The Analysis of Metabolism in *Saccharomyces cerevisiae* with Genome-scale Gene Expression Data

HUI Sheng

A thesis submitted in partial fulfillment of the requirements
for degree of
Master of Philosophy

Principal Supervisor: Dr. TANG Lei Han
Hong Kong Baptist University
April 2005

Declaration

I hereby declare that this thesis represents my own work which has been done after registration for the degree of MPhil at Hong Kong Baptist University, and has not been previously included in a thesis, dissertation submitted to this or other institution for a degree, diploma or other qualification.

Signature: _______________

Date: April 2005

Abstract

The availability of whole cell scale data makes analysis possible at the systems level. In this thesis, we try to gain a better understanding of the yeast metabolism by analyzing the metabolic network with the aid of genome-scale gene expression data. The metabolic network is examined topologically, at both the global and local levels. The network shows scale free property and is intrinsically modular. The loop structure is a statistically significant motif and may play an important role in network dynamics. By simulating the *in silico* cell, we predict theoretically the metabolic flux and investigate its patterns. The network is shown to consist of a backbone structure and its functionality is proven optimal. It is also shown that the microarray gene expression data can be used to reveal the dynamic organization of the metabolic network. Topologically different sub-networks are utilized to respond to different internal or external living environments.

Acknowledgement

I would like to express my greatest appreciation to my supervisor, Dr Lei Han TANG, for his supervision and constant encouragement during my study. I would like to thank him for his many days' patient tutoring and showing us what good research constitutes.

I would like to thank my co-supervisor, Prof. Nai Ho CHEUNG, for his encouragement and help. I am also very grateful to Dr. Gang HU for helping me in numerous cases throughout all these years.

I am also thankful to the staffs in the department for their support. I am thankful to my group members, especially Shenghua LIANG, for their valuable suggestions and comments.

I would specially thank all my friends for everything they have done for me.

Finally, I would like to thank my family for many years' caring and support.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction and Thesis Outline

1.1 The Biology of Yeast

Saccharomyces cerevisiae, commonly known as baker's yeast or budding yeast, is one of the major model organisms that have been under intense study for many decades. The yeast *Saccharomyces cerevisiae* is a unicellular eukaryotic organism, existing in two cell types, the haploid cell and the diploid cell. The haploid cells of opposite type may mate to form a diploid. The diploid may sporulate to generate haploid spores of both a and α types. *Saccharomyces cerevisiae* contains a haploid set of 16 chromosomes, ranging in size from 200 to 2,200 kb. The complete chromosomal genome is of 12,052 kb, released in April, 1996. A total of 6,183 ORFs of over 100 amino acids long were reported. Approximately 30% of the genes already have been characterized experimentally. About half of the remaining 70% ORFs either contain a motif of a characterized class of proteins or correspond to proteins that are related to functionally characterized gene products.

Basic biology on yeast is covered in most textbooks. We name a few of the popular ones here: *Biochemistry* (Mathews *et al.* 2000), *The World of the Cell* (Becker *et al.* 2000), *Genes VIII* (Lewin 2003), and *Molecular Biology of the Cell* (Alberts *et al.* 2002). Three online databases contain comprehensive information on yeast and they are SGD (http://www.yeastgenome.org/), CYGD (http://mips.gsf.de/genre/proj/yeast/) and YPD (http://www.incyte.com/control/tools/proteome).

Proteins are a major component in the cellular machinery. They have diverse functions ranging from catalyzing biochemical reactions, regulating transcription and translation, transmitting signals, transporting metabolites, to serving as structural components. A detailed list of the function distribution of yeast proteins can be found in Appendix 1. Proteins function at their specific locations inside the cell and the full list is in Appendix 2.

1.2 The Systemic Approach to Yeast Biology

Decades of painstaking work has accumulated a relatively complete set of biochemical reactions for yeast. These reactions tell how metabolites are inter-converted and describe the physiology of the cell. The so-called metabolic network is an equal representation of the reaction set. The analysis of such a network at the global level belongs to the newly emerging research field of systems biology. Recently high-throughput experimental techniques have been producing unprecedented large amount of data, among which the genome-scale microarray expression data are of our interest. The microarray experiment measures expression levels of virtually all genes of an organism simultaneously. The data are intrinsically of systemic nature and are subject to systemic analysis.

1.2.1 Work Done by Others

The work done by others can be generally classified into three categories. First there is the analysis of the topology of cellular networks. Both top-down and bottom-up approaches have been taken in the study of network topology. The work by Barabasi and coworkers (Barabasi *et al.* 1999; Farkas *et al.* 2003; Jeong *et al.* 2000; Ravasz *et al.* 2002) represents the first approach, where statistical characteristics of various networks are collected and analyzed at the global scale. They discovered the so-called scale-free properties of many cellular networks, which since become a fashionable topic in many fields of science and engineering. Alon and coworkers (Shen-Orr *et al.* 2002; Milo *et al.* 2002; Milo *et al.* 2004) take the other approach, focusing on the network motifs, defined as recurring basic network units. The regulatory network motifs identified by them are argued, both theoretically and experimentally, to be dynamically significant.

The second category is the interference of genetic regulatory network from high-throughput data (D'haeseleer *et al.* 2000; de la Fuente *et al.* 2002). This approach follows the tradition of bioinformatics, a field developed and thriving since the early 1990's. Various algorithms have been developed to mine 'meaningful' information from large data sets. The identification of regulatory motifs from genome sequences is one of the major study foci (Bussemaker *et al.* 2001; Wang *et al.* 2002; Chen *et al.* 2004; Kato *et al.* 2004). Another focus is the identification of functional modules from genome expression data (Ihmels *et al.* 2002; Ihmels *et al.* 2003).

The third category is the *in silico* modeling and simulation of the yeast cell. The work by Palsson and coworkers (Duarte *et al.* 2004) is the major force behind this way of understanding biology at the systemic level. The metabolic network is constructed by connecting all known biochemical reactions for a certain organism. Linear programming is used to optimize the biomass production under the steady state assumption. The solution yields quantitative information about the metabolic flux through each reaction. It is noted that systems biology is still in its infancy. Basic principles and main methodologies are yet to be formulated. A collection of manifestos can be found in the March 2002 issue of Science.

1.2.2 Our Goals and Strategies

The ultimate goal we are aiming at is a quantitative understanding of the yeast metabolic network. To achieve this goal, we characterize the network from several aspects. The network topology is of our first concern. Statistical measures are to be taken to reveal special properties of the network. With the genome-scale gene expression data, the investigation of network dynamics is attempted. An *in silico* model of yeast is constructed to simulate the cell's physiology in steady state. The fluxes obtained from the simulation are then subject to careful study.

1.3 Organization of Thesis

We will introduce the yeast metabolic network and the microarray expression data in Chapter 2 and Chapter 3, respectively. Chapter 4 is devoted to the analysis of the metabolic network in combination with the expression data. Chapter 5 continues the analysis, now focusing on an important concept, the network module. Both the searching algorithm and the results are to be discussed. An alternative way of network analysis, still of systemic nature, is presented in Chapter 6. Based on the iND750 model, we simulate the cell growth with the Flux Balance Analysis (FBA) and analyze the flux patterns generated by the model. In Chapter 7, we identify the network motifs and discuss their possible functions. Finally in Chapter 8, we summarize the whole thesis and provide prospects on future work.

Chapter 2

Yeast Metabolic network

2.1 Construction of Metabolic Network

The raw metabolic data collected at KEGG (Kanehisa 1997; Kanehisa *et al.* 2000; http://www.genome.jp/kegg/) consist of a detailed list of biochemical reactions. Besides annotations for genes and genomes, KEGG contains comprehensive information on biochemical reactions, enzymes, and pathways. Each reaction is assigned a unique reaction number. From this data, one may construct a network by making connections between metabolites (both substrates and products) and the reaction they participate in. In the actual graph representation, 'arrows' or 'arcs' are drawn from substrates to reactions and from reactions to products. For reversible reactions, 'edges' with no directions are used instead to represent the connections. For *S. cerevisiae*, which is the focus of present study, the resulting network consists of 1007 reactions and 1037 metabolites, with 1954 arcs and 2354 edges. In the unorganized form, the network is too complex to be presented for visual inspection. However, we are able to do some simple statistics on it.

2.2 Scale free Network

Barabasi and coworkers (Jeong *et al.* 2000) discovered the scale free property of metabolic networks. Following their work, we define the connection degree of a metabolite as the number of reactions connected to the metabolite. We calculated the connection degrees for all the 1037 metabolites and it displays a power law distribution, with exceptions for the tail part (Figure 2.1). Networks with a power law distribution of connection degrees are called scale free networks. The tail part suggests presence of a few highly connected nodes in our data set. They have been termed currency metabolites (Table 2.1).

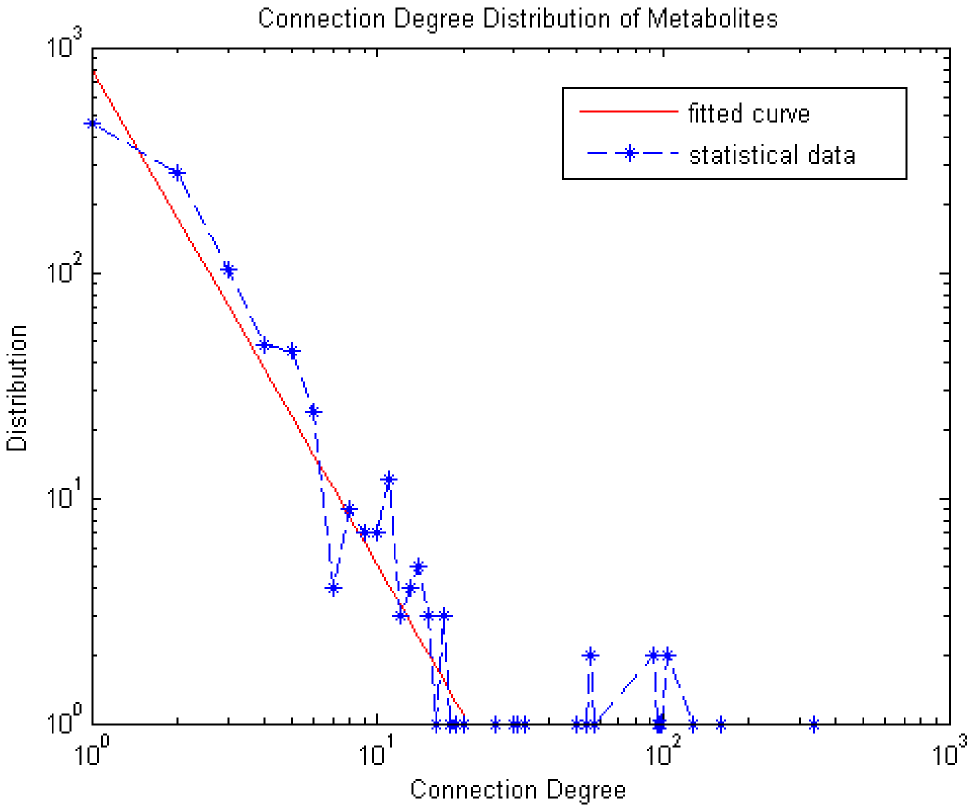

Figure 2.1: The connection degree distribution of metabolites. The fitted curve has a power of -2.2 .

H2O	Orthophosphate	NADH	NADPH	NH3
ATP	ADP	Pyrophosphate	CO2	
H+	NAD+	NADP+	AMP	

Table 2.1: The 13 most connected metabolites.

2.3 Hierarchical Modular Network

2.3.1 Definition of Clustering Coefficient

The clustering coefficient is a measure of the interrelatedness of the local neighborhoods. For a node i with k_i immediate neighbors, its clustering coefficient is given by

$$C_i = \frac{2N_i}{k_i(k_i - 1)},$$

where N_i is the number of links among the k_i nodes. Note that $k_i(k_i - 1)/2$ is the largest possible number of connections among the neighbors. The value of the clustering coefficient is 1 when the neighbors are maximally linked and zero when no links among them (Figure 2.2).

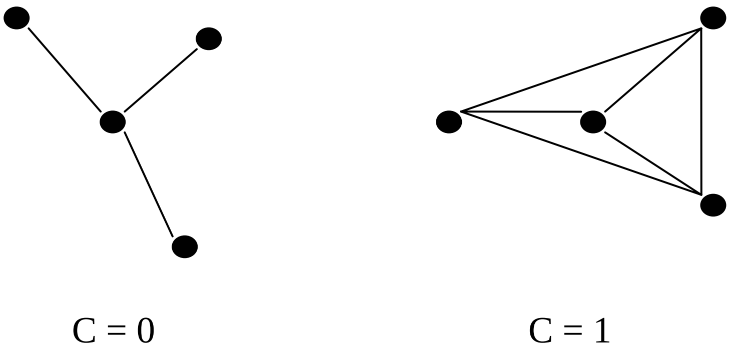

Figure 2.2: The clustering coefficient

The clustering coefficient for the whole network is taken as the mean of clustering coefficients of all nodes.

2.3.2 Clustering Coefficient of Random Networks

The metabolic network consists of two types of nodes, the reaction nodes and the metabolite nodes. A link can only occur between different types of nodes. The immediate neighbors of a node are of the same type and no links among them. This kind of network always has zero value of clustering coefficient. To measure the interrelatedness of network modules, we transform the original network into a network of only reaction nodes. In this new network, a link is established between any two reactions that are connected to at least one common metabolite in the original network.

Here we calculate the clustering coefficients for random networks of different sizes. A random network is selected by starting from a random node. One additional node is added by randomly choosing one node from the nodes that have connections to the nodes already included in the random network. The procedure repeats until the size of the random network reaches a specified value. For a given network size, we generate ten random networks and an average clustering coefficient is calculated for this size.

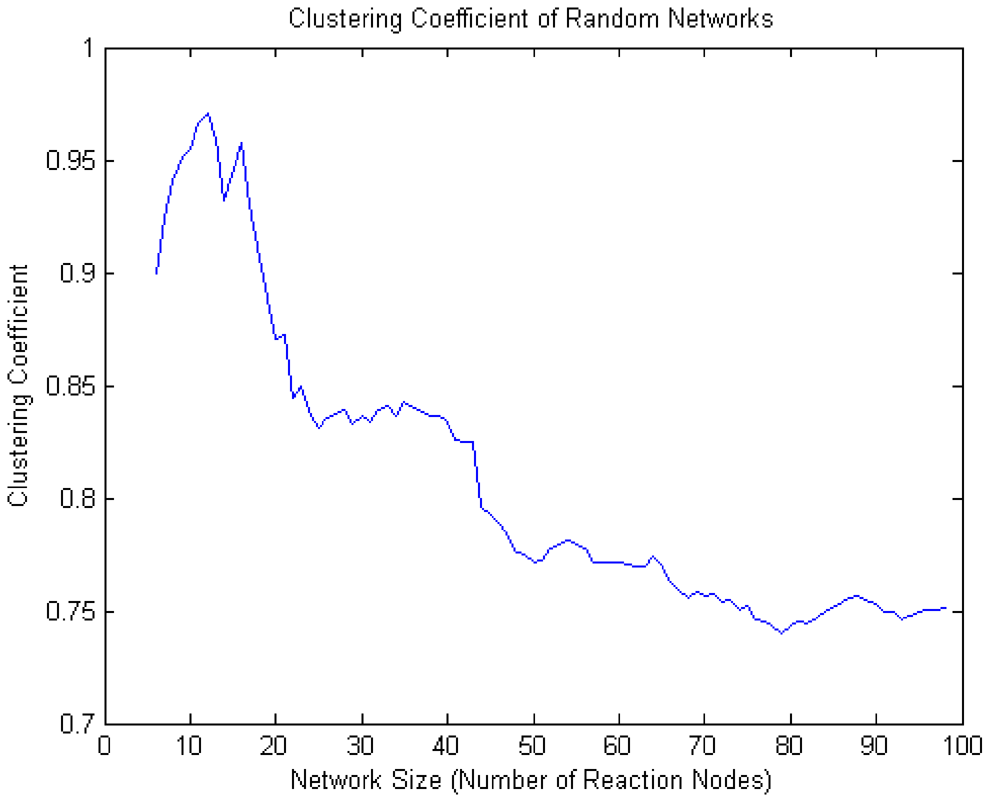

Figure 2.3: The clustering coefficients of random networks.

Figure 2.3 shows two dramatic changes of clustering coefficient as the network size increases gradually. The first such transition occurs at about 20 and the second at about 50. This phenomenon is best explained if the network has hierarchical modular structure.

2.4 Network with Main Metabolites

In most biochemical reactions, only one or two substrates and products are considered to be the ‘main’ metabolites, while others may serve as co-factors. The main metabolites are mostly carbon-carrying metabolites and may be later consumed for energy or transformed to certain products. For example, in the glycolysis reaction ATP + Pyruvate ⇔ ADP + Phosphoenolpyruvate (PEP), the carbon flow is between two main metabolites: Pyruvate and PEP, while ATP and ADP are only facilitating the flow by providing the necessary energy. The complexity of the network is significantly reduced while non-main metabolites are removed. However, the network is still quite involved (Figure 2.4).

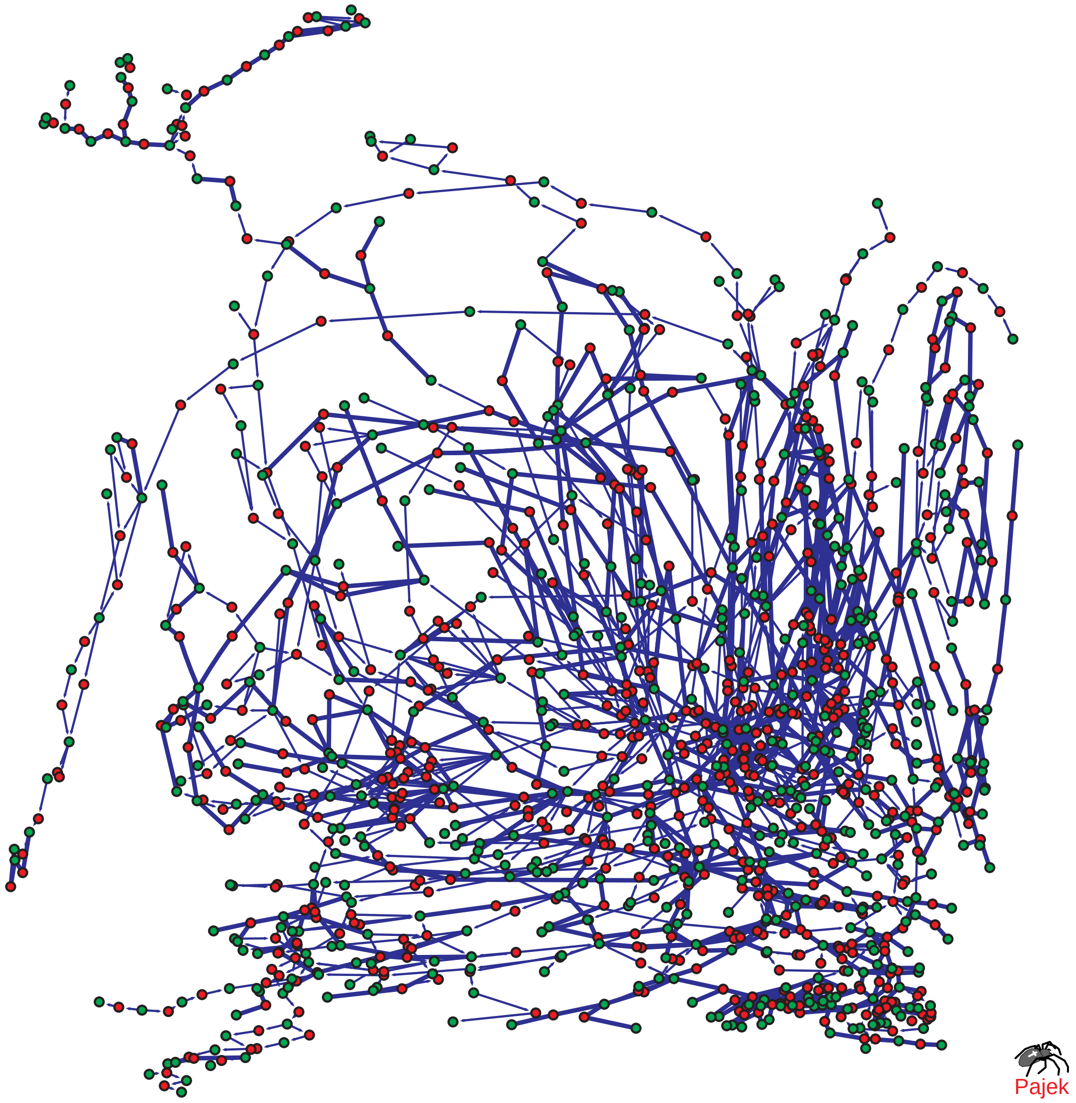

Figure 2.4: The network with only main metabolites. The graph is drawn with Pajek (Batagelj & Mrvar 1998).

2.5 Biochemical Pathways

Traditional biology has accumulated a large amount of knowledge on metabolism, among which is the concept of biochemical pathways. Biochemical pathway is a set of enzyme-catalyzed reactions that are closely related in such a way that the products of a reaction are the substrates of another reaction. A biochemical pathway is regarded as an organizational unit and all the reactions within it act together to achieve a biochemical function. For example, the TCA cycle (or Citric Acid Cycle) is a series of reactions that aerobic organisms use to release energy stored in acetyl-CoA, pyruvate and PEP(phosphoenolpyruvate) (Figure 2.5).

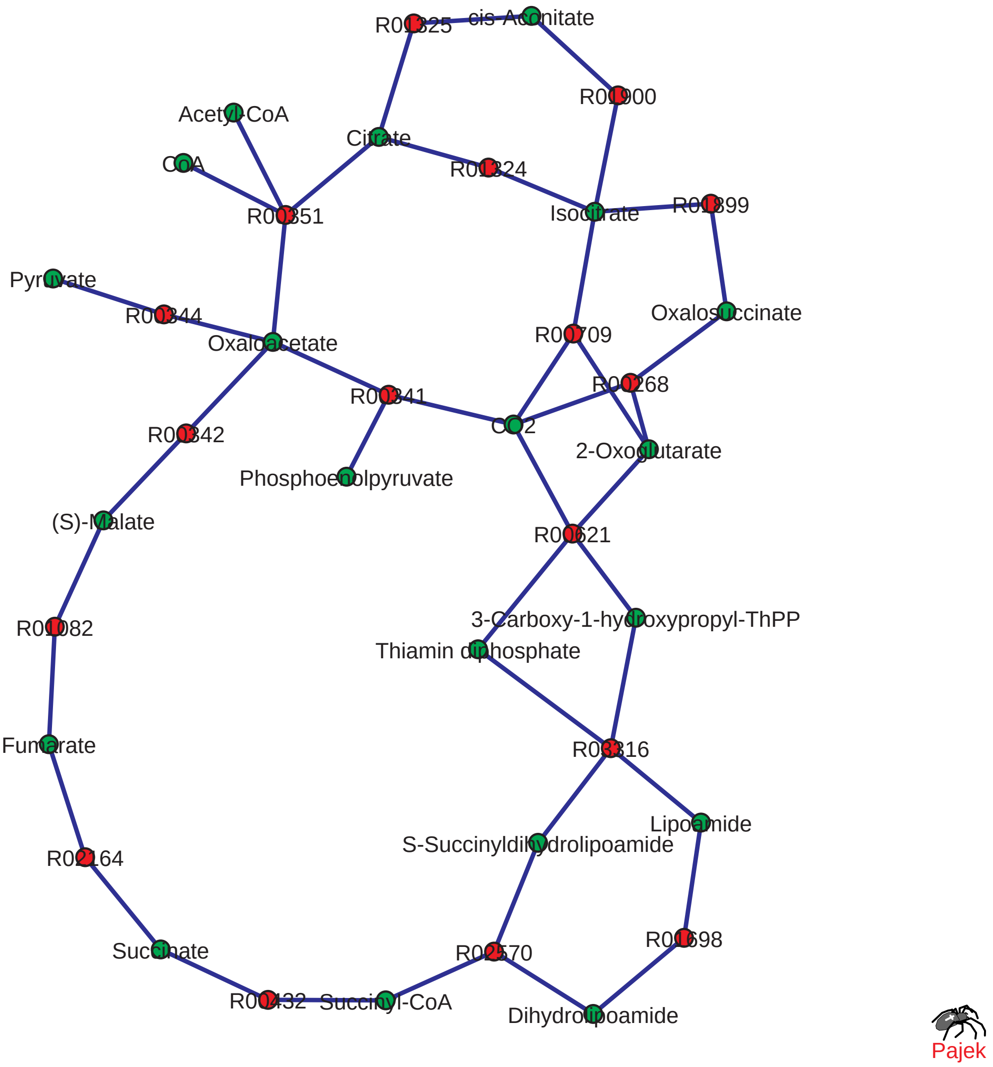

Figure 2.5: The biochemical pathway of TCA cycle.

The metabolic network shows how metabolic flux flows among the pathways.

Here we create a network of pathways by drawing connections between pathways that share one or more compounds.

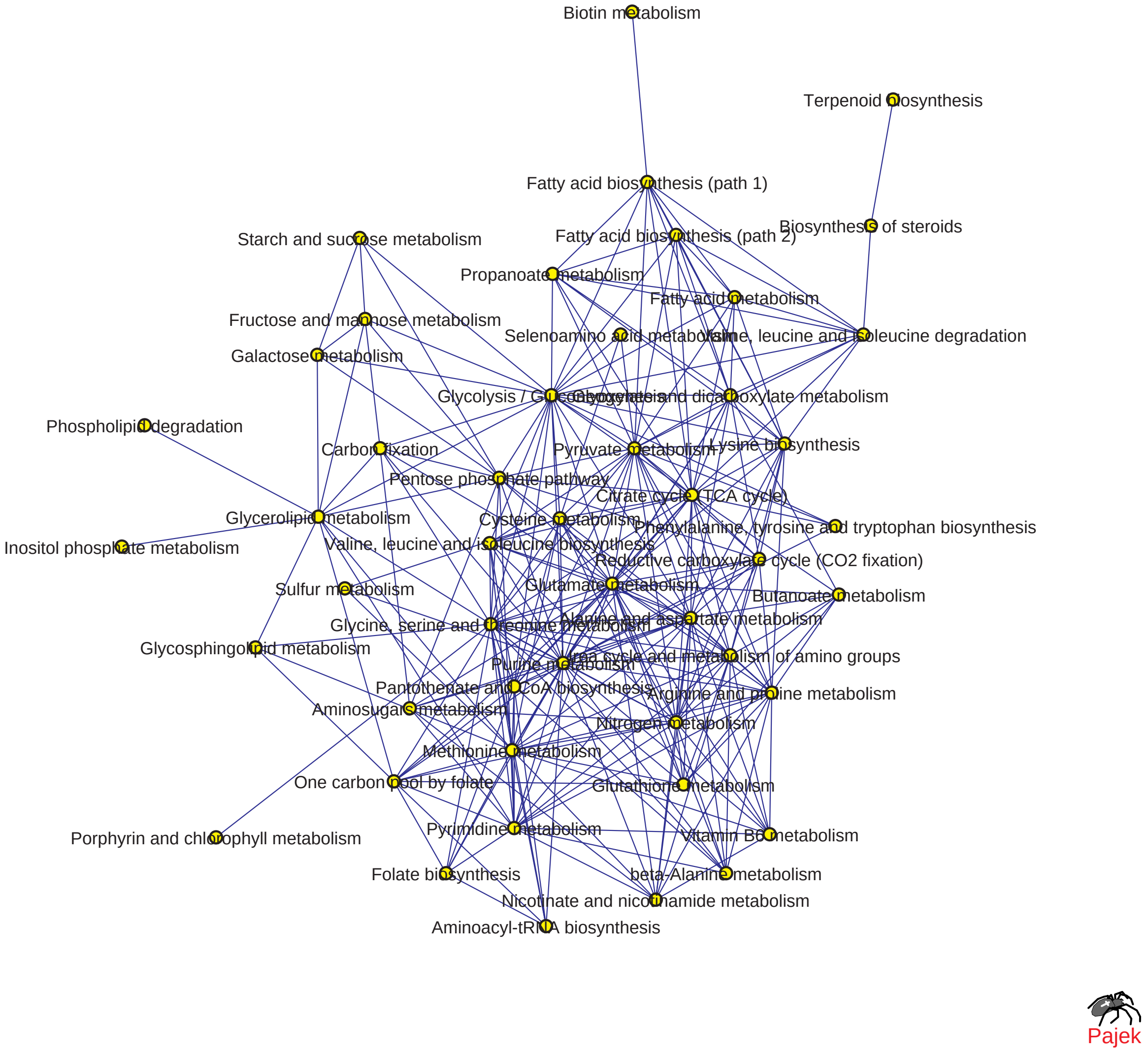
Biotin metabolism
Terpenoid biosynthesis
Fatty acid biosynthesis (path 1)
Starch and sucrose metabolism
Fatty acid biosynthesis (path 2)
Biosynthesis of steroids
Propanoate metabolism
Fatty acid metabolism
Fructose and mannose metabolism
Selenoamino acid metabolism
Valine, leucine and isoleucine degradation
Galactose metabolism
Glycolysis / Gluconeogenesis
Glyoxylate and dicarboxylate metabolism
Phospholipid degradation
Carbon fixation
Pyruvate metabolism
Lysine biosynthesis
Pentose phosphate pathway
Citrate cycle (TCA cycle)
Glycerolipid metabolism
Cysteine metabolism
Phenylalanine, tyrosine and tryptophan biosynthesis
Inositol phosphate metabolism
Valine, leucine and isoleucine biosynthesis
Reductive carboxylate cycle (CO2 fixation)
Glutamate metabolism
Sulfur metabolism
Butanoate metabolism
Glycine, serine and threonine metabolism
Alanine and aspartate metabolism
Glycosphingolipid metabolism
Urea cycle and metabolism of amino groups
Purine metabolism
Pantothenate and CoA biosynthesis
Arginine and proline metabolism
Aminosugars metabolism
Nitrogen metabolism
Methionine metabolism
One carbon pool by folate
Glutathione metabolism
Porphyrin and chlorophyll metabolism
Pyrimidine metabolism
Vitamin B6 metabolism
Folate biosynthesis
beta-Alanine metabolism
Nicotinate and nicotinamide metabolism
Aminoacyl-tRNA biosynthesis
Pajek

Figure 2.6: A network of pathways.

This network consists of 48 yeast metabolic pathways, with small pathways that have less than 5 reactions not included. The nodes are highly connected to each other, with an average number of connections of 11.3 and standard deviation of 7.5. The size of the node is proportional to its number of nodes it connects to. The eight pathways that have more than 20 connections to other pathways are Glycolysis / Gluconeogenesis, Citrate cycle (TCA cycle), Purine metabolism, Glutamate metabolism, Alanine and aspartate metabolism, Glycine, serine and threonine metabolism, Pyruvate metabolism, and Nitrogen metabolism. The Glycolysis / Gluconeogenesis, Citrate cycle (TCA cycle), and Pyruvate metabolism pathways provide most precursor metabolites to other pathways and thus are highly connected. The amino acids glutamate, alanine, aspartate and serine act like some kind of currency metabolites and are final products or starting substrates for many other pathways. Nitrogen is an important element indispensable to many pathways including amino acids metabolism and nucleotide metabolism. Purine metabolism is a big pathway involving many reactions and is inevitable to be highly connected.

Chapter 3

Expression Data

3.1 Introduction to Microarray Data

Here in this thesis we mainly use the cDNA microarray data. The genome-scale cDNA microarray experiment measures the expression differences under two different conditions, the control and the test conditions, for virtually all genes of an organism. An expression ratio is obtained for each gene in an experiment. The ratio tells how the gene's expression is tuned to respond to the change of conditions, which could be the change of external environment or the deletion of one or more genes. The simultaneous availability of the expression ratios of all the genes in a genome-scale is of great value and presumably contains precious information on the underlying cellular regulatory mechanisms in response to the environmental stress or alteration of its internal organization.

3.2 Preprocessing of Expression Data

Our data set is compiled from the Stanford Microarray Database (http://genome-www5.stanford.edu/). After some selections we finally have a data set of 6126 yeast genes for 990 experiments. As a convenient way of data analysis, log-ratios are used instead of normal ratios. The data are also normalized such that for each experiment the mean log-ratio is 0 and the standard deviation is 1. Figure 3.1 shows a portion of the raw data set using color coding of the expression value (bar on the right).

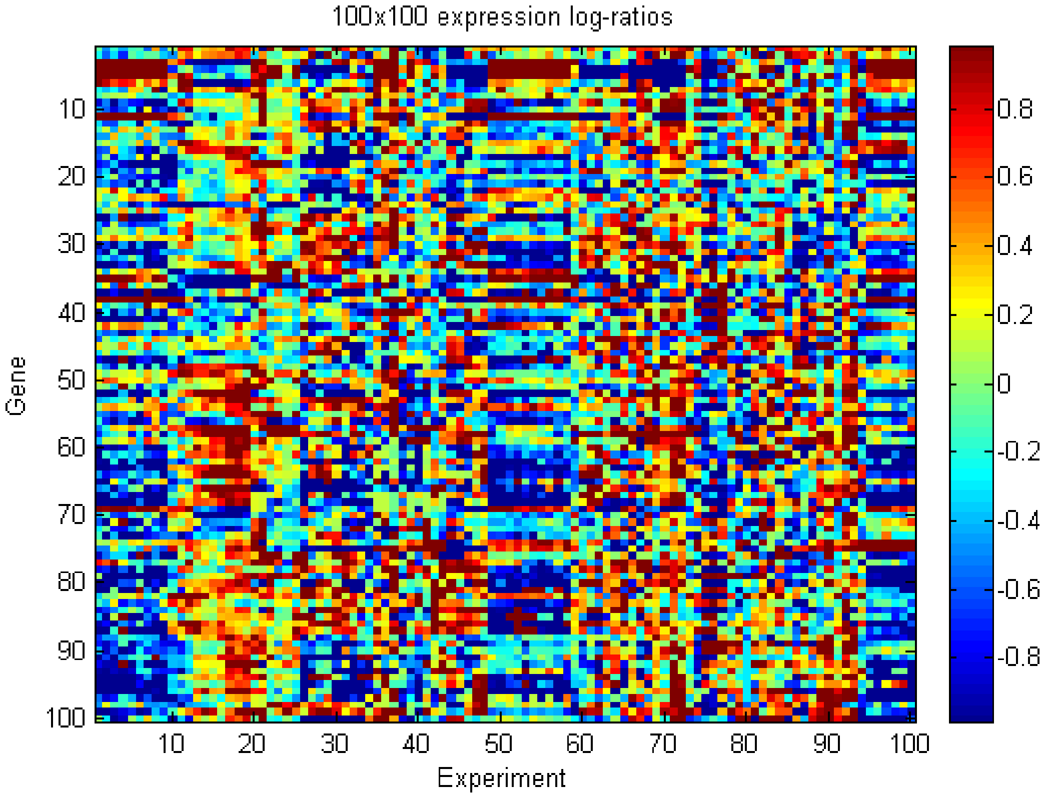

Figure 3.1: A portion of raw microarray data.

3.3 Overview of Expression Data Analysis

3.3.1 Condition-independent clustering

A simple-minded and popular way of analyzing the genome-scale gene expression data is to cluster the genes, where each gene is represented by a vector in a space whose dimension equals the number of experiments. The initial excitement generated by the papers using hierarchical clustering (Michaels *et al.* 1998; Eisen *et al.* 1998) and SOM (Tamayo *et al.* 1999; Toronen *et al.* 1999) lead to a large number of papers on fast and robust clustering algorithms (Ben-Dor *et al.* 1999; Sharan *et al.* 2000; Sasik *et al.* 2001; Heyer *et al.* 1999) Presumably genes within the same group are in some way related functionally (guilty by association). The results can then be verified experimentally. Along this line of thinking, algorithms are developed to uncover even more complicated relations among genes, forming the so-called regulatory network. This kind of analysis presumes the gene-gene interactions remain unchanged under all conditions. However, this is hardly the case since many genes have multi-functions, participating in different pathways under different conditions.

3.3.2 Condition-dependent clustering

A more advanced way of grouping genes is to cluster the conditions, together with the genes, the so-called biclustering (Cheng & Church; Madeira & Oliveira). A term that is more proper to describe what we're looking for here is called module. A module is defined by both a set of genes and a set of conditions (Figure 3.2). Different modules may have overlapping genes or conditions. Barkai and coworkers (Ihmels *et al.* 2003) identified some 80 such modules for yeast.

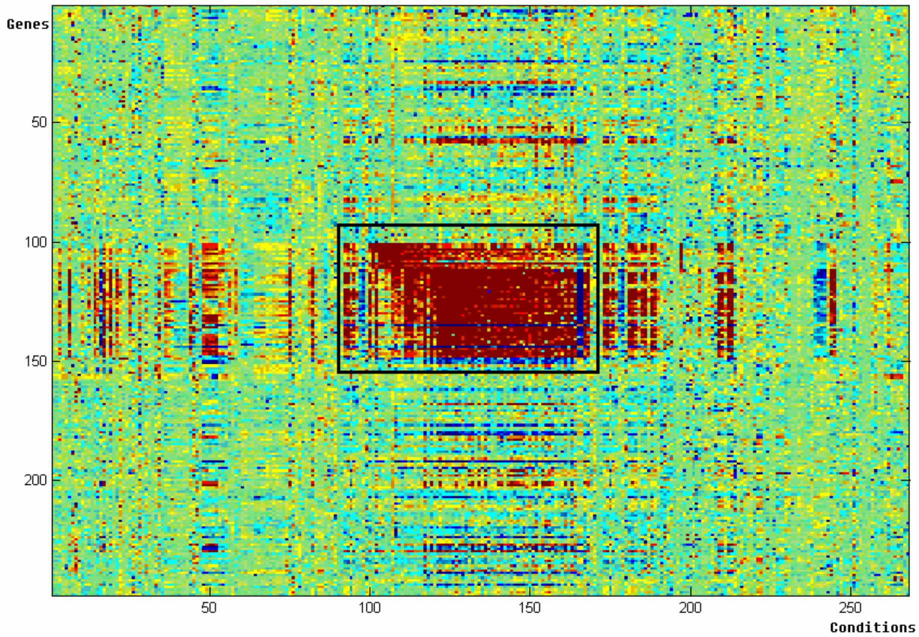

Figure 3.2: A condition-dependent cluster.

Chapter 4

Analysis of the Network with Expression Data

4.1 Integration of Network and Expression Data

The reaction-compound network is a network of compounds and reactions, while the expression data are all about genes. To combine the two types of data, we note that the reactions connecting compounds are catalyzed by enzymes, which of course are encoded by genes. A reaction is active if the genes coding for the enzymes associated with the reaction are up-regulated. This connection between the two types of data enables us to map the expression data onto the metabolic network. The number of genes that are involved in catalyzing biochemical reactions is 826. The less number of genes than reactions arises naturally from the fact that some enzymes catalyze more than one reaction.

In this chapter, we mainly address three questions: (1) how the network is activated; (2) how coherent biochemical pathways are; and (3) how network responds under different conditions.

The metabolic network we have is a static description of the cell physiology, which may not reflect the cell state under one particular condition. Here we are interested in the sub-network activated under each particular condition.

4.2 Connectivity

4.2.1 Definition of Connectivity

For a metabolic network with N nodes, its connectivity is defined as (Barthelemy *et al.* 2003)

$$C = \sum_{i=1}^{n} \left(\frac{m_i}{N} \right)^2 ,$$

where n is the number of clusters and m_i is the number of nodes in the *ith* cluster. A cluster is a set of nodes such that the distance between any two nodes is finite. It can be shown that the connectivity is 1 if all the network nodes form one cluster and C has the minimum value of $\frac{1}{N}$ if all the nodes are disconnected.

4.2.2 Connectivity of Genetically Activated Network

We calculate the connectivity for sub-networks activated under each condition. The activated network consists of reactions that are catalyzed by genes that have large expression ratios. To make the analysis statistically rigorous, for each condition, a total number of 12 sub-networks are produced, corresponding to 12 sets of genes ranging from the most expressed 10 to 120 genes. A mean value of connectivity for sub-networks corresponding to the same number of activated genes is obtained by averaging over all the 990 conditions. The significance of this measure of genetically activated networks is exhibited when compared with randomly activated networks. A randomly activated network is generated by first randomly (instead of referring to microarray data) selecting the top expressed genes.

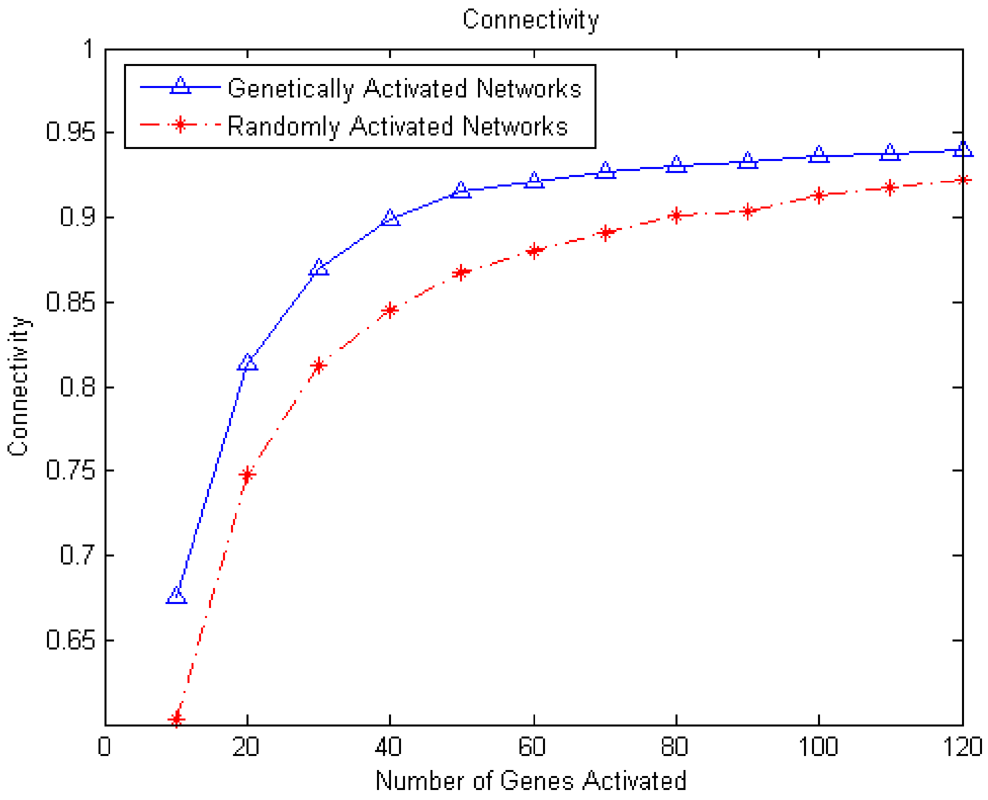

Figure 4.1: The comparison of connectivity of genetically activated networks with randomly activated networks.

The comparison (Figure 4.1) shows that at each snapshot the cell activates a well connected sub-network.

4.3 Internal Structures of Pathways

Biochemists have already divided the metabolic network into a number of groups which are called biochemical pathways. A full list of such pathways for yeast *Sacchromyces cerevisiae* is listed at Appendix 3. Each group carries out a specific function and genes within it are functionally closely related. For example, the lysine biosynthesis pathway depicts the steps that lead to the production of the amino acid lysine. Here we are interested in possible internal structures of such pathways.

With expression data, we are able to measure how coherently the genes within a pathway are expressed. The coherence of a pathway is taken as the average value of correlations between any two gene vectors within the pathway, with each condition as an observation. The correlation between two genes is calculated by the covariance function which is defined as

$$C_{ij} = \frac{\sum_{k=1}^{N}(G_{ik} - \bar{G}_i)\bullet(G_{jk} - \bar{G}_j)}{N}$$, where G_i and G_j are the gene vectors of gene i and j, respectively, and N is the number of conditions in the gene vector.

For each pathway, a second coherence is calculated for a subset of genes and conditions. The subset is selected by first choosing the top 10% 'active' conditions. The larger the variance of expression ratios of genes under a condition, the more active the condition. Then the top 30% 'active' genes are selected under the chosen top 10% 'active' conditions. Similarly, a gene is more active if the variance of its expression ratios is larger. The selected subset represents the core of the pathway and its coherence is compared with the coherence for the whole pathway (Table 4.1).

The comparison between the two coherence measures is best viewed with a scatter plot (Figure 4.2).

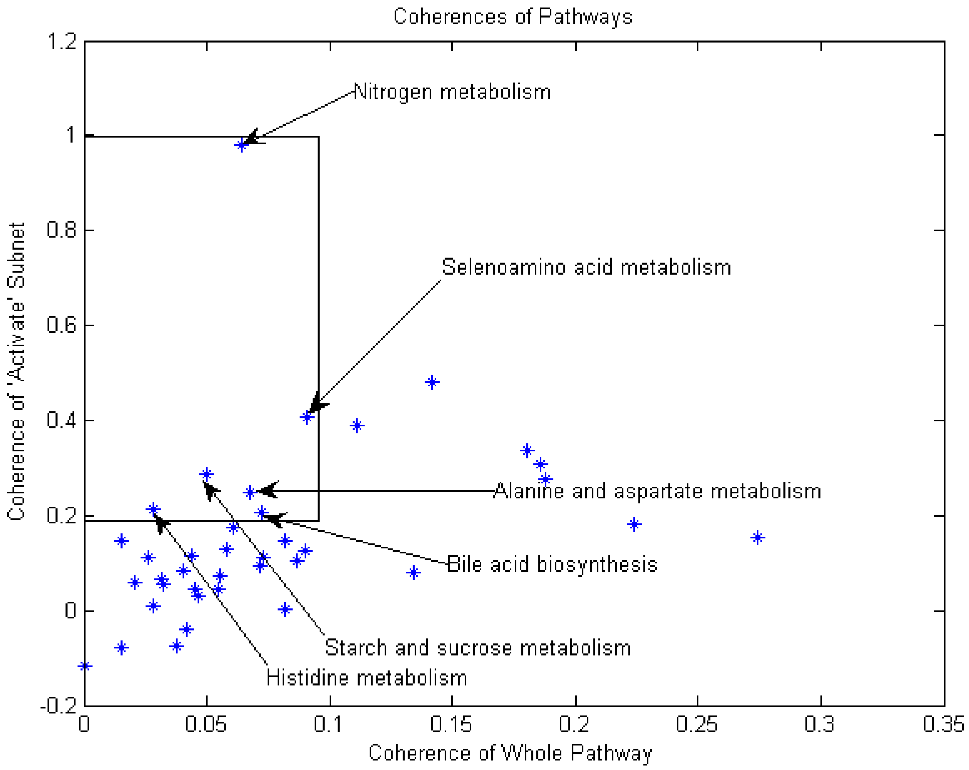

Figure 4.2: The coherences of pathways.

Pathway	Coherence of whole pathway	Coherence of 'active' subset
ATP synthesis	0.274525	0.152430
Valine, leucine and isoleucine biosynthesis	0.224259	0.180030
Citrate cycle (TCA cycle)	0.187592	0.277291
Oxidative phosphorylation	0.185859	0.306662
Glycolysis / Gluconeogenesis	0.180358	0.336866
Aminoacyl-tRNA biosynthesis	0.141385	0.480745
Phenylalanine, tyrosine and tryptophan biosynthesis	0.134067	0.081079
Galactose metabolism	0.110988	0.388516
Selenoamino acid metabolism	0.091038	0.407793
Pyrimidine metabolism	0.090149	0.125888
Pyruvate metabolism	0.086691	0.104087
N-Glycan biosynthesis	0.082091	0.001663
Purine metabolism	0.081809	0.145136
Carbon fixation	0.073007	0.110642
Bile acid biosynthesis	0.072376	0.206073
Lysine biosynthesis	0.071441	0.094757
Alanine and aspartate metabolism	0.067901	0.249354
Nitrogen metabolism	0.064497	0.981105
Pentose phosphate pathway	0.061125	0.173957
Butanoate metabolism	0.058344	0.129734
Glutamate metabolism	0.055127	0.072294
Fatty acid metabolism	0.054934	0.045056
Starch and sucrose metabolism	0.049627	0.288939
Glycosylphosphatidylinositol(GPI)-anchor biosynthesis	0.046884	0.028798
Arginine and proline metabolism	0.045517	0.042904
Fructose and mannose metabolism	0.043903	0.115142
Aminosugars metabolism	0.042060	-0.040296
Glycine, serine and threonine metabolism	0.040375	0.083125
Glycosphingolipid metabolism	0.037884	-0.073516
Lysine degradation	0.032220	0.053824
Glycerolipid metabolism	0.031745	0.064178
Histidine metabolism	0.028463	0.213006
Inositol phosphate metabolism	0.028112	0.010478
Benzoate degradation via CoA ligation	0.026282	0.111457
Nicotinate and nicotinamide metabolism	0.020474	0.057652
Tyrosine metabolism	0.015410	-0.077056
Folate biosynthesis	0.015083	0.147790
Tryptophan metabolism	0.000596	-0.117438

Table 4.1: The coherence of pathways.

Each of the six pathways inside the rectangle has low coherence for the whole pathway and high coherence for only a subset of it. This phenomena can be explained when these pathways have relatively independent internal structures. While the interrelatedness is high inside each of the internal structures, the overall coherence of the whole pathway may be low due to low dependence between these structures.

4.4 Network Dynamics

We try to get some glimpses of the network dynamics by investigating how the network responds to the change of conditions. We base our analysis on four distinct categories of conditions, namely, cell cycle, DNA damage, diauxic shift, and stress response.

4.4.1 Extent of Activation

We are interested to know how big the part of network that is activated in different conditions. The network size is a direct measure of the extent of activation. For a given threshold and a particular condition, the size of activated network is calculated by simply counting the number of nodes in the metabolic network corresponding to the genes with expression ratios above the threshold. The size for a condition category is then obtained by averaging over all the conditions in that category. A series of thresholds are taken and the results shown below (Figure 4.3).

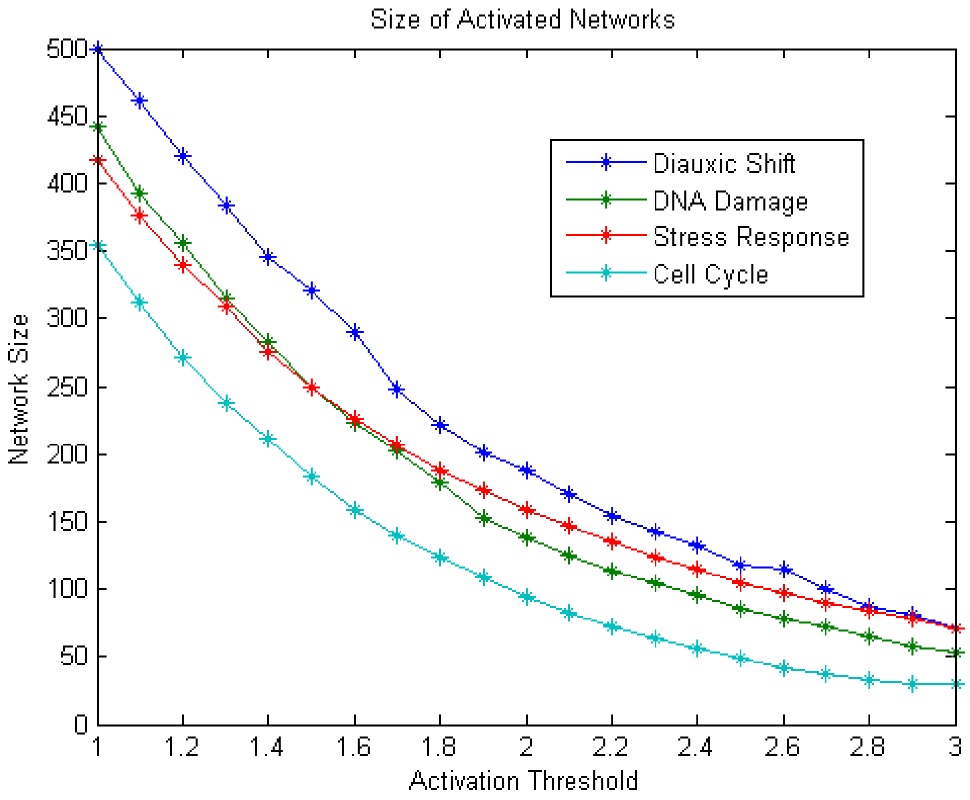

Figure 4.3: The size of activated networks under four categories of conditions.

The comparison may have the following several implications: the large network activated under diauxic shift may be explained that the cell undergoes a major mechanism change to accommodate the dramatic environmental shift; the small network of cell cycle makes sense since cell cycle involves only a part of the metabolic network; and the medium size networks of DNA damage and stress response reflect that the cell responds moderately to relatively small perturbations.

4.4.2 Connectivity of Activated Networks

We next measure the connectivity for networks activated under the four categories of conditions (Figure 4.4). The same number of top expressed genes is selected for each condition and the connectivity is calculated for the corresponding metabolic network. The mean value of connectivity is then computed for a category of conditions by averaging the connectivity of all the conditions in that category. The number of selected genes ranges from 10 to 60, increased by 2 at each step.

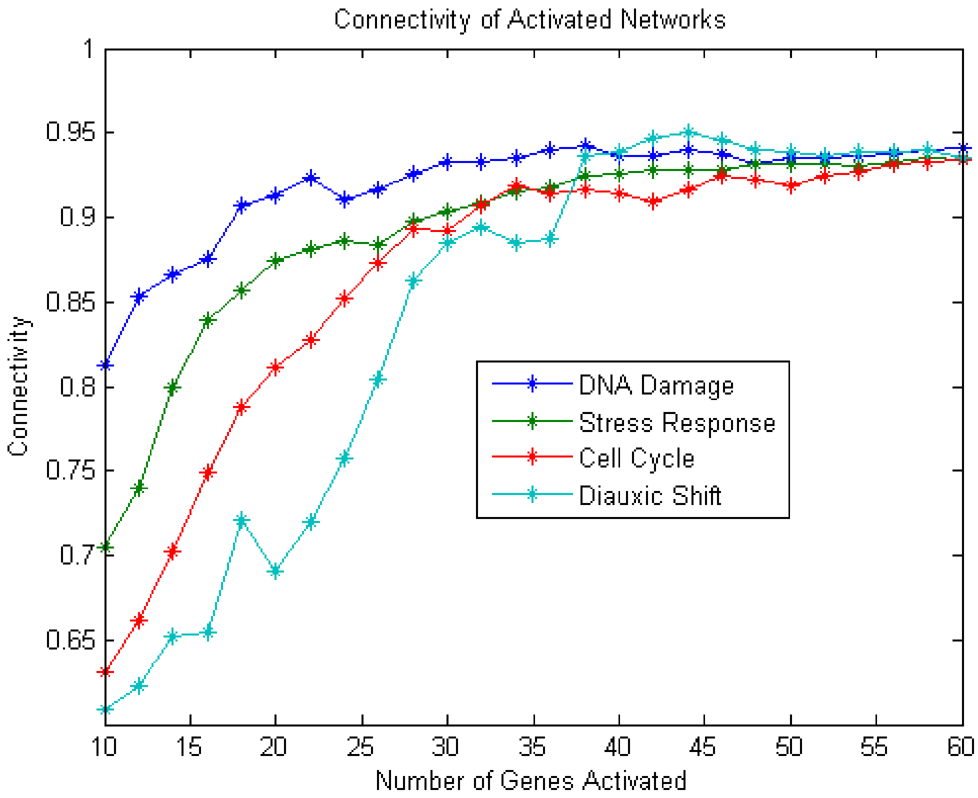

Figure 4.4: The connectivity of networks activated under four conditions.

The well-connected networks of DNA damage and stress response are expected because a part of network functioning for a common specific purpose is utilized to rescue the cell from damages or stringent environments. The highly specific nature of DNA damage explains the higher connectivity of its corresponding network. The diauxic shift involves too many genes which may be functionally remotely related. This gives poor connectivity of the network of diauxic shift.

Chapter 5

Network Modules

5.1 Definition of Network Module

Previous gene clustering analysis focuses on expression data alone. Here by combining both the genome-scale expression data and the metabolic network, we are to identify local blocks on the metabolic network. We call these blocks network modules. A network module consists of both a set of reactions that are connected on the metabolic network and a set of conditions under which the genes catalyzing the set of reactions are closely co-expressed. Reactions are connected on the network if the distance between any two of them is finite. The inclusion of metabolic network may make the results more biologically meaningful. Different network modules may have overlapping reactions. This is allowed because a reaction may participate in more than one module.

5.2 Searching Algorithm

The expression data are represented as a two dimensional matrix E, with rows as reactions and columns as conditions. Though the original expression data are only available for genes E can be easily obtained since each reaction corresponds to one or more genes. For each of 1007 reactions, the expression value for a condition is obtained by taking the maximal value of the expression ratios of all its corresponding genes under this condition. A module consists of a set of reactions (R) and a set of conditions (C), such that a function $F(R,C)$ is maximal. The function is defined as the following:

$$F(R,C)=\sum_{i\in R,\, j\in C}(E_{i,j}-Nr\cdot Nc\cdot T),$$

where Nr is the number of reactions in R and Nc is the number of conditions in C, and T is a threshold. Note that the reactions in R are required to be connected on the metabolic network. Here we are not only interested in the global maximum. Local maxima are also our concern. Our strategy is to start with a reaction and grow by including either a reaction or a condition, whichever makes F larger, in one-step. The reaction to be added is selected from the set of reactions that are connected to any of the reactions already included in the module. The growth process stops when F can no longer be increased by such move. The process is repeated with every reaction as a starting node. The modules found are then compared and similar modules are merged.

5.3 Analysis of Modules

5.3.1 Clustering Coefficients of Modules

A large pool of network modules is generated with different thresholds. The clustering coefficients of these modules are then calculated. An average value is obtained by averaging the clustering coefficients of networks of the same size. The clustering coefficient is plotted against the module size (Figure 5.1). Also on the same figure is the same plot for random modules, which are identical to random sub-networks in Chapter 2.

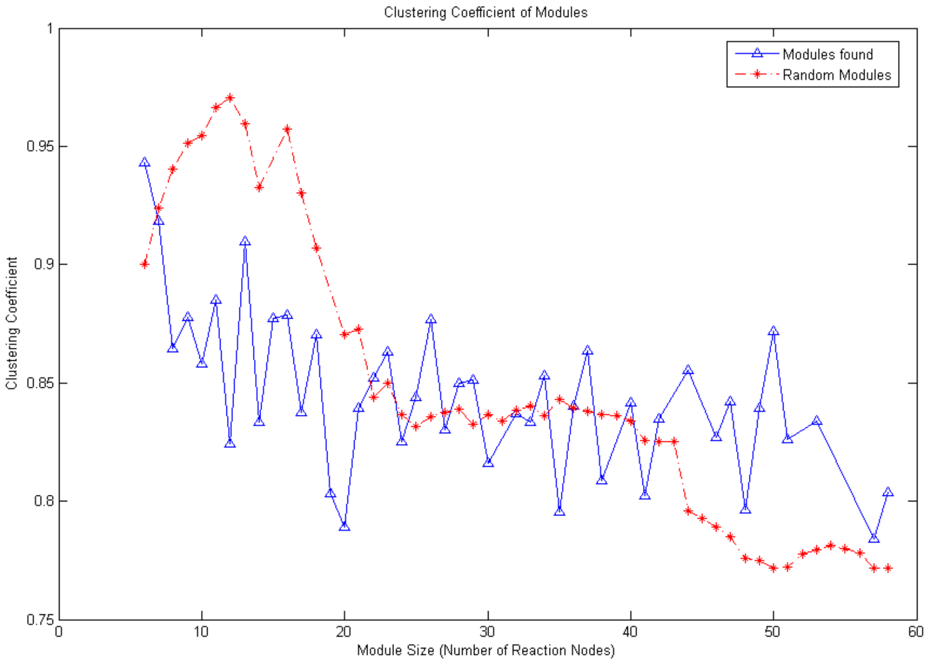

Figure 5.1: The clustering coefficients of network modules.

The comparison says that regardless of the hierarchical nature of the network topology the distribution of clustering coefficients of network modules is more uniform.

5.3.2 Examples of Modules

With a threshold of 2, 14 network modules are identified. In Table 5.1 we list the modules along with names of some biochemical pathways the modules overlap most.

Number of reactions	Number of conditions	Pathways the module overlaps most (Number of overlapped reactions)
49	65	Glycolysis / Gluconeogenesis (17) Carbon fixation (8) Purine metabolism (5)
43	15	Valine, leucine and isoleucine biosynthesis (11) Histidine metabolism (4)
88	27	Tyrosine metabolism (10) Pyruvate metabolism (8) Citrate cycle (TCA cycle) (7) Glycolysis / Gluconeogenesis (6)
15	7	Starch and sucrose metabolism (2) Pyrimidine metabolism (2) Purine metabolism (2) Galactose metabolism (2)
58	13	Purine metabolism (7) Urea cycle and metabolism of amino groups (7) Sulfur metabolism (5) Nitrogen metabolism (5) One carbon pool by folate (5) Arginine and proline metabolism (5)
14	13	Fatty acid metabolism (8) Fatty acid biosynthesis (path 2) (7)
8	14	Folate biosynthesis (5) One carbon pool by folate (3)
16	12	Purine metabolism (6) Pyrimidine metabolism (5)
19	13	Galactose metabolism (8)
31	13	Purine metabolism (10)
42	7	Fatty acid biosynthesis (path 1) (30)
16	19	C21-Steroid hormone metabolism (10) Androgen and estrogen metabolism (6)
6	134	Starch and sucrose metabolism (2)
11	187	Pentose phosphate pathway (4)

Table 5.1: The network modules and pathways they overlap most.

Chapter 6

Flux Balance Analysis of the Network

6.1 Introduction to FBA

Flux balance analysis (Varma & Palsson 1994; Bonarius *et al.* 1997; Edwards & Palsson 1999; Gombert & Nielsen 2000), or FBA, is to find the flux for each reaction in the network by linear programming, while the cell is in steady state. A more recent review on FBA is available (Kauffman *et al.* 2003). The stochiometric matrix S is a two-dimensional matrix with columns representing biochemical reactions and rows metabolites. The matrix elements are the reaction coefficients, with coefficients for substrates negative and products positive. Obviously, each column stands for one reaction and the matrix as a whole is a complete representation of the cell's biochemical reactions. Distinction is made between the so-called 'internal' and 'exchange' metabolites. The internal metabolites reside inside the cell and should be kept at constant concentrations in the steady state, while the exchange metabolites can be transported across the cell's membrane and their concentrations are not constrained. S_{in} is the stochiometirc matrix for the internal metabolites only. The reaction flux is represented as a column vector v. The steady state of the cell thus requires that $S_{in} \bullet v = 0$. Other constraints may include thermodynamic constraints (e.g. irreversibility of certain reactions) or capacity constraints (e.g. maximum uptake rate for a given compounds) (Figure 6.1).

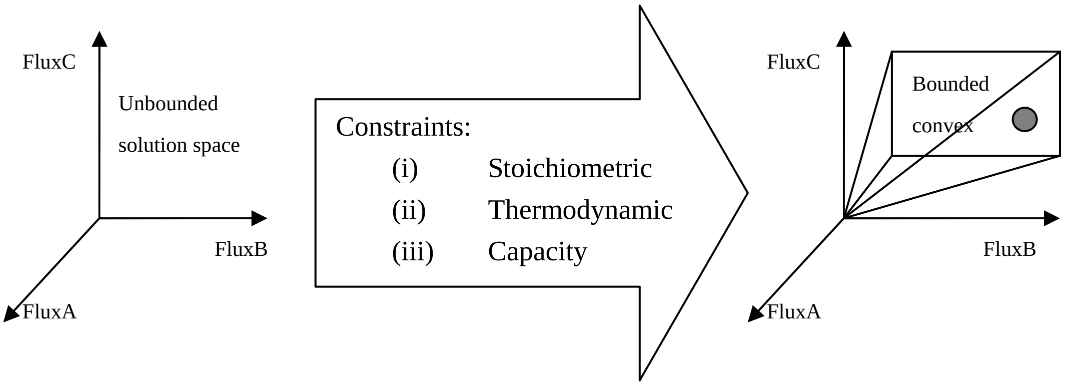

Figure 6.1: The flux balance analysis. The gray circle represents the optimal solution.

With all these constraints, an objective function, usually in the form of biomass production, is subject to maximization. This problem is a standard linear programming problem and can be solved easily with various software packages.

6.2 iND750 Model

The most recent *in silico* yeast model developed by Palsson's group is the iND750 model (Duarte *et al.* 2004; http://gcrg.ucsd.edu/organisms/yeast.html). The model includes 1149 reactions and 1061 metabolites. Note that the same compound that occurs in different compartments corresponds to more than one compound here. Regardless of the location, the number of chemically unique compounds is 646. Among the 1061 metabolites, there are 116 exchange metabolites. All the other 945 internal metabolites should thus be maintained at constant concentrations at steady state. The biomass consists of amino acids, nucleotides, carbohydrates, lipids and energy molecules in proper proportion. In the model, the biomass production is represented as the 1150th reaction. The model is then solved with LINDO (Lindo Systems, Inc.) linear programming package.

6.3 Flux Backbone

6.3.1 The Idea

We are interested in how the flux pattern changes when the cell is fed with different carbon sources. Before actual simulations are carried out, a reasonable conjecture would be that a flux backbone exists. The flux backbone should be utilized regardless what carbon sources are available. This structure reduces network complexity and increases the cell's efficiency and thus chance of survival. In the following two sections, we test the idea on both aerobic and anaerobic conditions.

6.3.2 Flux Backbone under Aerobic Conditions

In Palsson's aerobic glucose minimal medium simulation (http://gcrg.ucsd.edu/organisms/yeast/yeast_faqs.html), other than the eight unconstrained 'basic' compounds (Table 6.1) only glucose is available for the cell to consume.

O2	Ammonium	Sulfate	Phosphate
H2O	K+	Sodium	CO2

Table 6.1: The eight aerobic basic compounds.

With unlimited uptake rates of the eight basic compounds as in Palsson's aerobic glucose minimal medium example, we do the simulations by feeding the cell with each of the 108 remaining exchange compounds. Nonzero biomass is obtained for 43 of the compounds. For each of the 43 compounds fed to the cell, we list the number of reactions activated (Table 6.2).

Fed compounds	Number of activated reactions	Fed compounds	Number of activated reactions
Acetate	296	Trehalose	284
Ethanol	294	Guanosine	284
4-Aminobutanoate	294	Ornithine	283
Pyruvate	292	L-Arginine	283
Citrate	292	Xylitol	282
2-Oxoglutarate	292	D-Mannose	282
1,3-beta-D-Glucan	290	D-Glucose	282
Adenosine 3',5'-bisphosphate	289	L-Aspartate	282
L-Alanine	289	L-Asparagine	282
Acetaldehyde	289	Adenosine	282
Melibiose	288	D-Ribose	281
L-Glutamine	288	Uridine	280
Succinate	287	D-Galactose	280
Sucrose	286	Cytidine	280
D-Sorbitol	286	Inosine	279
L-Proline	286	D-Glucosamine 6-phosphate	279
D-Xylose	285	Fumarate	279
Maltose	285	D-Fructose	278
L-Malate	285	S-Adenosyl-L-methionine	277
Glycerol	285	L-Serine	276
L-Glutamate	285	Glycine	275
Xanthosine	284	MEAN	284.6

Table 6.2: Number of reactions activated for 43 carbon sources.

The flux backbone consists of 188 reactions, which are activated in all of the 43 cases.

6.3.3 Flux Backbone under Anaerobic Conditions

In Palsson’s anaerobic simulation, there are 13 basic compounds (Table 6.3).

Ammonium	Sulfate	Phosphate	H2O	K+
Sodium	CO2	Ergosterol	zymosterol	
octadecanoate (n-C18:0)	octadecenoate (n-C18:1)	octadecynoate (n-C18:2)	hexadecenoate (n-C16:1)	

Table 6.3: The 13 anaerobic basic compounds.

We then do the anaerobic simulations by feeding the cell with one of the 103 exchange compounds (the 13 basic compounds are excluded from the 116 exchange compounds). We get nonzero biomass productions for 17 of the compounds. Note that all the 17 compounds here are also present in 43 compounds which make the cell viable under the aerobic condition. The number of reactions activated for each of the 17 carbon sources is listed below (Table 6.4).

Fed compounds	Number of activated reactions	Fed compounds	Number of activated reactions
Melibiose	267	D-Ribose	264
Xanthosine	266	D-Mannose	264
1,3-beta-D-Glucan	266	Adenosine	264
Sucrose	265	D-Glucose	263
Maltose	265	D-Galactose	263
Guanosine	265	Inosine	259
D-Glucosamine 6-phosphate	265	L-Serine	257
D-Fructose	265	S-Adenosyl-L-methionine	255
Trehalose	264	MEAN	263.4

Table 6.4: Number of reactions activated for 17 carbon sources.

The flux backbone consists of 209 reactions, which are activated in all of the 17 cases. The number of the common reactions in both the aerobic backbone and anaerobic backbone is 144.

6.3.4 Discussions

The existence of this large backbone structure shows that various carbon sources can be converted to some common precursors in relatively few steps. The backbone is then utilized to produce biomass constituents from the small number of precursors, thus sustaining the cell growth.

6.4 Flux Patterns for Biomass with Different Constituents

6.4.1 Biomass Constituents

The 43 biomass constituents in the iND750 model can be grouped into 5 categories, namely, amino acids, nucleotides, carbohydrates, lipids, and other molecules (Table 6.5)

Amino acids (20)	Nucleotides (8)	Carbohydrates (4)	Lipids (8)	Others (3)
L-Alanine	dAMP	13BDglcn	ergst	atp
L-Arginine	dCMP	glycogen	pa_SC	h2o
L-Asparagine	dGMP	mannan	pc_SC	so4
L-Aspartate	dTMP	tre	pe_SC	
L-Cysteine	AMP		ps_SC	
L-Glutamine	CMP		ptdlino_SC	
L-Glutamate	GMP		triglyc_SC	
Glycine	UMP		zymst	
L-Histidine				
L-Isoleucine				
L-Leucine				
L-Lysine				
L-Methionine				
L-Phenylalanine				
L-Proline				
L-Serine				
L-Threonine				
L-Tryptophan				
L-Tyrosine				
L-Valine				

Table 6.5: The biomass constituents.

6.4.2 Simulations with Simple Biomass

Since the ‘real’ biomass has more than 40 constituents, we thus speculate that the activated sub-network should be much simpler if we simulate with a simpler version of biomass. To test the idea, we perform the simulation by each time choosing one of the

40 original constituents (atp, h2o and so4 are not included.) as the sole component of biomass under the aerobic glucose minimal medium. The rate of glucose uptake is set as 10. We obtain a sub-network, the part of the network that consists of the activated reactions, for each of the biomasses (Table 6.6).

Biomass Constituent	# of reactions activated	Biomass Constituent	# of reactions activated
Ergosterol	95	L-Lysine	69
Phosphatidylcholine	90	L-Histidine	69
phosphatidylethanolamine	84	GMP	68
L-Tryptophan	84	L-Isoleucine	67
zymosterol	83	L-Phenylalanine	65
phosphatidyl-1D-myo-inositol	81	L-Leucine	61
dCMP	80	L-Asparagine	60
L-Cysteine	80	L-Proline	57
phosphatidylserine	79	L-Aspartate	57
dGMP	78	L-Valine	55
triglyceride	77	Mannan	54
L-Methionine	77	L-Threonine	51
L-Arginine	77	glycogen	50
Phosphatidate	76	1,3-beta-D-Glucan	49
UMP	76	Trehalose	48
dTMP	76	L-Glutamine	45
CMP	75	L-Serine	42
AMP	75	L-Glutamate	42
dAMP	74	Glycine	39
L-Tyrosine	69	L-Alanine	35
		All 40 constituents	282

Table 6.6: Number of reactions activated with each of the constituents as biomass.

Here we define a quantity called centrality to measure how close each of the sub-network is from the ‘central part’ of the network. The central part consists of reactions that are activated in most cases. The higher the centrality, the less specific the sub-network. For sub-network i with n_i reactions, the centrality is calculated as

$$C_i = \frac{1}{(N-1)n_i} \sum_{j=1, j \neq i}^{N} M_{ij} ,$$

where M_{ij} is the number of overlapping reactions between sub-networks i and j, and N is total number of sub-networks, which is 40 here. The results are shown below (Table 6.7).

Biomass	Centrality	Biomass	Centrality
Glycine	0.74227	L-Proline	0.59019
L-Aspartate	0.73099	Phosphatidate	0.58974
L-Serine	0.69902	dTMP	0.58063
L-Asparagine	0.68932	L-Methionine	0.57742
L-Threonine	0.68175	triglyceride	0.57576
L-Glutamate	0.67216	dAMP	0.57554
Trehalose	0.67201	CMP	0.56991
1,3-beta-D-Glucan	0.65882	phosphatidylserine	0.5693
L-Glutamine	0.65242	L-Arginine	0.56743
L-Alanine	0.65201	dCMP	0.56474
glycogen	0.63487	phosphatidyl-1D-myo-inositol	0.55651
L-Phenylalanine	0.63471	dGMP	0.53649
L-Valine	0.61911	phosphatidylethanolamine	0.52595
Mannan	0.61349	L-Cysteine	0.52244
L-Isoleucine	0.60926	L-Tryptophan	0.51282
UMP	0.60493	Phosphatidylcholine	0.49715
L-Tyrosine	0.60126	zymosterol	0.49521
GMP	0.59389	L-Leucine	0.46742
L-Histidine	0.59123	L-Lysine	0.45373
AMP	0.59043	Ergosterol	0.45236

Table 6.7: The networks ranked according to their centralities.

6.5 Optimality of Network

6.5.1 Excretions of Network

Since the cell only maximizes its growth rate, we thus conceive that some ‘useful’ compounds may be excreted by the cell. But wasting some useful compounds is certainly not a good choice for the cell. To investigate this problem, we grow our *in silico* cell by feeding it with various carbon sources under minimal aerobic media and see what the cell excretes while maximizing the biomass production. The results are shown below (Table 6.8).

Carbon source	Excretion
1,3-beta-D-Glucan	CO2, H2O, H+
4-Aminobutanoate	CO2, Formate, H2O, Ammonium, Urea
Acetate	CO2, Formate, H2O, Urea
Acetaldehyde	CO2, H2O, H+
Adenosine	CO2, Formate, H2O, Hypoxanthine, Ammonium, Xanthine
2-Oxoglutarate	CO2, Formate, H2O, Urea
L-Alanine	CO2, Formate, H2O, Ammonium, Urea
S-Adenosyl-L-methionine	CO2, H+, Hypoxanthine, L-Methionine, Ammonium, Xanthine
L-Arginine	CO2, H2O, H+, Urea, Xanthine
L-Asparagine	CO2, Formate, Ammonium, Urea
L-Aspartate	CO2, Formate, H2O, Ammonium, Urea
Citrate	CO2, Formate, H2O, Urea
Cytidine	CO2, Formate, H2O, Ammonium, Thymine
Ethanol	CO2, H2O, H+
D-Fructose	CO2, H2O, H+
Fumarate	CO2, Formate, H2O, Urea
D-Galactose	CO2, H2O, H+
D-Glucosamine 6-phosphate	CO2, Formate, H2O, H+, Ammonium, Phosphate
D-Glucose	CO2, H2O, H+
L-Glutamine	CO2, Formate, H2O, Ammonium, Urea
L-Glutamate	CO2, Formate, H2O, Urea
Glycine	CO2, Formate, H2O, Ammonium, Urea
Glycerol	CO2, H2O, H+
Guanosine	CO2, Guanine, H2O, H+
Inosine	CO2, H2O, H+, Hypoxanthine
L-Malate	CO2, Formate, H2O, Urea
Maltose	CO2, H2O, H+
D-Mannose	CO2, H2O, H+
Melibiose	CO2, H2O, H+
Ornithine	CO2, Formate, H2O, Ammonium, Urea
Adenosine 3',5'-bisphosphate	CO2, Formate, H2O, H+, Phosphate, Xanthine
L-Proline	CO2, Formate, H2O, Ammonium, Urea
Pyruvate	CO2, Formate, H2O, Urea
D-Ribose	CO2, H2O, H+
D-Sorbitol	CO2, H2O, H+
L-Serine	CO2, Formate, H2O, Ammonium
Succinate	CO2, Formate, H2O, Urea
Sucrose	CO2, H2O, H+

Trehalose	CO2, H2O, H+
Uridine	CO2, H2O, H+, Thymine
Xanthosine	CO2, H2O, H+, Xanthine
D-Xylose	CO2, H2O, H+
Xylitol	CO2, H2O, H+

Table 6.8: The excretions of the cell when fed with different carbon sources.

The list shows that only a few compounds are excreted. Some of them are ‘natural’ wastes since they are either carbon-free metabolites (H2O, H+, Ammonium, Phosphate) or metabolites with carbons in very low energy states (CO2, Formate, Urea). The other wastes are Guanine, Thymine, Hypoxanthine, Xanthine, and L-Methionine. Most of the external compounds still are not excreted.

The results indicate that the metabolic network consists of highly independent modules and each of them is efficiently constructed. The conclusion may not be right if the network forbids the excretion of most of the external compounds. To find out whether this is the case, we maximize the production of the external compound and see if we get positive values (Table 6.9 & Table 6.10).

Acetate	Ethanol	Hypoxanthine	D-Sorbitol
Acetaldehyde	Formate	L-Isoleucine	L-Serine
2-Oxoglutarate	Fumarate	L-Leucine	Succinate
L-Alanine	D-Glucosamine 6-phosphate	L-Lysine	L-Threonine
L-Arginine	L-Glutamine	L-Malate	Thymine
L-Asparagine	L-Glutamate	L-Methionine	L-Tryptophan
L-Aspartate	Glycine	Ornithine	L-Tyrosine
Citrate	Glycerol	Adenosine 3',5'-bisphosphate	Urea
L-Cysteine	Guanine	L-Phenylalanine	L-Valine
dTTP	H+	(R)-Pantothenate	Xanthine
Ergosterol	L-Histidine	L-Proline	zymosterol

Table 6.9: The list of possible products under the aerobic glucose minimal medium.

1,3-beta-D-Glucan	Cytidine	Inosine	Spermidine
4-Aminobutanoate	Deoxyadenosine	L-Lactate	Spermine
5-Amino-4-oxopentanoate	7,8-Diaminononanoate	Maltose	L-Sorbose
8-Amino-7-oxononanoate	Deoxycytidine	D-Mannose	Sucrose
L-Arabinitol	Deoxyguanosine	Melibiose	Thiamin
Adenine	Deoxyinosine	S-Methyl-L-methionine	Thiamin monophosphate
Adenosine	Deoxyuridine	NMN	Thiamine diphosphate
Allantoin	FMN	octadecanoate (n-C18:0)	Thymidine
Allantoate	D-Fructose	octadecenoate (n-C18:1)	Trehalose
S-Adenosyl-L-methionine	D-Galactose	octadecynoate (n-C18:2)	tetradecanoate (n-C14:0)
D-Arabinose	Glycolaldehyde	peptide	Uracil
L-Arabinose	Guanosine	Putrescine	Uridine
Biotin	Oxidized glutathione	Pyruvate	Xanthosine
Choline	Hexadecanoate (n-C16:0)	D-Ribose	D-Xylose
L-Carnitine	hexadecenoate (n-C16:1)	Riboflavin	Xylitol
Cytosine	myo-Inositol	L-Sorbitol	

Table 6.10: The list of impossible products under the aerobic glucose minimal medium.

The results show that a total number of 63 external compounds are impossible to be made under the aerobic glucose minimal medium.

6.5.2 Leakages of Internal Metabolites

To further investigate the optimality of the network, we allow the internal metabolites to be leaked out while maximizing the biomass production. Surprisingly, the leakages of only two internal metabolites lead to larger biomass production. These two metabolites are h[m] (hydrogen ion in mitochondrion) and hco3[c] (bicarbonate in cytosol). The fact that most internal metabolites don’t leak out even if allowed to do so illustrates that the network is internal optimized.

6.5.3 Superposition of Solutions

In section 6.4.2, we simulate the cell with biomass set to each of the 40 original constituents. The biomass production is now our concern (Table 6.11). Along with the biomass production, we also list the coefficients of the 40 constituents in the original biomass reaction. At the last row of Table 6.11, we list the biomass production for biomass consisting of all the 40 constituents with original coefficients. Note the glucose uptake rate is 10 in all cases.

Biomass constituent	Coefficient in original biomass reaction	Biomass production
L-Alanine	0.458800	17.142858
L-Arginine	0.160700	6.938389
L-Asparagine	0.101700	12.577320
L-Aspartate	0.297500	15.844155
L-Cysteine	0.006600	8.913526
L-Glutamine	0.105400	10.000000
L-Glutamate	0.301800	10.000000
Glycine	0.290400	20.000000
L-Histidine	0.066300	6.765250
L-Isoleucine	0.192700	7.311828
L-Leucine	0.296400	6.666666
L-Lysine	0.286200	6.666667
L-Methionine	0.050700	6.182432
L-Phenylalanine	0.133900	5.335277
L-Proline	0.164700	9.539749
L-Serine	0.185400	17.142857
L-Threonine	0.191400	12.750000
L-Tryptophan	0.028400	4.073107
L-Tyrosine	0.102000	5.583524
L-Valine	0.264600	9.230769
dAMP	0.003600	4.295775
dCMP	0.002400	5.198863
dGMP	0.002400	4.246714
dTMP	0.003600	4.647619
AMP	0.046000	4.455265
CMP	0.044700	5.269761
GMP	0.046000	4.365904
UMP	0.059900	5.600612
1,3-beta-D-Glucan	1.134800	8.974359

glycogen	0.518500	8.974359
Mannan	0.807900	8.974358
Trehalose	0.023400	4.605263
Ergosterol	0.000700	0.919714
Phosphatidate	0.000006	0.009478
Phosphatidylcholine	0.000060	0.007994
phosphatidylethanolamine	0.000045	0.008954
phosphatidylserine	0.000017	0.008992
phosphatidyl-1D-myo-inositol	0.000053	0.008492
triglyceride	0.000066	0.006478
zymosterol	0.001500	0.966847
All 40 constituents		1.4098

Table 6.11: The biomass productions with different biomass constituents.

Given the biomass yield with the production of each of the 40 constituents optimized, we can get a superposition solution for the simultaneous optimization of all of them with the original proportion in the iND750 model. When constituent C_i is optimized only, the yield is denoted as y_i. Let Y_{op} be the optimal solution when all n constituents are included in the biomass formula, $M = \sum_{i=1}^{n} \alpha_i C_i$, where α_i is the proportion of C_i. The superposition solution satisfies $Y_{sp}\alpha_i = \beta_i y_i$, where β_i is the proportion of carbon source uptake and $\sum_{i=1}^{n} \beta_i = 1$. By 1 we mean a unit of carbon source uptake rate.

Solution to these equations are given as $\beta_i = Y_{sp}(\alpha_i / y_i)$, and $Y_{sp} = \frac{1}{\sum_{i=1}^{n} \alpha_i / y_i}$. We now compare Y_{sp} with the optimal solution Y_{op}. From Table 6.10, Y_{sp} is easily calculated as 1.3470 and the ratio Y_{sp} / Y_{op} equals 95.54%.

The fact that the superposition solution is so close to the optimal solution is quite a surprise. It may imply that the network consists of highly independent modules, a property we plan to explore further.

Chapter 7

Loop Structures in Network

7.1 Loops as Network Motifs

By definition, network motifs are significantly recurring units in the network. We apply Mfinder (Milo *et al.* 2002) to the network and identify the motifs with 4 nodes The only 4 node motif found out of a total number of 199 possible sub-graphs with size 4 is the loop structure (Table 7.1).

Number of random networks	Frequency in real network (F)	Mean frequency in random network (F_m)	Standard deviation of frequency in random network (STD)	Z score
1001	864	16.4	4.5	189.81

Table 7.1: Statistics generated by Mfiner. Z score is a measure of the distance in standard deviation of a sample from the mean, or $Z = (F - F_m)/STD$.

7.2 Loop Enumeration

Before finding loops, we simplify the network by first removing some currency compounds (Table 7.2) and then removing nodes that have zero or one connections and reaction nodes that connect to either only substrates or products.

Currency compounds
adp amp atp cdp cmp co2 coa ctp fad fadh2 gdp gmp gtp h h2o h2o2 k na1 nad nadh nadp nadph o2 pi ppi q6 q6h2 so4 udp ump utp

Table 7.2: The currency compounds removed from the network.

The simplified network has only 785 reactions and 588 compounds. An algorithm of loop enumeration (see Appendix 5 for Matlab® (MathWorks Inc.) code) is carried out on this simplified network. Loops with length up to 12 have been completely identified (Table 7.3).

Number of loops	645	313	1380	3423	17921
Loop length	4	6	8	10	12

Table 7.3: The number of loops with different lengths.

In the above network, we don't distinguish between reversible and nonreversible reactions, with all reactions treated as reversible. Here we remove non-physical loops (Figure 1) by put the reversibility of reactions into consideration.

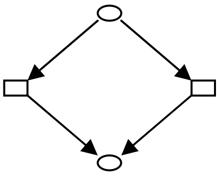

Figure 7.1: Non-physical loops. Circles represent for compounds and rectangles for reactions.

The number of loops is shown below (Table 7.4) after the removal.

Number of loops	290	110	502	665	4312
Loop length	4	6	8	9	12

Table 7.4: The number of loops after removal of non-physical loops.

Some loops may have the same set of compounds and differ only in their reaction set. We merge this kind of loops and the loop number is considerably reduced (Table 7.5).

Number of loops	78	44	119	157	452
Loop length	4	6	8	10	12

Table 7.5: The number of loops with unique set of compounds.

The loop number is further reduced when we don't distinguish between the same compounds at different cellular compartments (Table 7.6).

Number of loops	65	42	109	147	419
Loop length	4	6	8	10	12

Table 7.6: The number of loops when no distinction is made between different compartments.

7.3 2-Compound Loops

7.3.1 Grouping of 2-Compound Loops

If we think the loops as carriers of chemical parts, we then are able to identify these parts for 2-compound loops (or 2-C loops) (Table 7.7).

Chemical parts	Frequency in 2-C loops	Frequency in 2-C-2-R loops
HO-3P-1	9	9
*	8	19
H2	7	21
O3P	4	4
H4NO-1	3	167
H2NO-1	3	14
CO	2	5
C9H11N2O8P	2	2
C5H8O4	2	2
C4H3O-1	2	2
C2H2O	2	6
O300P100	1	1
HO	1	1
H	1	3
CH2O	1	1
CH2	1	4
CH-1O2	1	1
C5H6NO3	1	1
C5H4O4	1	1
C4H8NO2	1	1
C4H7NO2	1	2
C4H2	1	1
C3HO3	1	10
C3H5NO2	1	1
C3H4O2	1	1
C2HO2	1	1
C2H4O2	1	1
C14H17N6O13P3S	1	1
C10H14N3O6S	1	1
C10H12N5O3	1	1
C10H11N5O3	1	1
C	1	4

Table 7.7: The grouping of 2-compound loops.

Here the chemical parts, or load on the carriers, are defined as the constituent difference between the two compounds. The star symbol (*) in the table represents a null load, which corresponds to a transport loop. In such a loop, the same compound flows in and out of a compartment.

Some loops identified by our algorithm may not be what we're looking for. One of such situations arises when different reactions have the same sets of substrates and products. Totally 4 sets of such reaction pairs are present in the iND750 model (Table 7.8).

Reaction	ORF
accoa + crn --> acrn + coa	YML042W
acrn + coa --> accoa + crn	YAR035W
dhlam + nad <==> h + lpam + nadh	(YDR019C YMR189W YAL044C YFL018C)
dhlam + nad --> h + lpam + nadh	(YIL125W YDR148C YFL018C)
fad + succ <==> fadh2 + fum	(YDR178W YKL141W YKL148C YLL041C) or (YKL141W YKL148C YLL041C YLR164W) or (YDR178W YKL148C YLL041C YMR118C) or (YDR178W YJL045W YKL141W YLL041C)
fadh2 + fum --> fad + succ	YJR051W
34hpp + glu-L --> akg + tyr-L	YGL202W or YHR137W
akg + tyr-L <==> 34hpp + glu-L	YLR027C

Table 7.8: The four reactions that lead to 'fake' loops.

Now with transport loops and 'fake' loops removed, we have 54 2-compound loops left (Table 7.9).

Compound 1	Formula 1	Compound 2	Formula 2	Common part	Different part
5,6,7,8-Tetrahydrofolate	C19H22N7O6	5,10-Methylenetetrahydrofolate	C20H22N7O6	C19H22N7O6	C
L-Homocysteine	C4H9NO2S	S-Adenosyl-L-homocysteine	C14H20N6O5S	C4H9NO2S	C10H11N5O3
L-Methionine	C5H11NO2S	S-Adenosyl-L-methionine	C15H23N6O5S	C5H11NO2S	C10H12N5O3
Reduced glutathione	C10H16N3O6S	Oxidized glutathione	C20H30N6O12S2	C10H16N3O6S	C10H14N3O6S
Acetyl-ACP	C13H23N2O8PRS	acyl carrier protein	C11H21N2O7PRS	C11H21N2O7PRS	C2H2O
L-Carnitine	C7H15NO3	O-Acetylcarniti	C9H17NO4	C7H15NO3	C2H2O

		ne			
D-Xylulose 5-phosphate	C5H9O8P	Glyceraldehyde 3-phosphate	C3H5O6P	C3H5O6P	C2H4O2
L-Malate	C4H4O5	Citrate	C6H5O7	C4H4O5	C2HO2
(R)-S-Lactoyl glutathione	C13H20N3O8S	Reduced glutathione	C10H16N3O6S	C10H16N3O6S	C3H4O2
L-Homocysteine	C4H9NO2S	L-Cystathionine	C7H14N2O4S	C4H9NO2S	C3H5NO2
Malonyl-[acyl-carrierprotein]	C14H22N2O10PRS	acyl carrier protein	C11H21N2O7PRS	C11H21N2O7PRS	C3HO3
6,7-Dimethyl-8-(1-D-ribityl)lumazine	C13H18N4O6	4-(1-D-Ribitylamino)-5-aminouracil	C9H16N4O6	C9H16N4O6	C4H2
2-Oxoglutarate	C5H4O5	3-(4-Hydroxyphenyl)pyruvate	C9H7O4	C5H4O4	C4H3O-1
L-Cystathionine	C7H14N2O4S	L-Cysteine	C3H7NO2S	C3H7NO2S	C4H7NO2
O-Succinyl-L-homoserine	C8H12NO6	Succinate	C4H4O4	C4H4O4	C4H8NO2
Ammonium	H4N	L-Glutamate	C5H8NO4	H4N	C5H4O4
Ammonium	H4N	L-Glutamine	C5H10N2O3	H4N	C5H6NO3
Adenosine	C10H13N5O4	Adenine	C5H5N5	C5H5N5	C5H8O4
Guanine	C5H5N5O	Guanosine	C10H13N5O5	C5H5N5O	C5H8O4
UDPglucose	C15H22N2O17P2	D-Glucose 1-phosphate	C6H11O9P	C6H11O9P	C9H11N2O8P
UDPgalactose	C15H22N2O17P2	alpha-D-Galactose 1-phosphate	C6H11O9P	C6H11O9P	C9H11N2O8P
L-Glutamate	C5H8NO4	4-Aminobutanoate	C4H9NO2	C4H8NO2	CH-1O2
L-Glutamate	C5H8NO4	L-Aspartate	C4H6NO4	C4H6NO4	CH2
Glyceraldehyde 3-phosphate	C3H5O6P	D-Erythrose 4-phosphate	C4H7O7P	C3H5O6P	CH2O
Reduced glutathione	C10H16N3O6S	S-Formylglutathione	C11H16N3O7S	C10H16N3O6S	CO
5,6,7,8-Tetrahydrofolate	C19H22N7O6	10-Formyltetrahydrofolate	C20H22N7O7	C19H22N7O6	CO
Ferrocytochrome c	C42H53FeN8O6S2	Ferricytochrome c	C42H52FeN8O6S2	C42FeH52N8O6S2	H

Glycerol 3-phosphate	C3H7O6P	Dihydroxyacetone phosphate	C3H5O6P	C3H5O6P	H2
5,10-Methylenetetrahydrofolate	C20H22N7O6	5,10-Methenyltetrahydrofolate	C20H20N7O6	C20H20N7O6	H2
Oxaloacetate	C4H2O5	L-Malate	C4H4O5	C4H2O5	H2
Orotate	C5H3N2O4	(S)-Dihydroorotate	C5H5N2O4	C5H3N2O4	H2
Succinate	C4H4O4	Fumarate	C4H2O4	C4H2O4	H2
Reduced thioredoxin	XH2	Oxidized thioredoxin	X	X	H2
L-Aspartate	C4H6NO4	L-Asparagine	C4H8N2O3	C4H6NO3	H2NO-1
D-Glucosamine 6-phosphate	C6H13NO8P	D-Fructose 6-phosphate	C6H11O9P	C6H11O8P	H2NO-1
L-Glutamate	C5H8NO4	L-Glutamine	C5H10N2O3	C5H8NO3	H2NO-1
L-Glutamate	C5H8NO4	2-Oxoglutarate	C5H4O5	C5H4O4	H4NO-1
Pyruvate	C3H3O3	L-Alanine	C3H7NO2	C3H3O2	H4NO-1
L-Tyrosine	C9H11NO3	3-(4-Hydroxyphenyl)pyruvate	C9H7O4	C9H7O3	H4NO-1
5,10-Methenyltetrahydrofolate	C20H20N7O6	5-Formyltetrahydrofolate	C20H21N7O7	C20H20N7O6	HO
Deoxyuridine	C9H12N2O5	dUMP	C9H11N2O8P	C9H11N2O5	HO-3P-1
D-Fructose 6-phosphate	C6H11O9P	D-Fructose 2,6-bisphosphate	C6H10O12P2	C6H10O9P	HO-3P-1
D-Fructose 1,6-bisphosphate	C6H10O12P2	D-Fructose 6-phosphate	C6H11O9P	C6H10O9P	HO-3P-1
Glycerol	C3H8O3	Glycerol 3-phosphate	C3H7O6P	C3H7O3	HO-3P-1
Inosine	C10H12N4O5	IMP	C10H11N4O8P	C10H11N4O5	HO-3P-1
Phytosphingosine	C18H40NO3	Phytosphingosine 1-phosphate	C18H39NO6P	C18H39NO3	HO-3P-1
Pyridoxamine	C8H13N2O2	Pyridoxamine 5'-phosphate	C8H12N2O5P	C8H12N2O2	HO-3P-1
Sphinganine	C18H40NO2	Sphinganine 1-phosphate	C18H39NO5P	C18H39NO2	HO-3P-1
Thymidine	C10H14N2O5	dTMP	C10H13N2O8P	C10H13N2O5	HO-3P-1

Phosphatidate	C3540H6544O800P100	diacylglycerol pyrophosphate	C3540H6544O1100P200	C3540H6544O800P100	O300P100
dATP	C10H12N5O12P3	dADP	C10H12N5O9P2	C10H12N5O9P2	O3P
dGMP	C10H12N5O7P	dGDP	C10H12N5O10P2	C10H12N5O7P	O3P
dGTP	C10H12N5O13P3	dGDP	C10H12N5O10P2	C10H12N5O10P2	O3P
ITP	C10H11N4O14P3	IDP	C10H11N4O11P2	C10H11N4O11P2	O3P

Table 7.9: The 54 2-compound loops.

7.3.2 Removal of 2-Compound Loops

We remove the 290 2-compound loops and then enumerate 3,4,5,6-compound loops (Table 7.10).

Number of loops after removal of 2-C loops	0	22	65	27	54
Loop length	4	6	8	9	12

Table 7.10: The number of loops after removal of 2-compound loops.

Note that the loop numbers have been significantly reduced, and this observation lets us think that probably the 2-compound loops are the major contributor to the overall network complexity. The idea is further tested by removing the 2-compounds in a flux network, which is the part of the whole network that actually carries flux in a given simulation. Here the flux network we examine is the one under the aerobic glucose minimal medium (Section 6.3.2). Figure 7.2 shows the flux network after the removal of 2-compound loops. The network becomes so simple that it is now ready for visual inspection.

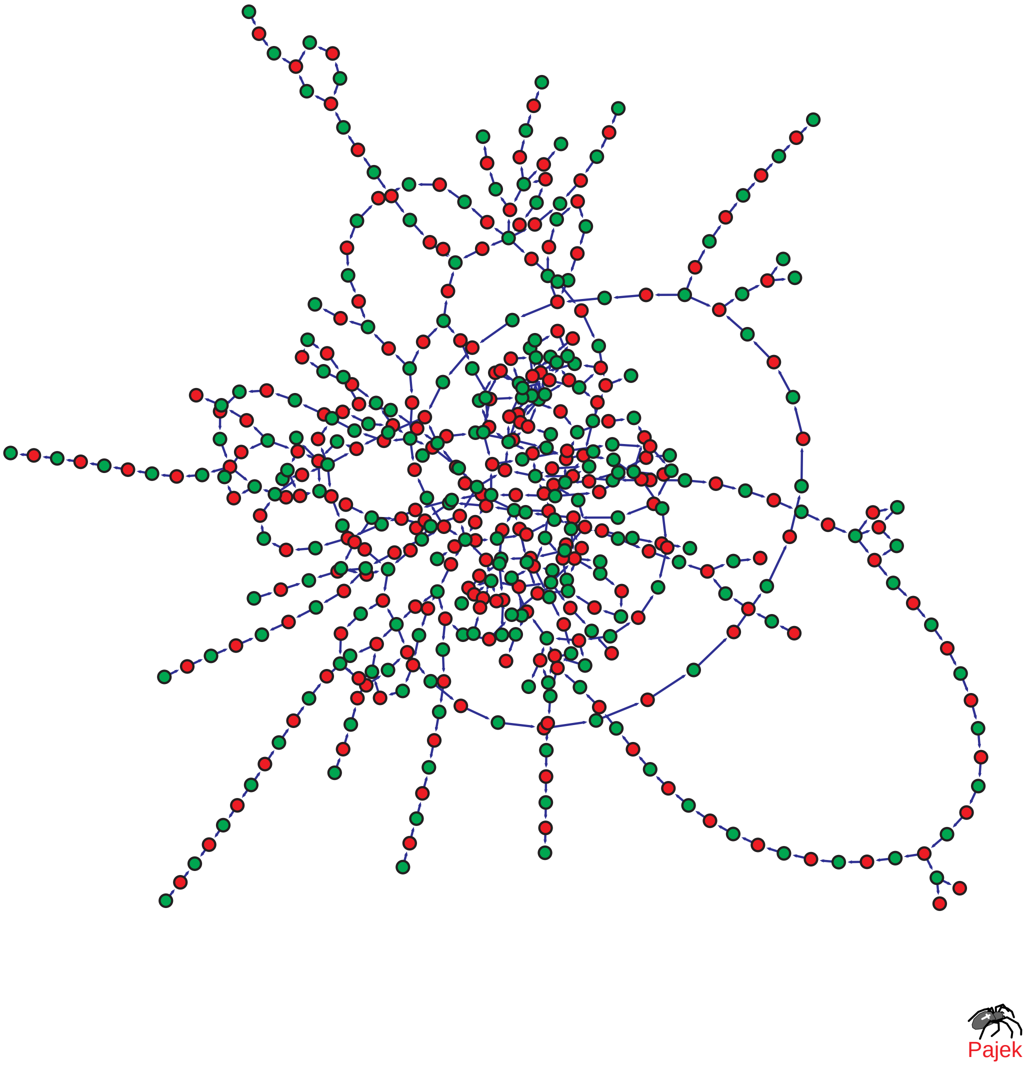

Figure 7.2: The flux network after the removal of 2-compound loops.

7.4 Conserved Parts in Loops

A first thought may be that there is no conserved common part for some loops, especially as the loops become longer. However, to our surprise, all loops with number of compounds up to six have conserved parts. A total number of 130 conserved parts or ‘carriers’ are found and are shown below along with their frequencies and possible corresponding compound names (Table 7.11). The frequency of a conserved part is simply the number of loops it belongs to.

Conserved part (130)	frequency	Compound with the same chemical formula (71)
H4N	97	Ammonium
C3H2O3	66	
C4H2O4	39	Fumarate
C3H3O3	38	Pyruvate
H4	35	
C3H5O6P	32	Glyceraldehyde 3-phosphate
C2H5NO2	32	Glycine
C4H2O5	29	Oxaloacetate
C2H3O2	29	Acetate
H3	21	
C11H21N2O7PRS	20	acyl carrier protein
C5H4O3	17	
C3H3O2	16	
C5H4O4	14	Itaconate
C4H4O5	13	L-Malate
C19H20N7O6	12	7,8-Dihydrofolate
C4H4O4	11	Succinate
C2HO3	11	
C5H8NO3	9	
C3H2O5	9	
C5H4O5	8	2-Oxoglutarate
C4H6NO3	8	
CH5N	6	
C6H10O5	6	Mannan
C3H2O2	6	
C2H5NO	6	
C2H3O	6	
C4H9NO2S	5	L-Homocysteine
C3H5O3	5	L-Lactate
C19H22N7O6	5	5,6,7,8-Tetrahydrofolate
C6H11O6	4	
C5H8NO4	4	L-Glutamate
C4H6NO4	4	L-Aspartate
C4H4O3	4	
C4H2O3	4	
C3H7NO3	4	L-Serine
C14H27O2	4	tetradecanoate (n-C14:0)
H2	3	
H	3	H+

C9H7O3	3	Phenylpyruvate
C8H15NOS2	3	Lipoamide
C7H15NO3	3	L-Carnitine
C6H5O7	3	Isocitrate
C6H11O9P	3	D-Tagatose 6-phosphate
C5H9O8P	3	D-Xylulose 5-phosphate
C5H9O5	3	
C5H9O4	3	(R)-2,3-Dihydroxy-3-methylbutanoate
C5H8NO2	3	
C5H4N4	3	
C4H9NO2	3	4-Aminobutanoate
C4H7O7P	3	D-Erythrose 4-phosphate
C3H2O4	3	
C2H2O2	3	
C14H20N6O5	3	
C10H16N3O6S	3	Reduced glutathione
CHO2	2	
C9H17NO4	2	O-Acetylcarnitine
C9H11N2O7P	2	
C6H10O9P	2	
C5H8O6	2	
C5H5N5O	2	Guanine
C4H6NO2	2	
C3H6NO2	2	
C3H5O2	2	
C3540H6544O800P100	2	Phosphatidate
C2H4O2	2	Glycolaldehyde
C20H20N7O6	2	5,10-Methenyltetrahydrofolate
C16H31O2	2	Hexadecanoate (n-C16:0)
C14H20N6O5S	2	S-Adenosyl-L-homocysteine
X	1	Oxidized thioredoxin
S2X	1	Lipoylprotein
HO	1	hydroxide ion
CH2NO3	1	
C9H16N4O6	1	4-(1-D-Ribitylamino)-5-aminouracil
C9H11N2O8P	1	dUMP
C9H11N2O5	1	
C9H11N2O4	1	
C8H8NO3	1	
C8H8NO2	1	
C8H12N2O2	1	
C7H9NO3	1	
C7H5O3	1	4-Hydroxybenzoate

C7H10O5	1	3-Carboxy-3-hydroxy-4-methylpentanoate
C6H9O9P	1	6-phospho-D-glucono-1,5-lactone
C6H9O4	1	2-Dehydropantoate
C6H8NOS	1	
C6H8N3O	1	
C6H12O6	1	L-Sorbose
C6H11O8P	1	
C5H9O3	1	
C5H8O5	1	D-Arabinono-1,4-lactone
C5H7O3	1	3-Methyl-2-oxobutanoate
C5H6NO3	1	L-1-Pyrroline-3-hydroxy-5-carboxylate
C5H5N5	1	Adenine
C5H5N2O4	1	(S)-Dihydroorotate
C5H4O2	1	
C5H4N4O2	1	Xanthine
C5H4N4O	1	Hypoxanthine
C5H4	1	
C5H3N2O4	1	Orotate
C5H13N2O2	1	Ornithine
C5H11NO2S	1	L-Methionine
C4H8O2	1	
C4H8NO2	1	
C4H5O3	1	Succinic semialdehyde
C4H4O2	1	
C42FeH52N8O6S2	1	
C4140H7544O1300P100	1	
C3H7O3	1	
C3H7NO2S	1	L-Cysteine
C3H4O2	1	Methylglyoxal
C3H2O6P	1	Phosphoenolpyruvate
C3740H7144N100O800P100	1	
C30H49O3	1	
C2H5O2	1	
C18H39NO3	1	
C18H39NO2	1	
C16H29O2	1	hexadecenoate (n-C16:1)

Table 7.11: The conserved parts in loops.

Notably, many (more than half) of the conserved parts have the same chemical formula as natural metabolites.

Chapter 8

Summary and Prospects

In this thesis we tackle the problem of yeast metabolism from several directions. We have confirmed the scale free and modular properties of the metabolic network and focused our effort on dissecting the network complexity from its functional perspective. We have discovered an important network motif, the internal compound and reaction loops, of which the 2-compound loop is the most evident and possibly crucial in flux coupling. Removal of these 2-compound loops greatly reduces the network complexity. This observation made by us serves an important step for simplifying the network flow. We also probe the network activation by mapping the genome-scale expression data onto the metabolic network. The simulations of the *in silico* cell under different growth conditions generate rich information, which allows for a quantitative assessment of the optimality of the network's functionality. A key finding in this regard is the near optimal solution constructed using superposition of single-compound synthetic fluxes.

We will continue our effort to elucidate the roles played by reaction loops in defining the flux patterns generated by *in silico* cell simulations. After this work is completed, we plan to integrate other genome-scale data, such as the transcriptional regulation network, protein abundance data, metabolite concentration data to construct a more complete picture of yeast metabolism, and to build quantitative and dynamic models, for this important and fundamental cellular process.

Bibliography

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K. & Walter, P. *Molecular Biology of the Cell*, Fourth Edition (2002).

Barabasi, A.-L. & Albert, R. Emergence of scaling in random networks. *Science* **286**, 509-512 (1999).

Barthelemy, M., Gondran, B. & Guichard, E. Spatial structure of the Internet traffic. *Physica A* **319**, 633-642 (2003).

Batagelj, V. & Mrvar, A. Pajek - program for large network analysis. *Connections* **21**, 47-57 (1998). (Home page: http://vlado.fmf.uni-lj.si/pub/networks/pajek/).

Becker, W.M., Kleinsmith, L.J. & Hardin, J. *The World of the Cell*, Fourth Edition (2000).

Ben-Dor, A., Shamir, R. & Yakhini, Z. Clustering gene expression patterns. *J Comput. Biol.* **6**, 281-297 (1999).

Bonarius, H. P. J., Schmid, G., & Tramper, J. Flux analysis of underdetermined metabolic networks: the quest for the missing constraints. *Trends Bio/Technol.* **15**, 308-314 (1997).

Bussemaker, H.J., Li, H. & Siggia, E.D. Regulatory element detection using correlation with expression. *Nature Genetics* **27**, 167-171 (2001).

Chen, G., Hata, N. & Zhang, M.Q. Transcription factor binding element detection using functional clustering of mutant expression data. *Nucl. Acids Res.* **32**, 2362-2371 (2004).

Cheng, Y. & Church, G.M. Biclustering of expression data. *Proc. Eighth Int'l Conf. Intelligent systems for molecular biology*, 93-103 (2000).

D'haeseleer, P., Liang, S. & Somogyi, R. Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics* **16**, 707-726 (2000).

de la Fuente, A., Brazhnik, P. & Mendes, P. Linking the genes: inferring quantitative gene networks from microarray data. *Trends in Genetics* **18**, 395-398 (2002).

Duarte, N.C., Herrgard, M.J. & Palsson, B.O. Reconstruction and validation of Saccharomyces cerevisiae iND750, a fully compartmentalized genome-scale metabolic model. *Genome Research* **14**, 1298-1309 (2004).

Edwards, J. S., & Palsson, B.O. Properties of the Haemophilus influenzae Rd metabolic genotype. *J. Biol. Chem.* **274**, 17410-17416 (1999).

Eisen, M.B., Spellman, P.T., Brown, P.O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci.* **95**, 14863-14868 (1998).

Farkas, I., Jeong, H., Vicsek, T., Barabasi, A.-L. & Oltvai, Z.N. The topology of the transcription regulatory network in the yeast, Saccharomyces cerevisiae. *Physica A* **318**, 601-612 (2003).

Gombert, A.K. & Nielsen, J. Mathematical modeling of metabolism. *Curr. Opin. Biotechnol.* **11**, 180-186 (2000).

Heyer, L.J., Kruglyak, S. & Yooseph, S. Exploring expression data: identification and analysis of coexpressed genes. *Genome Res.* **9**, 1106-1115 (1999).

Ihmels, J., Friedlander, G., Bergmann, S., Sarig, O, Ziv, Y. & Barkai, N. Revealing modular organization in the yeast transcription network. *Nature Genetics* **31**, 370-377 (2002).

Ihmels, J., Levy, R. & Barkai, N. Principles of transcriptional control in the metabolic network of Saccharomyces cerevisiae. *Nature Biotechnology* **22**, 86-92 (2003).

Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N. & Barabasi, A.-L. The large scale organization of metabolic networks. *Nature* **407**, 651-654 (2000).

Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucl. Acids Res.* **28**, 27-30 (2000).

Kanehisa, M. A database for post-genome analysis. *Trends Genet.* **13**, 375-376 (1997).

Kato, M., Hata, N., Banerjee, N., Futcher, B. & Zhang, M.Q. Identifying combinatorial regulation of transcription factors and binding motifs. *Genome Biol.* 5:R56 (2004).

Kauffman, K.J., Prakash, P. & Edwards, J.S. Advances in flux balance analysis. *Curr. Opin. Biotechnol.* **14**, 491-496 (2003).

Lewin, B. *Genes VIII* (2003).

Madeira, S.C. & Oliveira, A.L. Biclustering algorithms for biological data analysis: a survey. *IEEE Trans. Comput. Biology Bioinform.* **1**, 24-25 (2004).

Mathews, C.K., van Holde, K.E. & Ahern, K.G. *Biochemistry*, Third Edition (2000).

Michaels, G.S., Carr, D.B., Askenazi, M., Fuhrman, S., Wen, X. & Somogyi, R. Cluster analysis and data visualization of large-scale gene expression data. *Pac. Symp. Biocomput.* 42-53 (1998).

Milo, R., Itzkovitz, S., Kashtan, N., Levitt, R., Shen-Orr, S., Ayzenshtat, I., Sheffer, M. & Alon, U. Superfamilies of designed and evolved networks. *Science* **303**, 1538-1542 (2004).

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii., D. & Alon, U. Network motifs: simple building blocks of complex networks. *Science* **298**, 824-827 (2002).

Ravasz, E., Somera, A.L., Mongru, D.A., Oltvai, Z.N. & Barabasi, A.-L. Hierarchical organization of modularity in metabolic networks. *Science* **297**, 1551-1555 (2002).

Sasik, R., Hwa, T., Iranar, N. & Loomis, W. Percolation clustering: a novel algorithm applied to the clustering of gene expression patterns in Dictyostelium development. *Pac. Symp. Biocomput.* 335-347 (2001).

Sharan, R. & Shamir, R. CLICK: a clustering algorithm with applications to gene expression analysis. *Ismb* **8**, 307-316 (2000).

Shen-Orr, S., Milo, R., Mangan, S. & Alon, U. Network motifs in the transcriptional regulation network of Escherichia coli. *Nature Genetics* **31**, 64-68 (2002).

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E.S. & Golub, T.R. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci.* **96**, 2907-2912 (1999).

Toronen, P., Kolehmainen, M., Wong, G. & Castren, E. Analysis of gene expression data using self-organizing maps. *FEBS Lett.* **451**, 142-146 (1999).

Varma, A., & Palsson, B. O. Metabolic flux balancing: basic concepts, scientific and practical use. *Bio/Technology* **12**, 994-998 (1994).

Wang, W., Cherry, J.M., Botstein, D. & Li, H. A systematic approach to reconstructing transcription networks in Saccharomyces cerevisiae. *Proc. Natl. Acad. Sci.* **99**, 16893-16898 (2002).

Appendices

Appendix 1 – The function distribution of yeast proteins (CYGD: http://mips.gsf.de/genre/proj/yeast/).

Function	Number of ORFs
METABOLISM	1488
ENERGY	363
CELL CYCLE AND DNA PROCESSING	995
TRANSCRIPTION	1061
PROTEIN SYNTHESIS	473
PROTEIN FATE (folding, modification, destination)	1130
PROTEIN WITH BINDING FUNCTION OR COFACTOR REQUIREMENT (structural or catalytic)	1019
PROTEIN ACTIVITY REGULATION	237
CELLULAR TRANSPORT, TRANSPORT FACILITATION AND TRANSPORT ROUTES	1028
CELLULAR COMMUNICATION/SIGNAL TRANSDUCTION MECHANISM	233
CELL RESCUE, DEFENSE AND VIRULENCE	552
INTERACTION WITH THE CELLULAR ENVIRONMENT	457
INTERACTION WITH THE ENVIRONMENT (Systemic)	8
TRANSPOSABLE ELEMENTS, VIRAL AND PLASMID PROTEINS	123
CELL FATE	268
DEVELOPMENT (Systemic)	70
BIOGENESIS OF CELLULAR COMPONENTS	844
CELL TYPE DIFFERENTIATION	448
UNCLASSIFIED PROTEINS	2054
Total (average number of functions per ORF)	12851 (12851/6756 = 1.9022)

Appendix 2 – The location distribution of yeast proteins (CYGD: http://mips.gsf.de/genre/proj/yeast/).

Location	Number of ORFs
extracellular	52
bud	120
cell wall	37
cell periphery	190
plasma membrane	165
integral membrane / endomembranes (if not assigned to a specific membrane)	175
cytoplasm	2785
cytoskeleton	176
ER	529
golgi	117
transport vesicles	130
nucleus	2055
mitochondria	1013
peroxisome	49
endosome	51
vacuole	257
microsomes	5
lipid particles	24
punctate composite	135
ambiguous	220
Total (average number of locations per ORF)	8285 (8285/5200 = 1.5933)

Appendix 3 – The biochemical pathways of yeast (KEGG: http://www.genome.jp/kegg/).

Pathway	Number of ORFs
Purine metabolism	90
Starch and sucrose metabolism	71
Pyrimidine metabolism	71
Oxidative phosphorylation	62
Glycerolipid metabolism	52
Glycolysis / Gluconeogenesis	47
Glycine, serine and threonine metabolism	43
Benzoate degradation via CoA ligation	41
Aminoacyl-tRNA biosynthesis	37
Pyruvate metabolism	34
Nicotinate and nicotinamide metabolism	33
Galactose metabolism	32
Inositol phosphate metabolism	31
Fructose and mannose metabolism	31
Butanoate metabolism	30
N-Glycan biosynthesis	30
Citrate cycle (TCA cycle)	30
Lysine degradation	29
Alanine and aspartate metabolism	27
Glutamate metabolism	27
Pentose phosphate pathway	27
ATP synthesis	25
Phenylalanine, tyrosine and tryptophan biosynthesis	23
Bile acid biosynthesis	23
Arginine and proline metabolism	22
Tyrosine metabolism	21
Histidine metabolism	21
Lysine biosynthesis	20
Selenoamino acid metabolism	19
Carbon fixation	18
Glycosylphosphatidylinositol(GPI)-anchor biosynthesis	18
Aminosugars metabolism	18
Tryptophan metabolism	18
Glycosphingolipid metabolism	17
Fatty acid metabolism	17
Nitrogen metabolism	16

Folate biosynthesis	16
Valine, leucine and isoleucine biosynthesis	16
One carbon pool by folate	15
Glyoxylate and dicarboxylate metabolism	15
Nucleotide sugars metabolism	15
High-mannose type N-glycan biosynthesis	15
Urea cycle and metabolism of amino groups	15
Methionine metabolism	14
Riboflavin metabolism	13
Ascorbate and aldarate metabolism	13
Sulfur metabolism	12
Porphyrin and chlorophyll metabolism	12
Valine, leucine and isoleucine degradation	12
Ubiquinone biosynthesis	12
Biosynthesis of steroids	12
Reductive carboxylate cycle (CO2 fixation)	11
Propanoate metabolism	11
Glutathione metabolism	11
Cyanoamino acid metabolism	11
Aminophosphonate metabolism	11
Phenylalanine metabolism	11
Limonene and pinene degradation	10
Pantothenate and CoA biosynthesis	10
Tetrachloroethene degradation	10
Cysteine metabolism	10
gamma-Hexachlorocyclohexane degradation	9
Alkaloid biosynthesis II	8
beta-Alanine metabolism	8
Biotin metabolism	7
Methane metabolism	7
Vitamin B6 metabolism	6
Streptomycin biosynthesis	6
Pentose and glucuronate interconversions	6
Terpenoid biosynthesis	5
Phospholipid degradation	5
Androgen and estrogen metabolism	5
Thiamine metabolism	4
Nitrobenzene degradation	4
Globoside metabolism	4
Fatty acid biosynthesis (path 1)	4
Ganglioside biosynthesis	3
Blood group glycolipid	3

biosynthesis-neolactoseries	
O-Glycan biosynthesis	3
Novobiocin biosynthesis	3
Fatty acid biosynthesis (path 2)	3
Styrene degradation	2
1,4-Dichlorobenzene degradation	2
Prostaglandin and leukotriene metabolism	2
Taurine and hypotaurine metabolism	2
Benzoate degradation via hydroxylation	2
Synthesis and degradation of ketone bodies	2
Peptidoglycan biosynthesis	1
C21-Steroid hormone metabolism	1

Appendix 4 – The list of references to expression data.

Gerber AP, *et al.* (2004) PLoS Biol 2(3):E79 Extensive Association of Functionally and Cytotopically Related mRNAs with Puf Family RNA-Binding Proteins in Yeast

Hurowitz EH and Brown PO (2003) Genome Biol 5(1):R2 Genome-wide analysis of mRNA lengths in Saccharomyces cerevisiae

Fernandes PM, *et al.* (2004) FEBS Lett 556(1-3):153-60 Genomic expression pattern in Saccharomyces cerevisiae cells in response to high hydrostatic pressure

Shakoury-Elizeh M, *et al.* (2004). Mol Biol Cell 15(3):1233-43 Transcriptional Remodeling in Response to Iron Deprivation in Saccharomyces cerevisiae

Shepard KA, *et al.* (2003) . Proc Natl Acad Sci U S A 100(20):11429-34 Widespread cytoplasmic mRNA transport in yeast: identification of 22 bud-localized transcripts using DNA microarray analysis

Troyanskaya OG, *et al.* (2003). Proc Natl Acad Sci U S A 100, 8348-53 A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae)

Nagy PL, *et al.* (2003) Proc Natl Acad Sci U S A 100(11): 6364-6369 Genomewide demarcation of RNA polymerase II transcription units revealed by physical fractionation of chromatin

Segal E, *et al.* (2003). Nat Genet 34(2):166-176 Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data

Arava Y, *et al.* (2003). Proc Natl Acad Sci U S A 100(7):3889-94 Genome-wide analysis of mRNA translation profiles in Saccharomyces cerevisiae

Alter O, *et al.* (2003) Proc Natl Acad Sci USA 100(6):3351-3356 Generalized singular value decomposition for comparative analysis of genome-scale expression datasets of two different organisms

Dunham MJ, *et al.* (2002) . Proc Natl Acad Sci USA 99:16144-9. Characteristic genome rearrangements in experimental evolution of Saccharomyces cerevisiae

Yoshimoto H, *et al.* (2002). J Biol Chem 277(34):31079-31088 Genome-wide Analysis of Gene Expression Regulated by the Calcineurin/Crz1p Signaling Pathway in

Saccharomyces cerevisiae

Wang Y, *et al.* (2002). Proc Natl Acad Sci U S A 99(9):5860-5 Precision and functional specificity in mRNA decay

Rutherford JC, *et al.* (2001) Proc Natl Acad Sci U S A 98(25):14322-7 A second iron-regulatory system in yeast independent of Aft1p

Protchenko O, *et al.* (2001) J Biol Chem 276(52):49244-50 Three cell wall mannoproteins facilitate the uptake of iron in Saccharomyces cerevisiae

Gasch AP, *et al.* (2001).Mol Biol Cell 12(10):2987-3003 Genomic expression responses to dna-damaging agents and the regulatory role of the yeast atr homolog mec1p

Keller G, *et al.* (2001) J Biol Chem 276(42):38697-702 Haa1, a protein homologous to the copper-regulated transcription factor Ace1, is a novel transcriptional activator

Lieb JD, *et al.* (2001) Nat Genet 28(4):327-334 Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association

Carmel-Harel O, *et al.* (2001) Mol Microbiol 39(3):595-605 Role of thioredoxin reductase in the Yap1p-dependent response to oxidative stress in Saccharomyces cerevisiae

Iyer VR, *et al.*(2001) Nature 409:533-38 Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF

Kuhn KM, *et al.* (2001) Mol Cell Biol 21(3):916-27 Global and specific translational regulation in the genomic response of Saccharomyces cerevisiae to a rapid transfer from a fermentable to a nonfermentable carbon source

Ogawa N *et al.*(2000) Mol Biol Cell 11:4309-21 New components of a system for phosphate accumulation and polyphosphatemetabolism in saccharomyces cerevisiae revealed by genomic expressionanalysis

Gasch AP, *et al.* (2000) Mol Biol Cell 11(12):4241-57 Genomic expression programs in the response of yeast cells to environmental changes

Alter O, *et al.* (2000). Proc Natl Acad Sci USA 97(18):10101-6 Singular value decomposition for genome-wide expression data processing and modeling

Gross C, *et al.* (2000) J Biol Chem 275(41):32310-6 Identification of the copper regulon in Saccharomyces cerevisiae by DNA microarrays

Zhu G, *et al.* (2000) Nature 406(6791):90-4 Two yeast forkhead genes regulate the cell cycle and pseudohyphal growth

Casagrande R, *et al.* (2000). Mol Cell 5 (4):729-35 Degradation of proteins from the ER of S. cerevisiae requires an intact unfolded protein response pathway

Lyons TJ, *et al.* (2000) Proc Natl Acad Sci U S A 97(14):7957-62 Genome-wide characterization of the Zap1p zinc-responsive regulon in yeast

Diehn M *et al.*(2000) Nat Genet 25:58-62 Large-scale identification of secreted and membrane-associated gene products using DNA microarrays

Yun CW, *et al.* (2000) J Biol Chem 275(14):10709-15 Desferrioxamine-mediated iron uptake in Saccharomyces cerevisiae. Evidence for two pathways of iron uptake

Sudarsanam P *et al.*(2000) Proc Natl Acad Sci U S A 97:3364-9 Whole-genome expression analysis of snf/swi mutants of Saccharomycescerevisiae

Ferea TL, *et al.* (1999) Proc Natl Acad Sci U S A 96(17):9721-6 Systematic changes in gene expression patterns following adaptive evolution in yeast

Spellman PT *et al.*(1998) Mol Biol Cell 9:3273-97 Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization

Chu S, *et al.* (1998) Science 282(5389):699-705 The transcriptional program of sporulation in budding yeast

DeRisi JL, *et al.* (1997) Science 278(5338):680-6 Exploring the metabolic and genetic control of gene expression on a genomic scale

Appendix 5 – Matlab code for loop enumeration.

```
function [loops]=findloop(ajm,n_r,k)
%FINDLOOP    find loops
%    [LOOPS]=FINDLOOP(A,N_R,K) finds loops with length less than K for a
%    network given its adjacent matrix. N_R is the number of reactions. All
%    reaction indices are smaller than compound indices.
%
%    Tony
%    03/07/2005

t1=clock;

n=length(ajm);%number of nodes
n_c=n-n_r;%number of compounds

loops=cell(k-1,1);
for s_cpd=n_r+1:n%go through all compounds
    fprintf(1,'%d\n',n-s_cpd);
    paths_cpd=cell(k/2,1);
    for depth=2:2:k%go through all possible loop depths
        paths=[];
        %work only on the local network with maximal depth specified
        idx=s_cpd;
        for i=1:depth+1
            [zi,zj]=find(ajm(idx,:)~=0);
            idx_new=setdiff(zj,idx);
            idx=[idx idx_new];
        end
        if ~isempty(idx)
            %eliminate dead ends and reactions with only substrates or
            %products
            rcts=idx(find(idx<=n_r));
            cpds=idx(find(idx>n_r));

            delta1=1;
            while delta1>0
                %remove nodes with zero or one connection
                delta2=1;
                while delta2>0
                    idx_rm=idx(find(sum(abs(ajm(idx,idx)),1)<=1));
                    idx=setdiff(idx,idx_rm);
                    delta2=length(idx_rm);
                end
                %remove reactions with either only substrates or products
                rcts=idx(find(idx<=n_r));
                cpds=idx(find(idx>n_r));
                rcts_rm=rcts(find(abs(sum(ajm(cpds,rcts),1))==sum(abs(ajm(cpds,rcts)),1)));
                rcts=setdiff(rcts,rcts_rm);
                delta1=length(rcts_rm);
                idx=[rcts cpds];
            end
            ajm_local=ajm(idx,idx);
            paths=find(idx==s_cpd);%start with the s_cpd
            if ~isempty(paths)&&length(cpds)*length(rcts)>0
                paths=getpaths(ajm_local,paths,depth+1);
            end
            idx=idx(:);
            if ~isempty(paths)
                paths=idx(paths);%map back the indices of the original network
            end
        end
        paths_cpd{depth/2}=paths;
    end %end of for depth=2:2:k
    %compare paths and identify loops
    loops_cpd=cell(k-1,1);
    %find loops with length 2*depth
    for i=1:k/2
        loop=[];
        path=paths_cpd{i};
        if ~isempty(path)
            [zu,zf]=unique2(path(end,:));
            [ztf,zloc]=ismember(path(end,:),zu(find(zf==1)));
            path(:,ztf)=[];%eliminate paths with unique end compounds
            [n_d,n_p]=size(path);
            for zi=1:n_p-1
                for zj=zi+1:n_p
                    if path(end,zi)==path(end,zj)
                        if isempty(intersect(path(2:end-1,zi),path(2:end-1,zj)))
                            zloop=[path(:,zi);flipud(path(2:end-1,zj))];
                            loop=[loop zloop];
                        end
                    end
                end
            end
        end
        loops_cpd{2*i-1}=loop;
    end
    %find loops with length 2*depth-2
```

```
    for i=2:k/2
        loop=[];
        path1=paths_cpd{i-1};
        path2=paths_cpd{i};
        if ~isempty(path1)&&~isempty(path2)
            for zi=1:size(path1,2)
                for zj=1:size(path2,2)
                    if path1(end,zi)==path2(end,zj)
                        if isempty(intersect(path1(2:end-1,zi),path2(2:end-1,zj)))
                            zloop=[path1(:,zi);flipud(path2(2:end-1,zj))];
                            loop=[loop zloop];
                        end
                    end
                end
            end
        end
        loops_cpd{i*2-2}=loop;
    end

    %add to loops
    for i=1:k-1
        loops{i}=[loops{i} loops_cpd{i}];
    end
end
%eliminate redundant loops
for i=1:k-1
    loops_k=loops{i};
    if ~isempty(loops_k)
        sorted_loops=sort(loops_k,1);
        [zu_loops,zf,idx]=unique2(sorted_loops','rows');
        loops_k=loops_k(:,idx);
    end
    loops{i}=loops_k;
end

e=etime(clock,t1);
fprintf(1,'elapsed time: %d mins %d seconds\n',floor(e/60),mod(e,60));

%recursive algorithm to find all the possible paths starting from a given
%compound, with specified depth.
function [paths]=getpaths(ajm,paths,depth)
[nr,nc]=size(paths);
if nr>depth
    error('path length exceeds the maximal depth');
```

```
end
if nr==depth||nr==0
    paths=paths;
else
    zpaths=[];
    for i=1:size(paths,2)
        cpd=paths(end,i);
        rcts=find(ajm(cpd,:)~=0);
        rcts=setdiff(rcts,paths(:,i));%eliminate the paths that form smaller loops
        for j=1:length(rcts)
            rct=rcts(j);
            cpds_next=find(ajm(:,rct)==-ajm(cpd,rct));
            cpds_next=setdiff(cpds_next,paths(:,i));%eliminate the paths that form smaller loops
            if ~isempty(cpds_next)
                zt=[repmat(paths(:,i),1,length(cpds_next));repmat(rct,1,length(cpds_next));cpds_next'];
                zpaths=[zpaths zt];
            end
        end
    end
    paths=zpaths;
    paths=getpaths(ajm,paths,depth);
end
```

Curriculum Vitae

The Theoretical Biology Research Lab
Department of Physics, Hong Kong Baptist University
Kowloon Tong, Hong Kong
Phone: (852) 3411 5156, E-mail: shui@phys.hkbu.edu.hk
URL: http://cmt.hkbu.edu.hk/~tonyhui

RESEARCH INTERESTS

Microarray data analysis, Transcription regulatory network, Metabolic network, RNA folding, Protein folding

EDUCATION

02/2003-05/2005 **M. Phil.** in Physics, expected, Hong Kong Baptist University
Thesis title: *The analysis of metabolism in Saccharomyces cerevisiae with genome-scale gene expression data*
Supervisor: Prof. L.H. Tang, Co-supervisor: Prof. N.H. Cheung

09/1999-06/2002 **B. Sc** in Physics (First Class Hon., 1 out of 45), Hong Kong Baptist University
Final year honors project: *RNA second structure prediction with pseudoknots*
Supervisor: Dr. L.H. Tang
GPA: 3.67/4.0

10/1998-6/1999 Tsinghua University and Hong Kong Baptist University Foundation Course

PROFESSIONAL ASSOCIATION

- Member of American Physical Society
- Member of Hong Kong Physical Society

AWARDS/SCHOLARSHIP

- Foundation Scholarship for Undergraduate Students from Mainland, 1998~1999
- Hong Kong Jockey Club Scholarship for Outstanding Mainland Students, 1999~2002
- Wu Dayou Summer Camp with Nobel Laureates, Taiwan, 2001
- The Lui Ming Choi Scholarship, Hong Kong Baptist University, in 2002

RESEARCH/WORK EXPERIENCE

02/2003-present Department of Physics, Hong Kong Baptist University

Research student

- Yeast microarray data analysis, e.g., regulatory module finding

- Topological analysis of yeast and E.coli transcription regulatory network

- Construction and module finding of yeast metabolic network

02/2003-06/2003 Department of Physics, Hong Kong Baptist University

Coordinator for postgraduate student seminar

- Invite speakers and organize weekly lunch seminars

12/2002-01/2003 Institute of Physics, Chinese Academy of Sciences

Visiting student at the Soft Condensed Matter Lab

12/2001-01/2002 Peking University

Visiting student at the Center for Theoretical Biology

09/2000-09/2001 Hong Kong Baptist University,

Academic Executive of Physics Society

- Organize Academic Week of Physics

COMPUTER SKILLS

- Operating Systems: expert level on Windows, Linux/Unix
- Languages & Software: extensive experience on Fortran, Perl, Matlab

FOREIGN LANGUAGE PROFICIENCY

- Fluent in English
- TOEFL score: 270/300
- GRE score: 2140/2400

PUBLICATIONS/PRESENTATIONS

1. **Sheng Hui**, Lei-han Tang, *Analysis of metabolic network with genome expression data in yeast*. In preparation
2. **Sheng Hui**, Lei-han Tang, *RNA secondary structure prediction with pseudoknots*. Presented at the 2003 Annual Conference of Physical Society of Hong Kong

April 2005