcsZ	400	Endo-1,4-D-glucanase; breaks down carboxymethylcellulose; periplasmic cellulase
AgaR	yjbE	400	Extracellular polysaccharide production
AgaR	yjbF	400	Extracellular polysaccharide production, novel lipoprotein
AgaR	yjbG	400	Extracellular polysaccharide production
AgaR	yjbH	400	Extracellular polysaccharide production
AlsR	yjcS*	400	Function unknown

TF	Gene	Reconstructed network size	Gene Function
ArcA	maeB*	400	NADP-dependent malic enzyme; NADP-ME
ArcA	sthA*	400	Soluble pyridine nucleotide transhydrogenase
AscG	clpB*	400	Bichaperone with DnaK for protein disaggregation; protein-dependent ATPase; role in de novo protein folding under mild stress conditions
AscG	hslU*	400	Heat-inducible ATP-dependent protease HslVU, ATPase subunit; involved in the degradation of misfolded proteins; heat shock protein D48.5
AscG	hslV*	400	Heat-inducible ATP-dependent protease HslVU, protease subunit; involved in the degradation of misfolded proteins
AscG	ybbN*	400	DnaK co-chaperone, thioredoxin-like protein; has SXXC not CXXC motif
BaeR	katE	400	Catalase hydroperoxidase II, heme d-containing; response to oxidative stress; chromate resistance
CpxR	flgA	400	Flagellar basal body P-ring formation
CpxR	flgM	400	Anti-sigma 28 (FliA) factor; regulator of FlhD
CpxR	flgN	400	Initiation of flagellar filament assembly
CpxR	fliD	400	Hook-associated protein 2, axial family
CpxR	fliS	400	Flagellar chaperone, inhibits premature FliC assembly; cytosolic
CpxR	fliT	400	Flagellar synthesis, predicted chaperone, role unknown
CpxR	ves*	400	Cold and stress-inducible protein, function unknown
CpxR	ycgR	400	Cyclic-di-GMP receptor, regulates motility; mutation suppresses motility defect of hns and yhjH mutants
CpxR	yhjH	400	Cyclic-di-GMP phosphodiesterase, FlhDC-regulated; suppresses motility defect of hns mutants in multicopy
CynR	lacA	400	Thiogalactoside acetyltransferase
CynR	lacY	400	Lactose permease; galactoside permease
CynR	lacZ	400	beta-D-Galactosidase
CysB	yciW*	400	Function unknown
DcuR	yjjI*	400	Putative glycine radical enzyme, function unknown
FhlA	dmsA	400	DMSO reductase subunit A, anaerobic, periplasmic
FhlA	dmsB	400	DMSO reductase subunit B; apparent Fe-S binding subunit; anaerobic
FhlA	dmsC	400	DMSO reductase subunit C, periplasmic; has a membrane bound anchor
FhlA	cysG	400	Siroheme synthase, multifunctional enzyme; has three activities: uroporphyrinogen III methyltransferase, SAM-dependent; precorrin-2 dehydrogenase; sirohydrochlorin ferrochelatase
FhlA	nirB	400	Nitrite reductase [NAD(P)H] subunit
FhlA	nirC	400	Nitrite uptake transporter; membrane protein
FhlA	nirD	400	Nitrite reductase [NAD(P)H] subunit
FhlA	yhbU*	400	Function unknown, U32 peptidase family
FhlA	yhbV*	400	Function unknown, U32 peptidase family
FlhDC	cheB	400	Chemotaxis MCP protein-glutamate methylesterase; reverses CheR methylation at specific MCP glutamates

TF	Gene	Reconstructed network size	Gene Function
FlhDC	cheR	400	Chemotaxis MCP protein methyltransferase, SAM-dependent; binds C-terminus of chemoreceptors; makes glutamate methyl esters
FlhDC	cheY	400	Response regulator for chemotactic signal transduction; CheA is the cognate sensor protein
FlhDC	cheZ	400	CheY-P phosphatase
FlhDC	tap	400	Dipeptide chemoreceptor, methyl-accepting; MCP IV; flagellar regulon
FlhDC	tar	400	Aspartate, maltose chemoreceptor, methyl-accepting; MCP II; also senses repellents cobalt and nickel; flagellar regulon
FlhDC	flgK*	400	Flagellar synthesis, hook-associated protein
FlhDC	flgL*	400	Flagellar synthesis, hook-associated protein
Fur	bfd*	400	Bacterioferritin-associated ferredoxin; predicted redox component complexing with Bfr in iron storage and mobility [2Fe-2S]
Fur	bfr*	400	Bacterioferritin; negatively regulated by ryhB RNA as part of indirect positive regulation by Fur; 24-mer
Fur	efeB*	400	Deferrrochelatase, periplasmic; inactive acid inducible low-pH ferrous ion transporter EfeUOB; periplasmic acid peroxidase; heme cofactor
Fur	efeO*	400	Inactive acid-inducible low-pH ferrous ion transporter EfeUOB; acid-inducible periplasmic protein
Fur	ybaN*	400	Inner membrane protein, DUF454 family, function unknown
LexA	ymfJ*	400	Function unknown, e14 prophage
MarA	yhbW*	400	Function unknown, luciferase-like
MetR	metF	400	5,10-Methylenetetrahydrofolate reductase
Nac	ppc*	400	Phosphoenolpyruvate carboxylase; monomeric
RcsAB	ygaU*	400	Function unknown
Rob	exbB	400	Uptake of enterochelin; resistance or sensitivity to colicins; similarity with TolQ
Rob	exbD	400	Uptake of enterochelin; resistance or sensitivity to colicins; similarity with TolR
Rob	fhuF	400	Siderophore-iron reductase; releases iron from hydroxamate-type siderophores; cytoplasmic
Zur	lpxM*	400	Lipid A synthesis, KDO2-lauroyl-lipid IVA myristoyl-ACP acyltransferase
AgaR	gspC*	500	Part of H-NS-silenced gsp divergon, type II protein secretion;cloned gsp divergon secretes ChiA
AgaR	gspD*	500	Part of H-NS-silenced gsp divergon, type II protein secretion; OM secretin; cloned gsp divergon secretes ChiA
AgaR	gspE*	500	Part of H-NS-silenced gsp divergon, type II protein secretion; cloned gsp divergon secretes ChiA
AgaR	gspF*	500	Part of H-NS-silenced gsp divergon, type II protein secretion; cloned gsp divergon secretes ChiA
AgaR	gspG*	500	Pseudopilin in H-NS-silenced gsp divergon, type II secreton; cloned gsp divergon secretes ChiA
AgaR	gspH*	500	Part of H-NS-silenced gsp divergon, type II protein secretion; cloned gsp divergon secretes ChiA

TF	Gene	Reconstructed network size	Gene Function
AgaR	gspI*	500	Part of H-NS-silenced gsp divergon, type II protein secretion; cloned gsp divergon secretes ChiA
AgaR	gspJ*	500	Part of H-NS-silenced gsp divergon, type II protein secretion; cloned gsp divergon secretes ChiA
AgaR	gspK*	500	Part of H-NS-silenced gsp divergon, type II protein secretion; cloned gsp divergon secretes ChiA
AgaR	gspL*	500	Part of H-NS-silenced gsp divergon, type II protein secretion; cloned gsp divergon secretes ChiA
AgaR	gspM*	500	Part of H-NS-silenced gsp divergon, type II protein secretion; cloned gsp divergon secretes ChiA
AgaR	gspO	500	Prepilin peptidase in H-NS-silenced gsp divergon; type II protein secretion; cloned gsp divergon secretes ChiA
AgaR	phnC	500	Phosphonate uptake, ATP-binding protein; ABC transporter
AgaR	phnD	500	Phosphonate uptake, periplasmic binding protein; ABC transporter
AgaR	phnE	500	Function Unknown
AgaR	phnE	500	Function Unknown
AgaR	phnF	500	Phosphonate utilization, probable regulatory gene
AgaR	phnG	500	Carbon-phosphorus lyase complex subunit
AgaR	phnH	500	Carbon-phosphorus lyase complex subunit
AgaR	phnI	500	Carbon-phosphorus lyase complex subunit
AgaR	phnJ	500	Carbon-phosphorus lyase complex subunit
AgaR	phnK	500	Carbon-phosphorus lyase complex subunit
AgaR	phnL	500	Carbon-phosphorus lyase complex subunit
AgaR	phnM	500	Carbon-phosphorus lyase complex subunit
AgaR	phnN	500	Carbon-phosphorus lyase complex, ribose 1,5-bisphosphokinase subunit; also functions in an alternative pathway for PRPP formation
AgaR	phnO	500	Unknown role in C-P lyase complex; in phn operon for phosphonate utilization, putative acetyltransferase
AgaR	phnP	500	Carbon-phosphorus lyase complex membrane-bound subunit; 2',3'-cyclic nucleotide phosphodiesterase; bis(p-nitrophenyl)phosphate phosphodiesterase
AgaR	rhaA	500	L-Rhamnose isomerase
AgaR	rhaB	500	Rhamnulokinase
AgaR	rhaD	500	Rhamnulose-1-phosphate aldolase; homotetrameric
AgaR	xdhD*	500	Probable hypoxanthine oxidase; mutation confers adenine sensitivity
AgaR	ygfM*	500	Function unknown
AlsR	idnD	500	L-idonate 5-dehydrogenase
AlsR	idnO	500	5-keto-D-gluconate 5-reductase
AlsR	idnR	500	idn operon activator; represses GntR-regulated genes gntKU and gntT
AlsR	idnT	500	L-idonate transporter; also transports 5-keto-D-gluconate (Reed, 2006)
AlsR	ulaA	500	PTS Enzyme IIC transport protein; involved in L-ascorbate uptake
AlsR	ulaB	500	PTS Enzyme IIB; involved in L-ascorbate uptake
AlsR	ulaC	500	PTS Enzyme IIA; involved in L-ascorbate uptake

TF	Gene	Reconstructed network size	Gene Function
AlsR	ulaD	500	3-keto-L-gulonate-6-phosphate decarboxylase; involved in the utilization of L-ascorbate by anaerobic fermentation; dimeric
AlsR	ulaE	500	L-xylulose 5-phosphate 3-epimerase; involved in the utilization of L-ascorbate by anaerobic fermentation
AlsR	ulaF	500	L-ribulose 5-phosphate 4-epimerase; involved in the utilization of L-ascorbate by anaerobic fermentation
AlsR	ulaG	500	L-ascorbate-6-phosphate lactonase, involved in the utilization of L-ascorbate by anaerobic fermentation; has phosphodiesterase activity
AscG	grpE	500	Nucleotide exchange factor for the DnaKJ chaperone; heat shock protein; mutant survives lambda induction; stimulates DnaK and HscC ATPase
AscG	lipB*	500	Lipoyl-protein ligase; lipoyl-[ACP]:protein N-lipoyltransferase
AscG	ybeD*	500	Required for swarming phenotype, UPF0250 family, function unknown; structural similarity to the regulatory domain from d-3-phosphoglycerate dehydrogenase
BaeR	fbaB*	500	Fructose 1,6-bisphosphate aldolase, class I
BaeR	lsrA	500	Autoinducer 2 (AI-2) import ATP-binding protein; essential for aerobic growth; upregulated in biofilms
BaeR	lsrB	500	Autoinducer-2 (AI-2)-binding protein
BaeR	lsrC	500	Autoinducer 2 (AI-2) import system permease protein
BaeR	lsrD	500	Autoinducer-2 (AI-2) import system permease protein
BaeR	lsrF	500	Function unknown, involved in AI-2 catabolism
BaeR	lsrG	500	Autoinducer 2-degrading protein; ygiN paralog
BaeR	tam	500	Trans-aconitate 2-methyltransferase, SAM-dependent
CpxR	fliA	500	Transcription factor sigma 28 for class III flagellar operons
CpxR	fliY	500	Cystine-binding protein, periplasmic; not required for motility; may regulate FliA (sigma 28)
CpxR	fliZ	500	RpoS antagonist, transiently in post-exponential phase; timing factor allowing motility to continue for a while during starvation; not required for normal motility
CpxR	flxA	500	Member of FliA regulon, function unknown, Qin prophage
DcuR	ansB	500	L-Asparaginase II
DcuR	aspA	500	L-Aspartate ammonia-lyase; L-aspartase
DcuR	dcuA	500	C4-dicarboxylate transporter, anaerobic
DcuR	hybA	500	Hydrogenase 2 component, periplasmic; possibly electron acceptor for hydrogenase 2 small subunit; probably binds 4 4Fe-4S clusters
DcuR	hybB	500	Hydrogenase 2 cytochrome b type component, probably
DcuR	hybC	500	Hydrogenase 2 [Ni Fe] large subunit, periplasmic
DcuR	hybD	500	Maturation endoprotease for Ni-containing hydrogenase 2
DcuR	hybE	500	Hydrogenase 2-specific chaperone
DcuR	hybF	500	Accessory protein required for the maturation of hydrogenases 1 and 2; may be involved in nickel incorporation
DcuR	hybG	500	Hydrogenase 2 accessory protein; chaperone-like function
DcuR	hybO	500	Hydrogenase 2 [Ni, Fe], small subunit, periplasmic

TF	Gene	Reconstructed network size	Gene Function
EvgA	yfiB*	500	Verified lipoprotein, function unknown
EvgA	yfiN*	500	Predicted diguanylate cyclase, function unknown
EvgA	yfiR*	500	Function unknown
FhlA	dcuB	500	C4-dicarboxylate transporter, anaerobic
FhlA	fumB	500	Fumarase B, anaerobic
FhlA	nikA	500	Nickel-binding, heme-binding periplasmic protein; Tar-dependent Ni-repellant chemosensor; Fnr-dependent
FhlA	nikB	500	Nickel transport system permease
FhlA	nikC	500	Nickel transport system permease
FhlA	nikD	500	Nickel transport ATP-binding protein
FhlA	nikE	500	Nickel transport ATP-binding protein
FhlA	nikR	500	Nickel-responsive regulator of the nik operon; homodimer
Fis	valV*	500	Valine tRNA(GAC) 2B
Fis	valW*	500	Valine tRNA(GAC) 2A
FlhDC	cheA	500	Histidine protein kinase sensor of chemotactic response; CheY is cognate response regulator; autophosphorylating; CheAS is a short form produced by an internal start at codon 98
FlhDC	cheW	500	Chemotaxis signal transducer; bridges CheA to chemoreceptors to regulate phophotransfer to CheY and CheB
FlhDC	motA	500	H+-driven stator protein of flagellar rotation
FlhDC	motB	500	H+-driven stator protein of flagellar rotation
FlhDC	flhC	500	Transcriptional activator of flagellar class II operons; forms heterotetramer with FlhD; CsrA regulon; may be allosteric effector of FlhD
FlhDC	flhD	500	Transcriptional activator of flagellar class II operons; forms heterotetramer with FlhC; possible role in regulation of cell division; can function in vivo independently of FlhC, but does not bind DNA by itself; contains HTH motif
FNR	priB*	500	Primosomal protein n; ssDNA-binding protein
FNR	rplI*	500	50S ribosomal subunit protein L9
FNR	rpsF*	500	30S ribosomal subunit protein S6; suppressor of dnaG-Ts
FNR	rpsR*	500	30S ribosomal subunit protein S18
FruR	tpiA*	500	Triosephosphate isomerase
GadE	flxA	500	Member of FliA regulon, function unknown, Qin prophage
GadE	cheB	500	Chemotaxis MCP protein-glutamate methylesterase; reverses CheR methylation at specific MCP glutamates
GadE	cheR	500	Chemotaxis MCP protein methyltransferase, SAM-dependent; binds C-terminus of chemoreceptors; makes glutamate methyl esters
GadE	cheY	500	Response regulator for chemotactic signal transduction; CheA is the cognate sensor protein
GadE	cheZ	500	CheY-P phosphatase
GadE	tap	500	Dipeptide chemoreceptor, methyl-accepting; MCP IV; flagellar regulon
GadE	tar	500	Aspartate, maltose chemoreceptor, methyl-accepting; MCP II; also senses repellents cobalt and nickel; flagellar regulon

TF	Gene	Reconstructed network size	Gene Function
GadE	yjcZ*	500	Function unknown
GadE	yjdA*	500	Function unknown
IHF	bdm	500	Osmoresponsive gene with reduced expression in biofilms; function unknown
IHF	yahO*	500	Predicted periplasmic protein, YhcN family, function unknown
IHF	yeaG	500	Protein kinase, function unknown; autokinase
IHF	yeaH	500	Function unknown
IscR	gntX	500	Required for the utilization of DNA as a carbon source; H. influenzae competence protein ComF homolog
LexA	rhsE*	500	Function Unknown
LexA	ydcD*	500	Function unknown
LexA	yebF*	500	Exported protein, function unknown
Lrp	gdhA	500	Glutamate dehydrogenase
MarA	phr	500	Deoxyribodipyrimidine photolyase; DNA photolyase; monomeric
MarA	ybgA*	500	Function unknown, DUF1722 family
Nac	asd*	500	Aspartate semialdehyde dehydrogenase
Nac	ilvH	500	Acetohydroxy acid synthase III (AHAS-III); acetolactate synthase III (ALS-III); valine sensitive; small subunit
Nac	ilvI	500	Acetohydroxy acid synthase III (AHAS-III); acetolactate synthase III (ALS-III); valine sensitive; large subunit
Nac	aroA	500	5-enolpyruvyl shikimate-3-phosphate synthase; ESPS synthase; 3-phosphoshikimate-1-carboxyvinyltransferase
Nac	serC	500	Phosphoserine aminotransferase
NanR	isrB*	500	Function Unknown
NanR	yjhB*	500	Predicted transporter, function unknown; N-acetylneuraminic acid inducible
NanR	yjhC*	500	predicted oxidoreductase, function unknown; N-acetylneuraminic acid inducible
NarL	pykA*	500	Pyruvate kinase II, minor
PhoB	nlpD*	500	Lipoprotein, function unknown; may be OM protein involved in cell wall formation and may have murein hydrolytic activity
PhoB	rpoS	500	RNA polymerase subunit, stress and stationary phase sigma S; Sigma38
RcsAB	msyB*	500	In multicopy restores growth and protein export functions of secY and secA mutants
RcsAB	ydiZ*	500	Function unknown
RstA	narK	500	Nitrate/nitrite antiporter; promotes nitrite extrusion and uptake
SgrR	sgrT*	500	Inhibitor of glucose uptake
TorR	aspA	500	L-Aspartate ammonia-lyase; L-aspartase
TorR	dcuA	500	C4-dicarboxylate transporter, anaerobic
TorR	rbsA	500	D-ribose high-affinity transport system
TorR	rbsB	500	D-ribose binding protein, periplasmic; substrate recognition for transport and chemotaxis
TorR	rbsC	500	D-ribose high-affinity transport system, membrane component

TF	Gene	Reconstructed network size	Gene Function
TorR	rbsD	500	D-ribose pyranase; interconverts beta-pyran and beta-furan forms of D-ribose; related to fucose mutarotase FucU
TorR	rbsK	500	Ribokinase
TorR	rbsR	500	Regulatory gene for rbs operon
YoeB-YefM	chpB*	500	ChpB toxin and mRNA interferase, antitoxin is ChpS; reversible inhibitor of translation, by mRNA cleavage
YoeB-YefM	chpS*	500	ChpS antitoxin, toxin is ChpB
YoeB-YefM	hokD	500	Small toxic membrane polypeptide, Qin prophage; homologous to plasmid-encoded plasmid stabilization toxins regulated by antisense RNA; functional relevance of chromosomal homologs is unknown
YoeB-YefM	relB	500	Antitoxin for RelE, Qin prophage; transcriptional repressor of relB operon; mutants have a delayed relaxed regulation of RNA synthesis and slow recovery from starvation
YoeB-YefM	relE	500	Sequence-specific mRNA endoribonuclease, Qin prophage; toxin-antitoxin (TA) pair RelEB, RelE inhibitor of translation cleaves mRNA in A site; binds to its antitoxin RelB and to ribosomes; co-repressor of relB operon transcription; stress-induced
AgaR	agaA*	600	Function Unknown
AgaR	agaW*	600	Function Unknown
AgaR	glnA	600	Glutamine synthase
AgaR	glnG	600	Nitrogen regulator I
AgaR	glnL	600	Bifunctional protein kinase/phosphatase, nitrogen regulator II, NRII; homodimeric
AgaR	argK	600	Required to convert succinate to propionate, Arg transport?; reported to have ATPase and protein kinase activity
AgaR	scpA	600	Methylmalonyl-CoA mutase, B12-dependent
AgaR	scpB	600	Methylmalonyl-CoA decarboxylase
AgaR	scpC	600	Propionyl CoA:succinate CoA transferase
AgaR	ssnA*	600	Causes decline of viability at early stationary phase; negatively regulated by RpoS; cloned product slows growth and causes enlarged filamentous cell morphology; related to chlorohydrolases and aminohydrolases
AgaR	ygfK*	600	Putative selenate reductase subunit; mutants impaired in selenium reduction
AlsR	alsK*	600	Allose kinase
AlsR	nanC	600	N-acetylneuraminic acid outer membrane channel protein
AlsR	nanM	600	N-acetylneuraminic acid mutarotase
ArcA	msrB*	600	Methionine sulfoxide reductase B; specific for met-R-(o) diastereoisomers within proteins; mutant is cadmium sensitive; free met-R-(o) is inefficiently reduced by MsrB
ArsR	yihO*	600	Putative transporter, function unknown
ArsR	yihP*	600	Putative transporter, function unknown, membrane protein
AscG	lnt*	600	Apolipoprotein N-acetyltransferase; copper sensitivity

TF	Gene	Reconstructed network size	Gene Function
AscG	ybeX*	600	Salmonella ortholog involved in Co2+ and Mg2+ efflux; contains two CBS domains; integral membrane protein; possible hemolysin (by homology)
AscG	ybeY*	600	Required for translation at 42C, function unknown; metal-binding heat shock protein
AscG	ybeZ*	600	PhoH paralog, function unknown
BaeR	otsA*	600	Trehalose phosphate synthase; cold- and heat- induced; required for viability at 4C; rpoS regulon
BaeR	otsB*	600	Trehalose phosphate phosphatase; cold- and heat- induced; required for viability at 4C; rpoS regulon; HAD17
CdaR	nanA	600	N-Acetylneuraminate lyase (aldolase)
CdaR	nanE	600	Probable N-acetylmannosamine-6-phosphate 2-epimerase
CdaR	nanK	600	N-acetyl-D-mannosamine (ManNAc) kinase
CdaR	nanT	600	Sialic acid transporter
CdaR	yhcH	600	Required for swarming phenotype, function unknown; last gene in nanATEK-yhcH operon, but not required for growth on sialic acid
CpxR	trg	600	Ribose, galactose chemoreceptor, methyl-accepting; MCP III; flagellar regulon
CpxR	yjcZ*	600	Function unknown
CpxR	yjdA*	600	Function unknown
CpxR	ynjH*	600	Putative secreted protein, function unknown; DUF1496 family
CysB	gsiA*	600	Glutathione transporter ATP-binding protein; GsiABCD is an ABC transporter system
CysB	gsiB*	600	Glutathione periplasmic binding protein; GsiABCD is an ABC transporter system
CysB	gsiC*	600	Glutathione transporter permease; GsiABCD is an ABC transporter system
CysB	gsiD*	600	Glutathione transporter permease; GsiABCD is an ABC transporter system
CysB	iaaA*	600	isoAsp aminopeptidase, cleaves isoAsp-X dipeptides; Ntn hydrolase; glutathione utilization; weak L-asparaginase activity in vitro (EcAIII); precursor is cleaved into an alpha and beta subunit; heterotetrameric
CysB	ydjN*	600	Predicted symporter, function unknown
DcuR	dmsA	600	DMSO reductase subunit A, anaerobic, periplasmic
DcuR	dmsB	600	DMSO reductase subunit B; apparent Fe-S binding subunit; anaerobic
DcuR	dmsC	600	DMSO reductase subunit C, periplasmic; has a membrane bound anchor
DcuR	fhlA	600	Formate hydrogen lyase system activator, global regulator
DcuR	hypA	600	Hydrogenase 3 accessory protein required for activity
DcuR	hypB	600	Required for metallocenter assembly in Hydrogenases 1,2,3; guanine-nucleotide-binding protein; Ni donor for Hyd-3 large subunit; homodimeric
DcuR	hypC	600	Hydrogenase 3 chaperone-type protein; required for Hyd-3 metallocenter assembly, binds HycE subunit

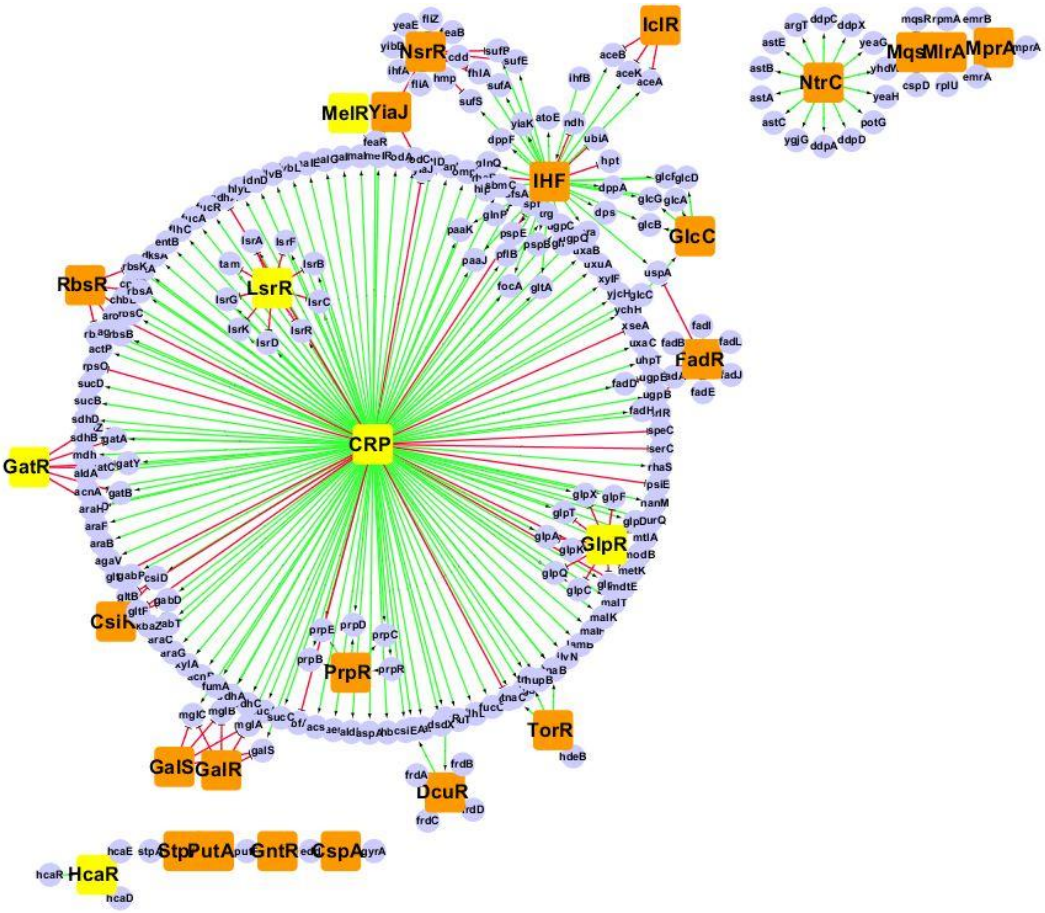
TF	Gene	Reconstructed network size	Gene Function
DcuR	hypD	600	Hydrogenases 1,2,3 accessory protein; required for metallocenter assembly
DcuR	hypE	600	Hydrogenases 1,2,3 accessory protein, carbamoyl dehydratase; required for CN ligand synthesis at the metallocenter; converts thiocarbamate at its C-terminal Cys to a thiocyanate
FadR	astA	600	Arginine succinyltransferase, arginine catabolism
FadR	astB	600	Succinylarginine dihydrolase, arginine catabolism
FadR	astC	600	Succinylornithine transaminase, mutant cannot catabolize arginine, overproduction complements argD mutants; carbon starvation protein
FadR	astD	600	Succinylglutamic semialdehyde dehydrogenase, NAD-dependent; arginine catabolism
FadR	astE	600	Succinylglutamate desuccinylase, arginine catabolism
FhlA	yjjI*	600	Putative glycine radical enzyme, function unknown
FlhDC	ymdA*	600	Function unknown
FNR	rplE*	600	50S ribosomal subunit protein L5; 5S rRNA-binding
FNR	rplF*	600	50S ribosomal subunit protein L6; gentamicin sensitivity
FNR	rplN*	600	50S ribosomal subunit protein L14
FNR	rplO*	600	50S ribosomal subunit protein L15
FNR	rplR*	600	50S ribosomal subunit protein L18; 5S rRNA-binding
FNR	rplX*	600	50S ribosomal subunit protein L24
FNR	rpmD*	600	50S ribosomal subunit protein L30
FNR	rpmJ*	600	50S ribosomal subunit protein X (L36)
FNR	rpsE*	600	30S ribosomal subunit protein S5
FNR	rpsH*	600	30S ribosomal subunit protein S8
FNR	rpsN*	600	30S ribosomal subunit protein S14
FNR	secY*	600	SecYEG inner membrane translocon core subunit; preprotein translocase secAYEG subunit; core translocon secYE subunit
FruR	yeaD*	600	Function unknown, mutarotase homolog, low abundance protein
FruR	pgi	600	Glucose-6-phosphate isomerase
FruR	ybgE*	600	Function unknown, cydAB operon, expressed in minicells
GadE	aer	600	Aerotaxis and redox taxis sensor; flavoprotein; senses intracellular energy (redox) levels, and mediates energy taxis, as does Tsr
GadE	ycgR	600	Cyclic-di-GMP receptor, regulates motility; mutation suppresses motility defect of hns and yhjH mutants
LeuO	aroP	600	General aromatic amino acid transport
Lrp	ilvG_1*	600	Function Unknown
Lrp	ilvG_2*	600	Function Unknown
Lrp	thrA*	600	Aspartokinase I and homoserine dehydrogenase I, bifunctional
Lrp	thrB*	600	Homoserine kinase
Lrp	thrC*	600	Threonine synthase
Lrp	thrL*	600	Regulatory leader peptide for thrABC operon
MarA	talA*	600	Transaldolase A; creBC regulon

TF	Gene	Reconstructed network size	Gene Function
MarA	tktB*	600	Transketolase B; binds Zn(II)
MngR	ycbC*	600	conserved protein, DUF218 superfamily, function unknown
Nac	hisA*	600	1-(5'-phosphoribosyl)-5-[(5'-phosphoribosylamino)methylideneamino] imidazole-4-carboxamide isomerase
Nac	hisB*	600	Imidazoleglycerolphosphate dehydratase/histidinol phosphatase; bifunctional enzyme; HAD21
Nac	hisC*	600	Histidinol-phosphate aminotransferase
Nac	hisD*	600	Histidinol dehydrogenase
Nac	hisF*	600	Imidazole glycerol phosphate (IGP) synthase, cyclase subunit
Nac	hisG*	600	ATP-phosphoribosyltransferase
Nac	hisH*	600	Imidazole glycerol phosphate (IGP) synthase, amidotransferase
Nac	hisI*	600	PR-ATP pyrophosphatase/PR-AMP cyclohydrolase, bifunctional
Nac	hisL*	600	his operon leader peptide
NanR	lldD	600	L-lactate dehydrogenase, FMN dependent
NanR	lldP	600	L-lactate permease; also involved in glycolate uptake
NanR	lldR	600	Dual role activator/repressor for lldPRD operon
NanR	ytfJ*	600	Expressed periplasmic protein, function unknown
NrdR	treB	600	Trehalose permease PTS EIIBC component
NrdR	treC	600	Trehalose-6-phosphate hydrolase, osmoprotectant
PutA	aldA	600	Aldehyde dehydrogenase, NAD-dependent; active on lactaldehyde, glycolaldehyde, and other aldehydes
QseB	agaB	600	Putative PTS system N-acetylgalactosamine-specific enzyme IIB component
QseB	agaC	600	Enzyme IIC Nag, PTS system; N-acetylgalactosamine-specific enzyme IIC; EIIC-Nag
QseB	agaD	600	Enzyme IID Nag, PTS system; N-acetylgalactosamine-specific enzyme IID; EIID-Nag; remnant of aga operon
QseB	agaI	600	Galactosamine-6-phosphate isomerase
QseB	agaS	600	Tagatose-6-phosphate ketose/aldose isomerase
QseB	kbaY	600	Ketose 1,6-bisphosphate aldolase, class II; D-tagatose 1,6-bisphosphate aldolase; requires KbaZ subunit for full activity and stability
RcsAB	elaB*	600	Function unknown
RcsAB	fbaB*	600	Fructose 1,6-bisphosphate aldolase, class I
RcsAB	osmE	600	Osmotically inducible lipoprotein, function unknown
RcsAB	yegP*	600	Function unknown, UPF0339 family
RutR	dos*	600	Function Unknown
SoxS	ydhC*	600	Putative transporter, function unknown; no overexpression resistances found
TorR	pck	600	Phosphoenolpyruvate carboxykinase [ATP]
TyrR	thrA*	600	Aspartokinase I and homoserine dehydrogenase I, bifunctional
TyrR	thrB*	600	Homoserine kinase
TyrR	thrC*	600	Threonine synthase
TyrR	thrL*	600	Regulatory leader peptide for thrABC operon

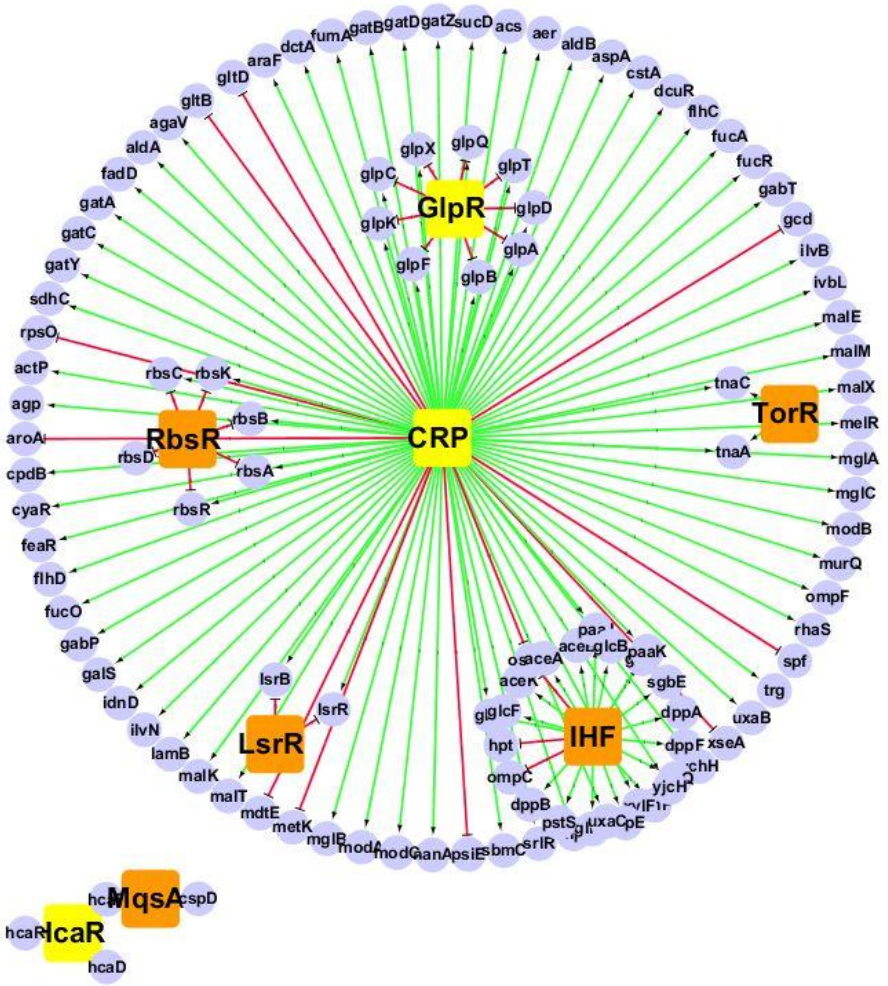
TF	Gene	Reconstructed network size	Gene Function
YoeB-YefM	gntT	600	High-affinity gluconate transport

APPENDIX C:

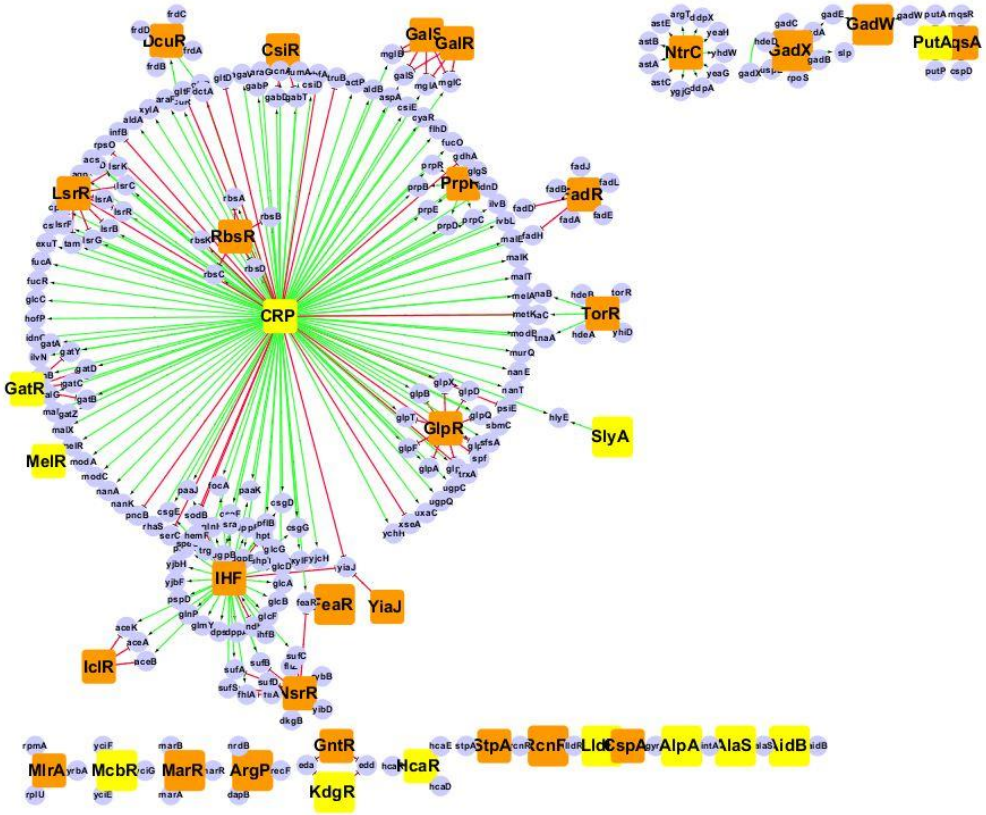
Effective regulatory networks (ERNs) of *E. coli* at condition change from wild type glucose (MOPS media)



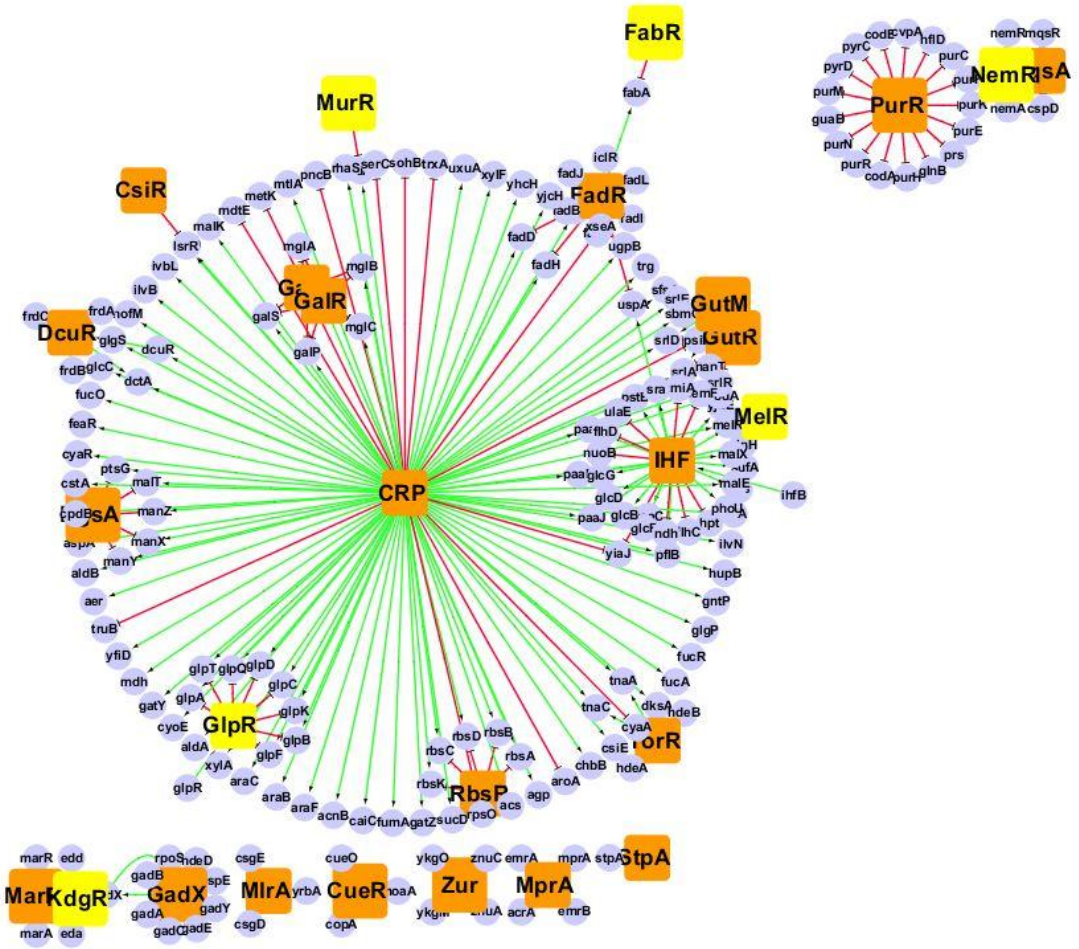
A. Wild Type Acetate



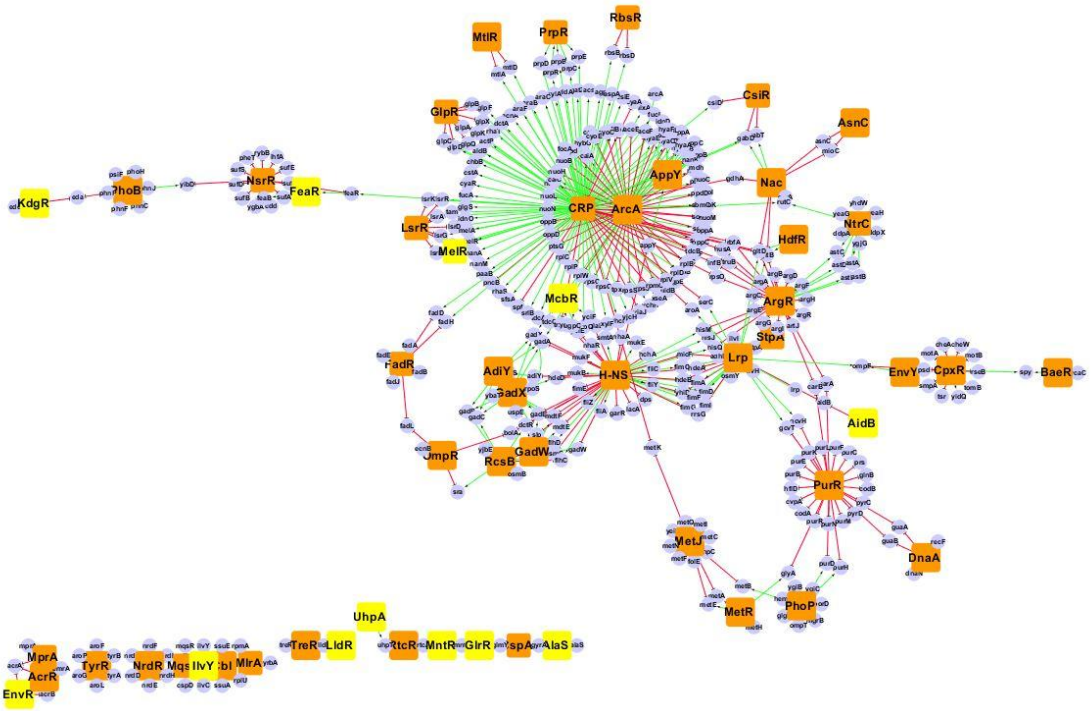
B. Wild Type Glycerol



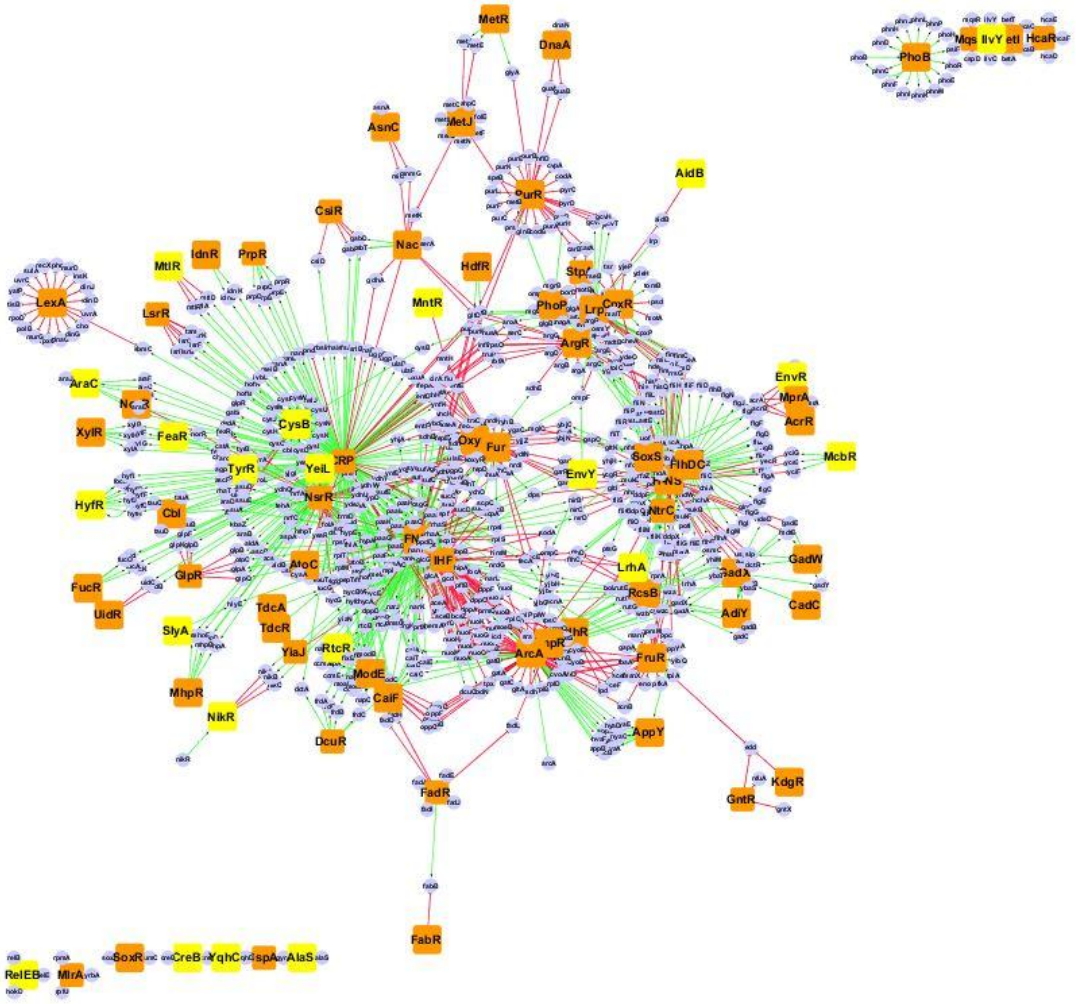
C. Wild Type Proline



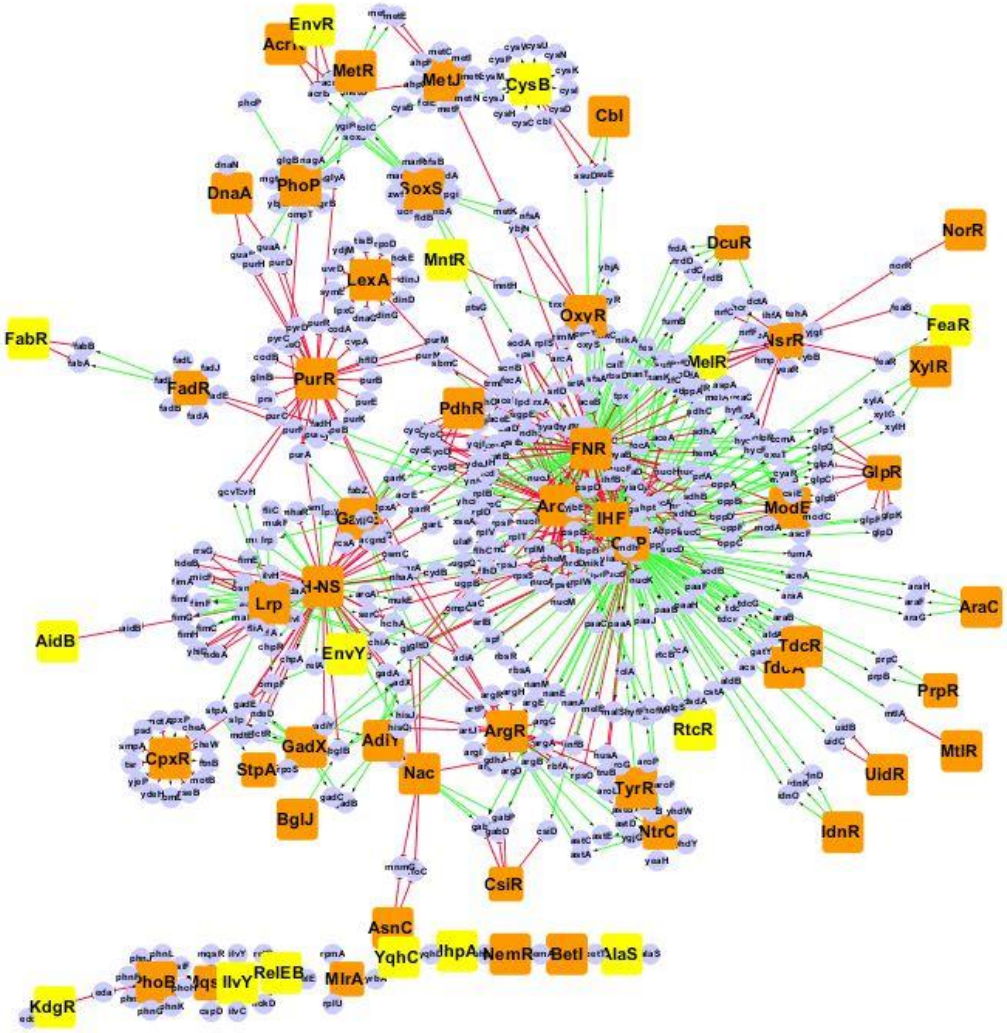
D. Wild Type Stationary



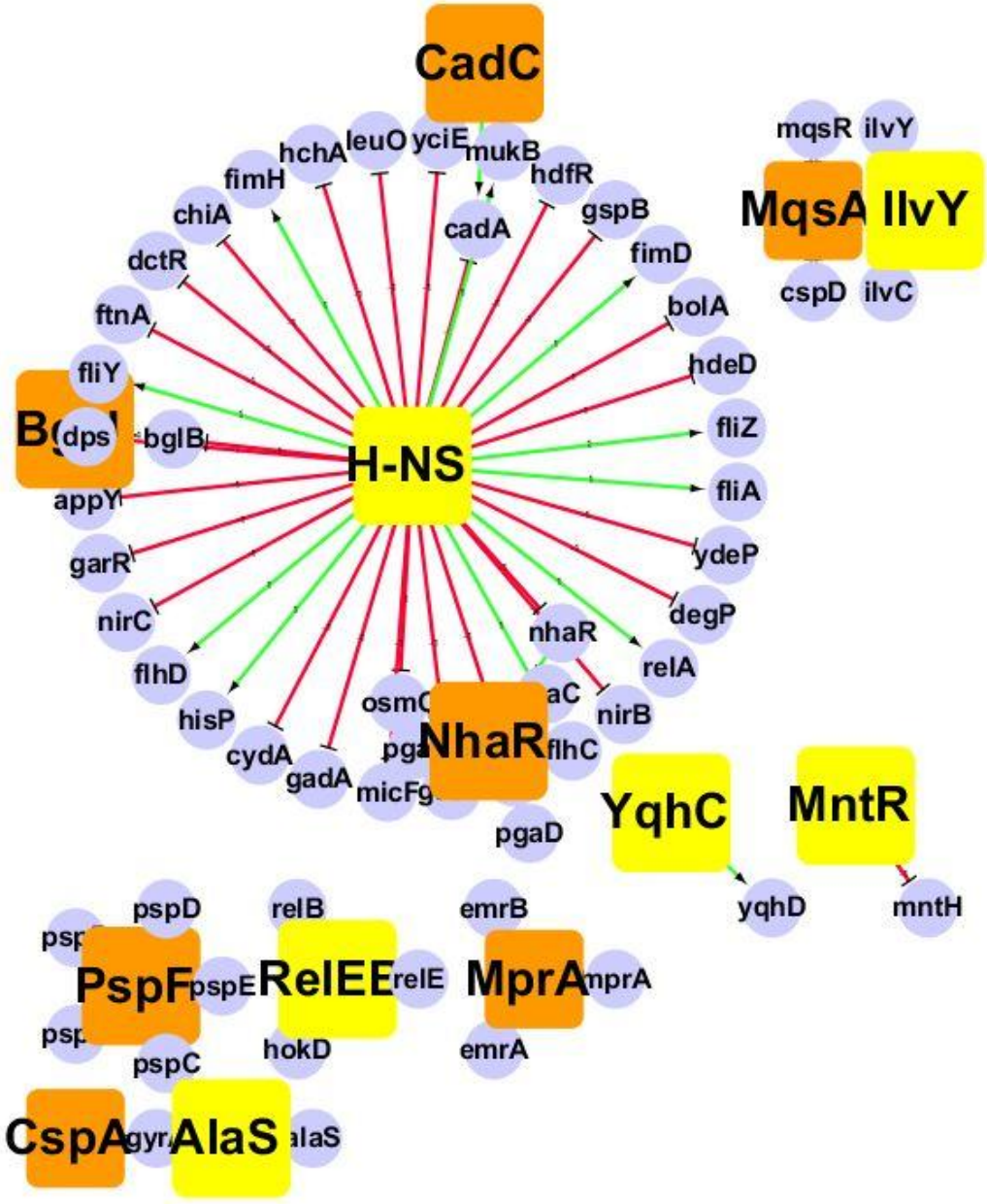
E. Wild Type Stationary2



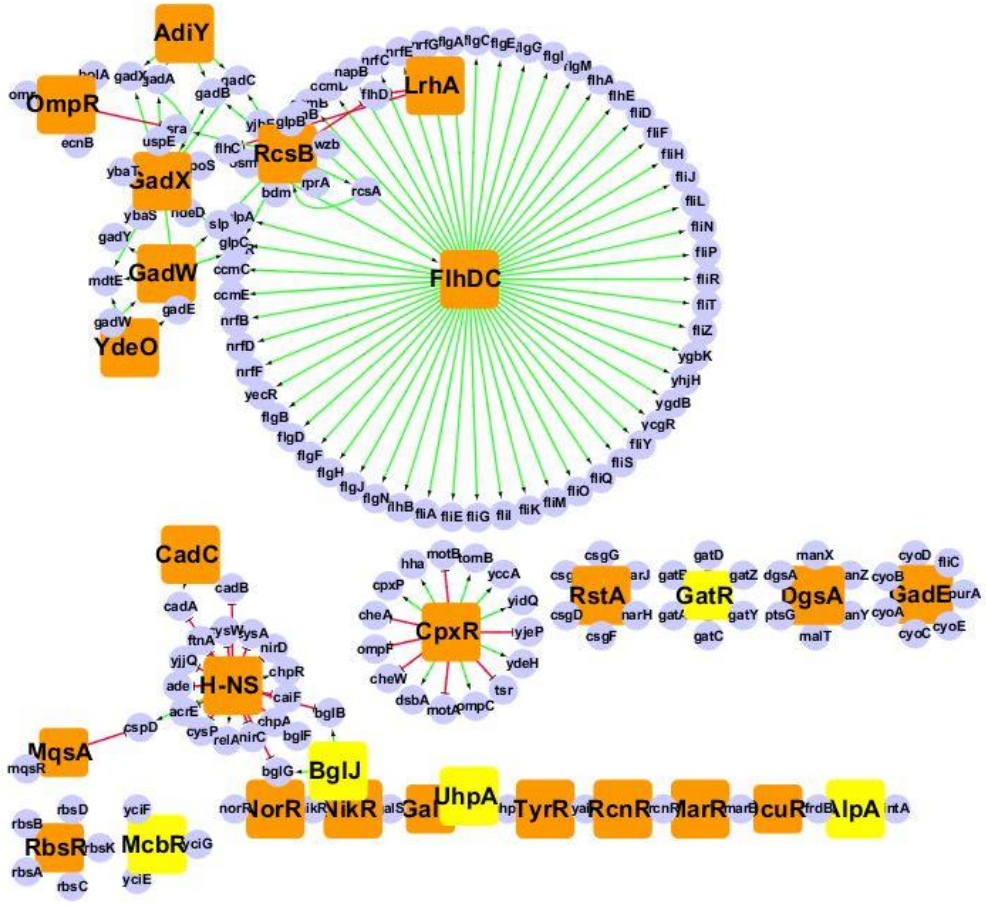
F. Wild Type Stationary3



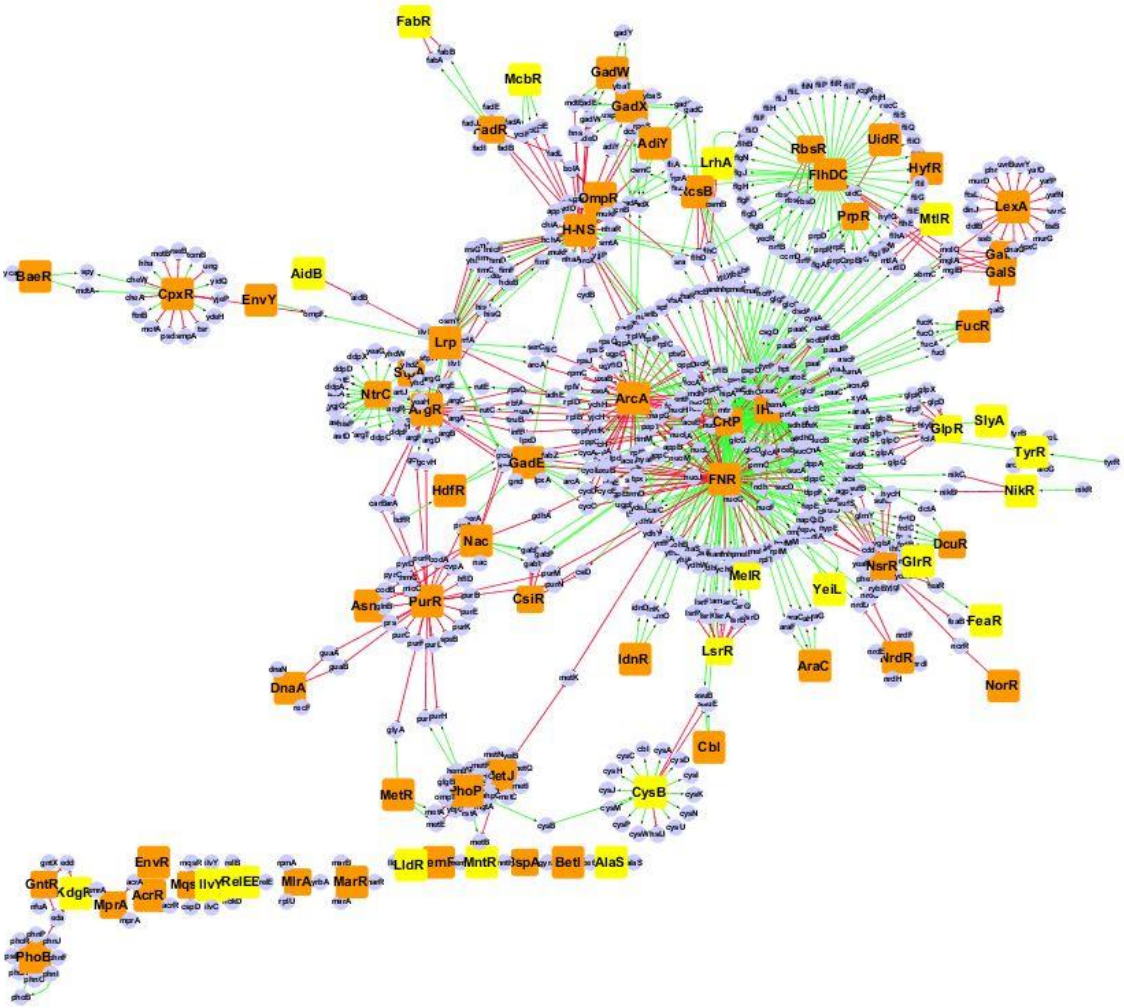
G. Wild Type Stationary4



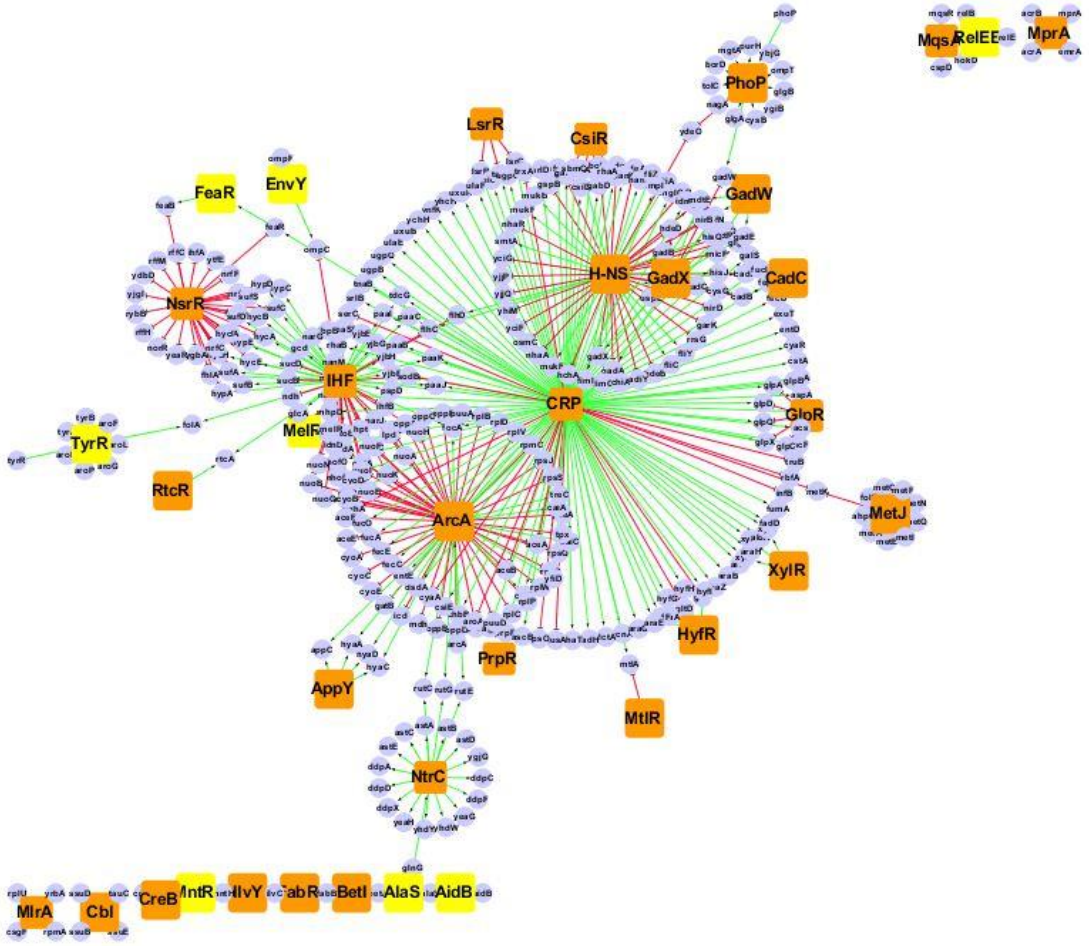
H. Wild Type Heat Shock



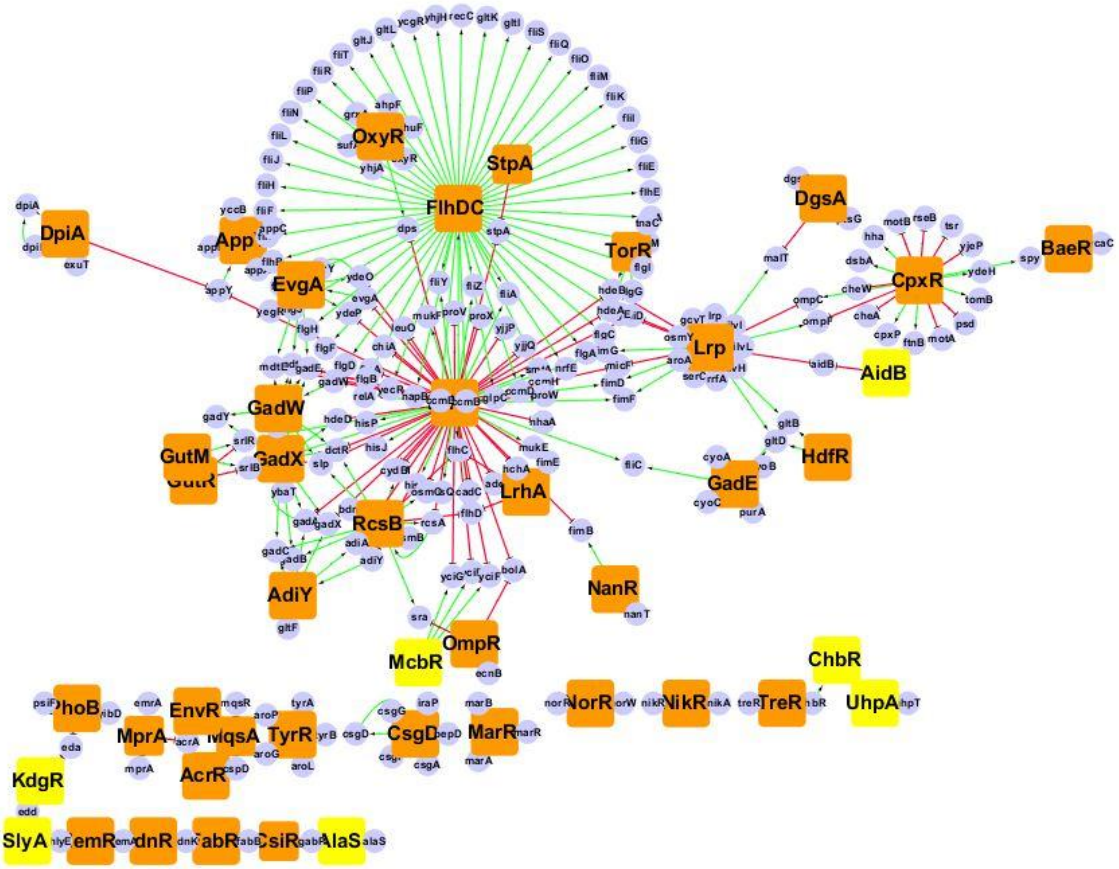
I. *crp* Knockout



J. *dps* Knockout Stationary



K. *dps* Knockout Stationary2



L. *hns* Knockout

APPENDIX D:

Key TFs under each MOPs medium condition

ERN network properties significantly changed effective TFs while experimental condition changes from control (MOPS medium, wild-type *E. coli* K-12, and Glucose carbon source) to other experimental conditions.

Key TFs of MOPS media experiments compare to the WT_MOPS_glucose, and their network properties					
Experimental Condition	**TF**	**Degree (Normalized)**	**Betweenness (Normalized)**	**Closeness (Normalized)**	**Network Size Normalized Degree**
WT_MOPS_acetate	GatR	1	0	1	0.0194
	GlpR	0.9	0	1	0.0290
	LsrR	0.9	0	1	0.0290
	MelR	1	0	0	0.0032
	HcaR	0.428571429	0.4	0.997716896	0.0097
	CRP	0.414248021	0	0.542213131	0.5065
WT_MOPS_glycerol	GlpR	0.9	0	1	0.0647
	HcaR	0.428571429	0.4	0.997716896	0.0216
	CRP	0.253298153	0	0.518017142	0.6906
WT_MOPS_proline	AlpA	1	0	1	0.0031
	GatR	1	0	1	0.0189
	KdgR	1	0	1	0.0063
	LldR	1	0	0	0.0031
	McbR	1	0	1	0.0094
	MelR	1	0	0	0.0031
	SlyA	1	0	1	0.0031
	HcaR	0.428571429	0.4	0.997716896	0.0094
	AidB	0.5	0	1	0.0031
	AlaS	0.5	0	1	0.0031
	PutA	0.666666667	0	1	0.0063
	CRP	0.364116095	0	0.534122359	0.4340
WT_MOPS_stationary	GlpR	0.9	0.888888889	0.999427263	0.0363
	KdgR	1	0	1	0.0081
	MelR	1	0	0	0.0040
	FabR	0.5	0	0.99942955	0.0040
	MurR	0.333333333	0	0.99942955	0.0040

	TF	Degree (Normalized)	Betweenness (Normalized)	Closeness (Normalized)	Network Size Normalized Degree
	NemR	0.666666667	0	1	0.0081
WT_MOPS_stationary2	EnvR	1	0	1	0.0039
	FeaR	1	0.285714286	1	0.0039
	GlrR	1	0	1	0.0019
	KdgR	1	0	1	0.0039
	LldR	1	0	0	0.0019
	McbR	1	0	1	0.0058
	MelR	1	0	0	0.0019
	MntR	1	0	1	0.0019
	UhpA	1	0	1	0.0019
	AidB	0.5	0	1	0.0019
	AlaS	0.5	0	1	0.0019
	IlvY	0.666666667	0	1	0.0039
WT_MOPS_stationary3	CreB	1	1	1	0.0017
	EnvR	1	0	1	0.0017
	EnvY	1	0	1	0.0017
	FeaR	1	0.285714286	1	0.0017
	McbR	1	0	1	0.0025
	MntR	1	0	1	0.0008
	RtcR	1	0	1	0.0017
	SlyA	1	0	1	0.0008
	YeiL	1	0	1	0.0017
	YqhC	1	0	1	0.0008
	AraC	0.75	0.461538462	0.99885649	0.0050
	CysB	0.72	0.472222222	0.995394688	0.0151
	HyfR	0.692307692	0.636363636	0.99770905	0.0076
	LrhA	0.8	0.530120482	0.977205659	0.0034
	NikR	0.571428571	0.6	0.998286694	0.0034
	TyrR	0.666666667	0.7	0.997710361	0.0067
	AidB	0.5	0	1	0.0008
	AlaS	0.5	0	1	0.0008
	IlvY	0.666666667	0	1	0.0017
	MtlR	0.75	0	1	0.0025
	RelEB	0.75	0	1	0.0025
WT_MOPS_stationary4	EnvR	1	0	1	0.0026
	EnvY	1	0	1	0.0026
	FabR	1	0	1	0.0026

	TF	Degree (Normalized)	Betweenness (Normalized)	Closeness (Normalized)	Network Size Normalized Degree
	FeaR	1	0.285714286	1	0.0026
	KdgR	1	0	1	0.0026
	MelR	1	0	0	0.0013
	MntR	1	0	1	0.0013
	RtcR	1	0	1	0.0026
	UhpA	1	0	1	0.0013
	YqhC	1	0	1	0.0013
	CysB	0.6	0.583333333	0.993678491	0.0196
	AidB	0.5	0	1	0.0013
	AlaS	0.5	0	1	0.0013
	IlvY	0.666666667	0	1	0.0026
	RelEB	0.75	0	1	0.0039
	MntR	1	0	1	0.0156
WT_MOPS_heatShock	YqhC	1	0	1	0.0156
	AlaS	0.5	0	1	0.0156
	IlvY	0.666666667	0	1	0.0313
	RelEB	0.75	0	1	0.0469
	H-NS	0.223602484	0	0.787836637	0.5625
	AlpA	1	0	1	0.0054
MOPS_K_crp	GatR	1	0	1	0.0323
	McbR	1	0	1	0.0161
	UhpA	1	0	1	0.0054
	BglJ	0.75	0	1	0.0161
	FabR	1	0	1	0.0022
MOPS_K_dps_stationary	FeaR	1	0.285714286	1	0.0022
	GlrR	1	0	1	0.0011
	KdgR	1	0	1	0.0022
	LldR	1	0	0	0.0011
	McbR	1	0	1	0.0034
	MelR	1	0	0	0.0011
	MntR	1	0	1	0.0011
	SlyA	1	0	1	0.0011
	YeiL	1	0	1	0.0022
	CysB	0.68	0.444444444	0.994821964	0.0191
	LrhA	0.8	0.578313253	0.979495536	0.0045
	NikR	0.428571429	0.4	0.997716896	0.0034
	TyrR	0.5	0.5	0.996569469	0.0067

	TF	Degree (Normalized)	Betweenness (Normalized)	Closeness (Normalized)	Network Size Normalized Degree
	AidB	0.5	0	1	0.0011
	AlaS	0.5	0	1	0.0011
	GlpR	0.9	0	1	0.0101
	IlvY	0.666666667	0	1	0.0022
	LsrR	0.9	0	1	0.0101
	MtlR	0.75	0	1	0.0034
	RelEB	0.75	0	1	0.0034
	EnvY	1	0	1	0.0043
	FeaR	1	0.285714286	1	0.0043
MOPS_K_dps_stationary2	MelR	1	0	0	0.0022
	MntR	1	0	1	0.0022
	TyrR	0.75	0.8	0.998281788	0.0194
	AidB	0.5	0	1	0.0022
	AlaS	0.5	0	1	0.0022
	RelEB	0.75	0	1	0.0065
	ChbR	1	0	0	0.0034
	KdgR	1	0	1	0.0068
MOPS_K_hns	McbR	1	0	1	0.0102
	SlyA	1	0	1	0.0034
	UhpA	1	0	1	0.0034
	AidB	0.5	0	1	0.0034
	AlaS	0.5	0	1	0.0034

APPENDIX E

Medium and Experiments Details

MOPS Medium

Recipe Substances:

Substances	Concentration	Role
β-D-glucose	20.0 g/l	Source of C
3-(N-morpholino)propanesulfonate	8.372 g/l	pH Buffer
tricine	0.717 g/l	pH Buffer
dipotassium phosphate	1.32 mg/l	Source of P
ammonium chloride	9.5 mM	Source of N
ammonium molybdate	3.6000001 µM	Source of N
borate	24.732 mM	
calcium chloride	50.0 nM	
cobalt chloride	7.2000003 µM	
copper sulfate	2.4000003 µM	Source of S
iron sulfate	0.01 µM	Source of S
manganese chloride	16.0 µM	
MgCl2	0.52500004 µM	
potassium sulfate	276.0 µM	Source of S
sodium chloride	50.0 mM	
zinc sulfate	2.8 µM	Source of S

Composition:

Constituents	Concentration
β-D-glucose	111.01 mM
chloride	59.55 mM
Na+	50.00 mM
3-(N-morpholino)propanesulfonate	40.01 mM
borate	24.73 mM
ammonium	9.51 mM
tricine	4.00 mM
K+	567.07 µM
sulfate	281.21 µM
Mn2+	16.00 µM
phosphate	7.53 µM
Co2+	7.20 µM
molybdate	3.60 µM
Zn2+	2.80 µM
Cu2+	2.40 µM
Mg2+	525.00 nM
Ca2+	50.00 nM
Fe2+	10.00 nM

pH: 7.2

Osmolarity (approximate, computed from constituents): 0.3 Osm/L

Wildtype growth observations:

T (°C)	O_2	Growth?
37	Aerobic	Yes

Experiments															
Experiment Name	M3D ID	acetate	aeration	Ciprofloxacin (mM)	culture_pH	culture temperature (Celsius)	glycerol (mM)	glucose (mM)	l-proline (mM)	growth phase	MOPS (mM)	perturbation	perturbation gene	strain	time point (minutes)
Glucose (Control)	WT_MOPS_glucose		aerobic	0	7.2	37	0	5.5506	0	log	40	N/A	N/A	MG1655	0
Carbon source: acetate	WT_MOPS_acetate	16.6519	aerobic	0	7.2	37	0	0	0	log	40	N/A	N/A	MG1655	0
Carbon source: glycerol	WT_MOPS_glycerol		aerobic	0	7.2	37	11.1012	0	0	log	40	N/A	N/A	MG1655	0
Carbon source: prolin	WT_MOPS_proline		aerobic	0	7.2	37	0	0	6.6607	log	40	N/A	N/A	MG1655	0
Late Log phase	WT_MOPS_lateLog		aerobic	0	7.2	37	0	5.5506	0	late_log	40	N/A	N/A	MG1655	90
Stationary	WT_MOPS_stationary		aerobic	0	7.2	37	0	5.5506	0	stationary	40	N/A	N/A	MG1655	135
Stationary 2	WT_MOPS_stationary2		aerobic	0	7.2	37	0	5.5506	0	stationary	40	N/A	N/A	MG1655	330
Stationary 3	WT_MOPS_stationary3		aerobic	0	7.2	37	0	5.5506	0	stationary	40	N/A	N/A	MG1655	480
Stationary 4	WT_MOPS_stationary4		aerobic	0	7.2	37	0	5.5506	0	stationary	40	N/A	N/A	MG1655	720
Heat Shock	WT_MOPS_heatShock		aerobic	0	7.2	50	0	5.5506	0	log	40	N/A	N/A	MG1655	N/A
Acid Shock	WT_MOPS_acidShock		aerobic	0	2	37	0	5.5506	0	log	40	N/A	N/A	MG1655	N/A
Cipro	WT_MOPS_cipro		aerobic	6.04E-05	7.2	37	0	5.5506	0	log	40	N/A	N/A	MG1655	N/A
Cipro 2	WT_MOPS_cipro2		aerobic	6.04E-05	7.2	37	0	5.5506	0	log	40	N/A	N/A	MG1655	N/A
Knockout: crp	MOPS_K_crp		aerobic	0	7.2	37	0	5.5506	0	log	40	knockout	crp	MG1655	N/A
Knockout: cspA	MOPS_K_cspA		aerobic	0	7.2	37	0	5.5506	0	log	40	knockout	cspA	MG1655	N/A
Knockout: dps	MOPS_K_dps		aerobic	0	7.2	37	0	5.5506	0	log	40	knockout	dps	MG1655	N/A
Knockout: dps stationary	MOPS_K_dps_stationary		aerobic	0	7.2	37	0	5.5506	0	stationary	40	knockout	dps	MG1655	N/A
Knockout: dps stationary 2	MOPS_K_dps_stationary2		aerobic	0	7.2	37	0	5.5506	0	stationary	40	knockout	dps	MG1655	N/A
Knockout: hns	MOPS_K_hns		aerobic	0	7.2	37	0	5.5506	0	log	40	knockout	hns	MG1655	N/A
Knockout: hupB	MOPS_K_hupB		aerobic	0	7.2	37	0	5.5506	0	log	40	knockout	hupB	MG1655	N/A