

**METABOLIC NETWORK BASED GENE ESSENTIALITY
ANALYSIS**

MA JING

NATIONAL UNIVERSITY OF SINGAPORE

2012

**METABOLIC NETWORK BASED GENE ESSENTIALITY
ANALYSIS**

MA JING

(B.Sc., Tsinghua University)

**A THESIS SUBMITTED FOR THE DEGREE OF DOCTOR OF
PHILOSOPHY**

**NUS GRADUATE SCHOOL FOR INTEGRATIVE SCIENCES
AND ENGINEERING**

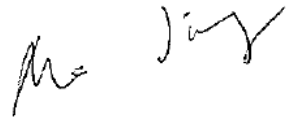
NATIONAL UNIVERSITY OF SINGAPORE

2012

Declaration

I hereby declare that the thesis is my original work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the thesis.

This thesis has also not been submitted for any degree in any university previously.

A handwritten signature in black ink, appearing to read 'Ma Jing', is positioned above a horizontal line.

Ma Jing

2 SEP 2013

Acknowledgements

Many people have helped in this dissertation, and it is my pleasure to present my thanks to all of them who made this thesis possible.

First and foremost, I would like to present my sincerely gratitude to my supervisor, Professor Li Baowen, for his invaluable guidance and continuous support during these four years. Without his advice, I could not have finished this project. Especially, I am impressed by his personality and very grateful to my supervisor for his unique way to train post-graduates. I learned how to do projects independently. I cannot imagine a better supervisor for my Ph.D. study.

Besides my supervisor, I would like to thank the members of my thesis advisory committee, Professor Chen Yu-Zong, Professor Gong Zhiyuan, and Professor Low Boon Chuan for their insightful comments and all the suggestions during the oral examination. Their insights and comments are really appreciated.

Many thanks go to Mr. Zhang Xun, who is not only my best friend, but also a great collaborator. Thank you for being there during the time I was bothered by disease and all your encouragement when I was frustrated by the project. I would also like to thank Dr. Ung Choong Yong, for the valuable advice and pleasant cooperation. I benefit a lot from his positive attitude. I am also grateful to Prof.

Yang Huijie for introducing me to the world of biological networks. I would also like to thank Dr. Ren Jie for his insightful ideas and discussions. I am inspired by his passion for academia. I would also thank Mr. Liu Sha for sharing experiences in writing code, which is especially important for my project.

My appreciation likewise extends to all the CCSE members. I would like to thank Ms. Zhang Kaiwen for all the pleasant time we have spent and Ms. Zhu Guimei for the care and the delicious soup when I was sick. Many thanks go to Dr. Ni Xiaoxi, and Dr. Wu Xiang for sharing job hunting experience. I would also appreciate all the happy lunch time spent with Mr. Zhu Feng, Mr. Tang Qinglin, Ms. Yang Lina, Ms. Qin Chu, Mr. Hou Ruizhen, Mr. Tao Lin, and Mr. Feng Ling. I am also grateful to all the other members for their comments, advice during the group meeting.

Last but most importantly, I would like to thank my beloved parents for giving birth to me and supporting me throughout my life. I am also grateful to my dearest brothers, for all the encouragement and care when I met difficulties during my life. I love you all. To them I dedicate this thesis.

Table of Contents

Acknowledgements.....	i
Table of Contents.....	iii
Summary.....	viii
List of Figures.....	i
List of Tables.....	i
List of Publications.....	i
Chapter 1 Introduction.....	1
1.1 An introduction to systems biology.....	2
1.2 Overview of network biology.....	4
1.3 Mathematical representation of biological network (adjacency matrix).....	8
1.4 Characterizing the topological features of complex networks.....	10
1.4.1 Degree and degree distribution.....	10
1.4.2 Path and shortest path.....	14
1.4.3 Clustering coefficient.....	14
1.4.4 Centrality.....	15
1.5 The application of network biology in drug development.....	16
1.6 Identification and analysis of essential genes.....	18
1.6.1 Molecular basis of gene mutation.....	19
1.6.2 Experimental studies on essential genes.....	22
1.6.3 Identification of essential genes via computational approaches.....	23

1.6.4 Single gene deletion analysis.....	25
1.6.5 Evolution of essential and non-essential genes	27
1.7 Genetic interactions.....	29
1.8 Objectives and outline of this thesis.....	31
1.8.1 Objectives of this thesis	31
1.8.2 Outline of this thesis	32
Chapter 2 Materials and methods	35
2.1 Biological databases.....	35
2.1.1 Kyoto Encyclopedia of Genes and Genomics (KEGG)	35
2.1.2 BiGG database.....	36
2.2 Metabolic network reconstruction.....	37
2.2.1 Data source for network reconstruction.....	37
2.2.2 Gene-protein-reaction (GPR) association.....	38
2.2.3 Metabolic network representation and visualization.....	39
2.3 Characterize gene deletion effects.....	43
2.3.1 Corresponding reactions	43
2.3.2 Cascading failure procedures.....	45
2.4 Flux balance analysis (FBA).....	49
2.4.1 Mathematical representation of FBA	50
2.5 Essential gene lists	56
Chapter 3 Single gene deletion analysis	58

3.1 Introduction	58
3.2 Materials and methods	61
3.2.1 Statistical hypothesis test.....	61
3.2.2 Null model	62
3.2.3 Gene essentiality information.....	62
3.2.4 Metabolic network reconstruction	62
3.2.5 Computational algorithm of cascading failure in metabolic network ...	63
3.3 Single gene deletion	63
3.3.1 Single gene deletion in metabolic network.....	63
3.3.2 Essential gene deletion induces large damage list.....	64
3.3.3 Analysis of components within damage list	66
3.3.4 Genes with similar damage lists share the same essentiality.....	69
3.3.5 Associated gene sets are necessary for survival	76
3.3.6 Structural organization of essential and non-essential genes in metabolic network	87
3.4 Discussions.....	90
Chapter 4 Essential gene prediction.....	95
4.1 Overview of essential gene predictions.....	95
4.2 Materials and Methods	98
4.2.1 Identification of functional and network topological features associated with gene essentiality.....	98
4.2.2 Self-devised algorithm for predicting gene essentiality	100
4.3 Gene essentiality prediction	104

4.3.1 Results of gene essentiality prediction	104
4.3.2 The effect of iteration number on prediction accuracy.....	105
4.3.3 The effect of threshold on prediction accuracy	106
4.3.4 The effect of top N similar genes on prediction accuracy.....	110
4.3.5 The effect of cycle number on prediction accuracy	111
4.3.6 Prediction accuracy for different number of unknown genes.....	112
4.3.7 Strategies to increase the prediction accuracy	116
4.3.8 Cross-species validation	120
4.4 Discussions.....	123
Chapter 5 Large-scale epistasis analysis in <i>E. coli</i> and <i>S. cerevisiae</i>	127
5.1 Introduction	127
5.2 Materials and methods	130
5.2.1 Metabolic network.....	130
5.2.2 Deletion effects in response to double gene removal.....	131
5.2.3 Shortest path distance	132
5.3 Double gene deletions in <i>E. coli</i>	133
5.3.1 Comparisons between damage lists for single- and double- gene deletions.....	133
5.3.2 Pathway analysis for gene pairs with enhanced and reduced deletion effects.....	137
5.4 Double gene deletions in <i>S. cerevisiae</i>	141
5.5 Gene pairs with reduced deletion effects arise from the same pathway and are conserved between species	146

5.6 Gene pairs with enhanced deletion effects disturb key reactions or metabolites	152
5.7 Discussions.....	157
Chapter 6 Summary and future work.....	162
6.1 Major findings and contributions	162
6.2 Future work	166
Bibliography	169
Appendix 1: C++ Code for identifying corresponding reactions	192
Appendix 2: Pseudo Code for identifying damage lists for single gene deletion	193

Summary

Essential genes are widely recognized as ideal drug target candidates since their deletion can lead to lethality phenotypes. In most organisms, however, only a small fraction of genes are required for viability while the majority of genes are dispensable. Studies on the cause of the preponderance of dispensable genes revealed some functional backup mechanisms used for non-essential genes, such as genetic buffering by duplicate genes and the presence of alternative pathways. To further enrich our understanding on the cause of gene essentiality/dispensability, it is necessary to explore the altered biological network in response to single/multiple gene deletions and the epistatic interactions between mutants. Understanding the causes and evolution of gene essentiality can also be beneficial to explore more efficient way to identify essential gene.

To understand the distinct deletion effects between essential and non-essential genes, we proposed a measure called 'damage list' to characterize single gene deletion effects in the context of metabolic networks. Our analysis showed that the size of damage list for essential genes is generally larger than non-essential genes. Moreover, it was observed that while essential genes can exert its deletion effects on both essential and non-essential genes, non-essential genes can only impact on other non-essential genes. Further analysis suggested that genes sharing highly similar damage lists tend to possess the same essentiality. It was also found out that essential genes spread its deletion effects through certain 'associated gene sets' whose cooperative effects is to block the production of key metabolites. The

failure of these metabolites finally disrupted normal cellular growth. Structural organization of essential and non-essential genes also supported our findings since essential genes preferred to be interconnected through low-degree metabolites while non-essential genes preferred to be interconnected through high-degree metabolites. Our analysis suggested a possible mechanism regarding how essential and non-essential genes are differentiated.

We then moved our analysis to gene essentiality prediction. Some features that are tightly associated with essentiality were revealed. Using these features, a method was proposed to predict gene essentiality in the context of metabolic networks, by using limited number of known type genes. With optimized parameters, we computed the prediction accuracy and found that the prediction accuracy is quite high and robust, compared with other approaches, such as Flux Balance Analysis and other machine-learning based approaches. Our studies therefore emphasized that understanding the topological and functional features that tightly associated with gene essentiality is crucial for high-accuracy prediction of essential genes.

Studies on epistasis between mutants identified gene pairs with either enhanced or reduced deletion effects. Further investigation on gene pairs with reduced deletion effects indicated that they are generally from the same pathway and their path distance in metabolic network is significantly reduced. Essential genes pairs with reduced deletion effects (i.e. those candidates for synthetic rescue) are located in pathways such as Metabolism of Terpenoids and Polyketides and Metabolism of Cofactors and Vitamins. Intriguingly, essential genes pairs from these two

pathways of *Escherichia coli* and *Saccharomyces cerevisiae* are orthologous. On the contrary, analysis of gene pairs with enhanced deletion effects suggested that they are mainly arising from diverse pathways. Further investigations on non-essential gene pairs with enhanced deletion effects indicated that some additional metabolites/reactions/genes are disrupted compared with the damage caused by individual gene removal. Their loss can sometimes lead to cellular lethality, implying their suitability as candidates for synthetic lethal gene pairs. The strong dependency between gene essentiality and biological network topological organization emphasized the importance and necessity to integrate the network analysis into gene essentiality research.

List of Figures

Figure 1 Adjacency matrix for (A) undirected network, (B) directed network and (C) weighted network.	6
Figure 2 Topological features of complex network.....	11
Figure 3 Degree distribution of scale-free network.	13
Figure 4 Samples of exported SBML file from BiGG database.....	39
Figure 5 The glycolysis pathway which converts glucose to pyruvate	41
Figure 6 Reconstructed metabolic network.	42
Figure 7 Gene-protein-reaction Association.....	45
Figure 8 Schematic diagram for network cascading failure for single gene deletion.	47
Figure 9 Overview of the self-devised algorithm.	49
Figure 10 Cumulative distribution of damage size	66
Figure 11 Distinct deletion effects between essential and non-essential genes.....	68
Figure 12 Overview of damage list similarity	70
Figure 13 Essentiality consistency within gene subnetworks sharing similar damage lists.....	71
Figure 14 Evolution of the largest subnetwork.....	74
Figure 15 Different types of metabolites.	88
Figure 16 Structural organizations of essential and non-essential genes.....	89
Figure 17 Proposed mechanisms for the differential single gene deletion effects.	94
Figure 18 Overview of the algorithm.....	103
Figure 19 Prediction results for different number of unknown genes.	116
Figure 20 Prediction accuracy combined with FBA.....	118
Figure 21 The prediction accuracy corresponding to different level of essential genes to be predicted.....	119

Figure 22 Prediction results for yeast <i>S. cerevisiae</i>	122
Figure 23 Epistatic interactions in <i>E. coli</i> between (A) E-E pairs, (B) E-N pairs, and (C) N-N pairs.....	137
Figure 24 Pathway analysis of gene pairs in <i>E. coli</i>	139
Figure 25 Distribution of shortest path distance between E-E pairs.....	140
Figure 26 Distribution of shortest path distance between N-N pairs.....	141
Figure 27 Epistatic interactions in <i>S. cerevisiae</i> between (A) E-E pairs, (B) E-N pairs, and (C) N-N pairs.....	144
Figure 28 Pathway analysis of gene pairs in <i>S. cerevisiae</i>	145

List of Tables

Table 1 Genome-scale metabolic networks.	5
Table 2 13 categories of metabolism classification in KEGG database	36
Table 3 26 currency metabolites.	42
Table 4 Biomass components.	54
Table 5 Damage list composition.	67
Table 6 Essentiality consistency within genes sharing highly similar damage lists in <i>E. coli</i>	73
Table 7 Essentiality consistency within genes sharing highly similar damage lists in <i>S. Cerevisiae</i>	73
Table 8 Top five GO functions in the largest subnetwork.....	75
Table 9 Associated gene sets	79
Table 10 Comparison between Keio Collection and PEC database.	86
Table 11 The number of gene pairs with similarity coefficient at different level threshold.....	100
Table 12 The effect of iteration number on prediction accuracy.....	106
Table 13 The effect of threshold on the prediction accuracy (800 unknown).....	108
Table 14 The effect of threshold on the prediction accuracy (100 unknown).....	109
Table 15 The effect of Top N similar genes on the prediction accuracy.....	110
Table 16 The effect of cycle number on prediction accuracy.....	112
Table 17 Summary of the prediction results for <i>E. coli</i>	114
Table 18 Summary of the prediction results for yeast <i>S. cerevisiae</i>	121
Table 19 Summary of pathway analysis in <i>E. coli</i>	138
Table 20 Summary of pathway analysis for <i>S. cerevisiae</i>	144
Table 21 E-E gene pairs with reduced deletion effects (<i>E. coli</i>)	148

Table 22 E-E gene pairs with reduced deletion effects (<i>S. cerevisiae</i>).....	149
Table 23 Gene pairs from Metabolism of Terpenoids and Polyketides.....	151
Table 24 N-N gene pairs with enhanced deletion effects (<i>E. coli</i>)	154
Table 25 N-N gene pairs with enhanced deletion effects (<i>S. cerevisiae</i>)	155
Table 26 Detailed analysis of N-N gene pairs with enhanced deletion effects ...	156

List of Publications

1. **Ma J**, Zhang X, Ung CY, Chen YZ, Li B: Metabolic network analysis revealed distinct routes of deletion effects between essential and non-essential genes. *Mol Biosyst* 2012, 8(4):1179-1186.
2. Ung CY, Lam SH, Zhang X, Li H, **Ma J**, Zhang L, Li B, Gong Z: Existence of inverted profile in chemically responsive molecular pathways in the zebrafish liver. *PLoS ONE* 2011, 6(11):e27819.
3. Zhang X, Ung CY, Lam SH, **Ma J**, Chen YZ, Zhang L, Gong Z and Li B: Toxicogenomic analysis suggests chemical-induced sexual dimorphism in the expression of metabolic genes in zebrafish liver. *PLoS ONE* 2012, 7(12): e51971.

Chapter 1 Introduction

Molecular biology has uncovered a multitude of biological facts, such as providing explanations for phenomena observed at the molecular level, resolving proteins functions, and revealing the molecular basis of diseases. Yet, the traditional reductionist approaches seldom provide a comprehensive understanding of systemic issues related to genetic interactions, robustness of biological systems and evolutionary genomics. This is because a biological system is not a simple assembly of genes and proteins, but involves intricate interactions between them. Fortunately, the holistic exploratory approaches combined with high throughput data collection technologies make it possible to analyze biological processes and regulations from a systematic perspective.

Robustness is one of the key characteristics of biological systems, which refers to the phenomenon that many biological systems keep functioning in response to a random attack. On the other hand, indispensable components which make the systems relatively fragile and easily attacked are considered as ideal drug targets, especially for infectious diseases. An in-depth understanding on these essential components may help us to unveil the structural and functional organizations of complex networks and gain new insights into drug design and development.

In this thesis, we focused on essential genes analysis in the context of metabolic networks. Problems such as how deletion of essential and non-essential genes

exerts their distinct phenotypes (lethality/survival) and how they are associated with the structural organizations were addressed. The network topological features associated with gene essentiality were applied to predict essential and non-essential genes, aiming to explore some efficient computational approaches that can handle with limited data. Beyond single gene analysis, the interactions between pairs of genes were also studied to unveil a higher order organization of biological networks and obtain a holistic understanding of biological robustness.

1.1 An introduction to systems biology

Systems biology is a newly emerging, multi-disciplinary field which studies the mechanisms regarding how the components of complex biological systems interact functionally.

High throughput technologies [1-4] developed in the post-genomic era such as next-generation sequencing, and microarrays have generated huge quantities of multidimensional data and made it possible to analyze biological processes systematically. For example, next-generating sequencing can sequence a genome up to 1.5G base pairs in 2.5 days with more than 99% accuracy, which is rather more compared with the traditional Sanger-based sequencing approaches, for which only 3M base pairs can be generated every three days [5]. These data can be used for multi-dimension analysis, such as mRNA expression study [6-8],

transcriptional active sites identification [9, 10], and protein-DNA interactions [11, 12] and so on.

The focus of systems-level understanding of biological processes is to gain a comprehensive insight into a system's structural organization, topological features and collective dynamics, rather than an individual's. Two types of research methodologies are widely used in system biology studies: (1) knowledge discovery, aiming to explore the hidden patterns or information from the available experimental data, and then form a hypothesis; (2) simulation-based analysis, which tests hypothesis using *in silico* experiments.

Knowledge discovery is the process of data integration and interpretation which extracts hidden patterns or new links by exploring the huge amount of data, and thereby derives some fundamental and applied biological information about the whole system. Approaches such as machine learning, statistical discriminators, and other sophisticated algorithms are widely used [13, 14]. By applying these approaches, researchers have elicited some interesting patterns and enabled this information or data that resides in large data repositories to be converted into new biological insights.

Simulation-based studies are hypothesis-driven, which attempt to make predictions based on the hypothesis that has been validated by experimental data.

It is composed two steps: (1) Hypothesis validation. Based on the analysis scope and the available biological data, a hypothesis-driven model is constructed. The results of *in silico* experiments are compared with experimental data. Inconsistency between them indicates that the assumption made within the model is incomplete, and as a consequence the model will be rejected or modified. (2) Predictions. Models that can survive the initial validation can be used to make predictions which will further be tested by wet experiments. Approaches such as systems of ordinary differential equations (ODEs), partial differential equations (PDEs), stochastic differential equations (SDEs), and Petri Nets can be applied in simulation-based studies.

1.2 Overview of network biology

One way to systematically understand biological systems is to represent biological interactions as computable networks, which can facilitate our analysis on how these interactions contribute to the structural and functional organization of a living cell. The massive information emerging from the fields of genomics, transcriptomics, proteomics, and metabolomics makes it possible to reconstruct genome-scale networks (**Table 1**) and understand the biological processes and regulations on the system-level [15].

Table 1 Genome-scale metabolic networks.

Organism	Strain	Gene	Type	Reference
Bacillus subtilis	168	4114	Bacteria	[16]
Escherichia coli	K-12 MG1655	4405	Bacteria	[17]
Geobacter sulfurreducens		3530	Bacteria	[18]
Haemophilus influenza	Rd	1775	Bacteria	[19]
Helicobacter pylori	26695	1632	Bacteria	[20]
Lactococcus lactis	lactisIL1403	2310	Bacteria	[21]
Lactobacillus plantarum	WCF51	3009	Bacteria	[22]
Mannheimia succiniciproducens	MBEL55E	2384	Bacteria	[23]
Mycobacterium tuberculosis	H37Rv	4402	Bacteria	[24]
Mycoplasma genitalium	G-37	521	Bacteria	[25]
Neisseria meningitidis	serogroup B	2226	Bacteria	[26]
Pseudomonas aeruginosa	PA01	5640	Bacteria	[27]
Rhizobium etli	CFN42	3168	Bacteria	[28]
Staphylococcus aureus	N315	2588	Bacteria	[29]
Streptomyces coelicolor	A3(2)	7825	Bacteria	[30]
Methanosarcina barkeri	Fusaro	5072	Archaea	[31]
Halobacterium salinarum	R-1	2867	Archaea	[32]
Homo sapiens		28783	Eukaryotes	[33]
Mus musculus		28287	Eukaryotes	[34]
Saccharomyces cerevisiae	Sc288	6183	Eukaryotes	[35]

Biochemical processes that convert environmental stimuli into internal responses can be modeled as graph (or network) for computational convenience [36]. There are two basic components in each graph (or network): node and edge, where the node represents the element involved, such as metabolite, gene or protein while the edge represents interactions between these nodes. Two nodes are linked as long as they interact with each other, either physically or biochemically. According to the types of edges, each network can be distinguished as undirected network (**Figure 1A**) or directed network (**Figure 1B**). Sometimes, a certain ‘weight’ is assigned to the edge to indicate interaction strength (**Figure 1C**).

Though for most networks, nodes are of the same category (homogeneous node), the bipartite network is an exception, in which two kinds of nodes (heterogeneous node) are available and each node can only connect to nodes of the other kind.

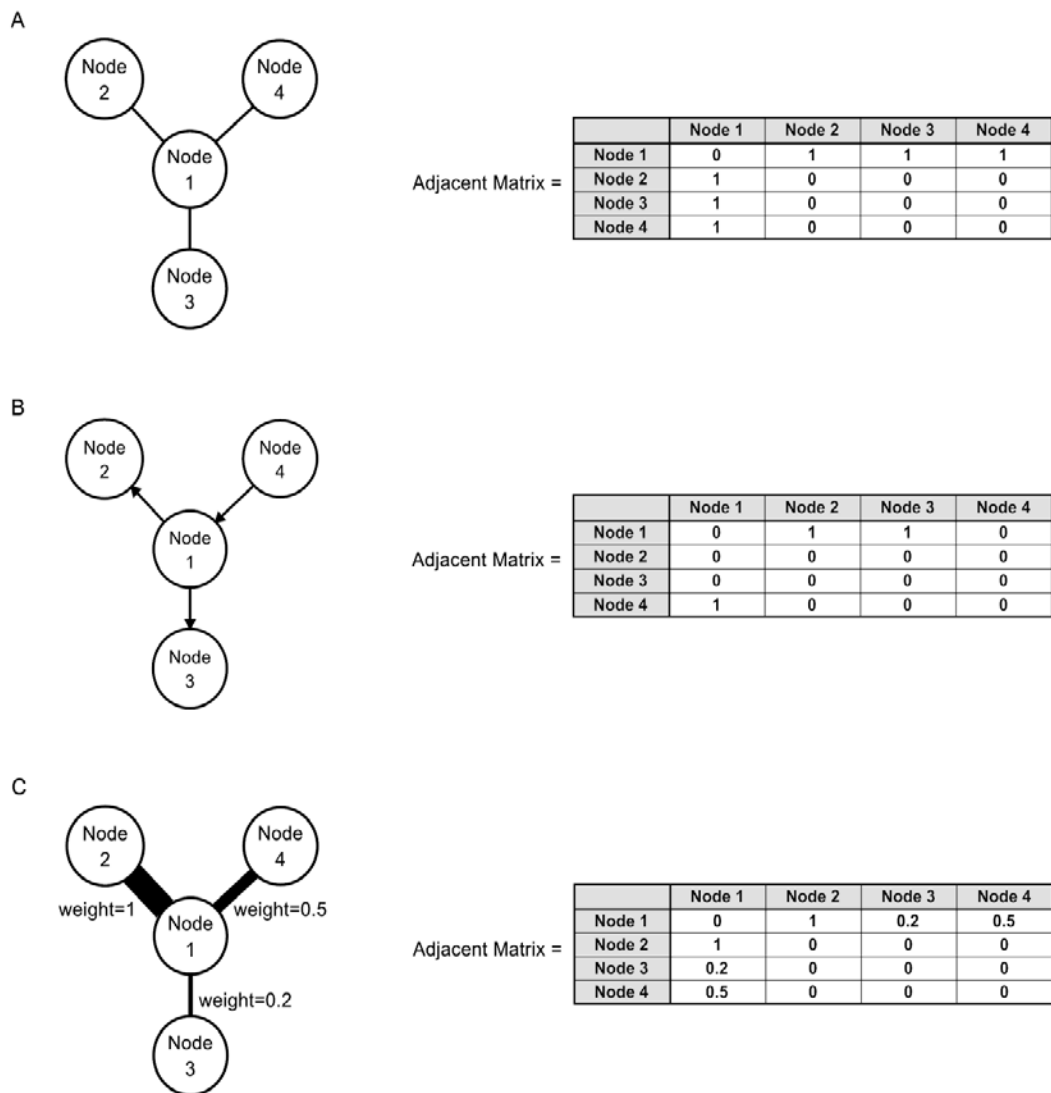


Figure 1 Adjacency matrix for (A) undirected network, (B) directed network and (C) weighted network.

Basically, these biological networks can be classified into three categories depending on the elements involved: protein-protein interaction (proteins), transcriptional networks (genes or transcriptional factors) and metabolic networks (metabolites or biochemical reactions). Protein-protein interaction network describes the physical interactions between proteins, within which each protein is represented as a node and each edge connects two interacting proteins. The transcriptional network describes the regulatory relationship between genes, where gene acts as node and edge denotes the regulation relations between gene pairs. Similarly, the metabolite network is consisted of a set of biochemical reactions / metabolites / enzymes.

While protein-protein interaction networks are usually illustrated as undirected graphs, the transcriptional and metabolic networks are frequently modeled as directed graph. This is due to the nature of interactions. For instance, in irreversible reactions, the flux can only flow from substrates to products, indicating that each edge should start at the substrates and finish at products if using a metabolite network. Similarly, in a transcriptional network, if gene A regulates gene B, then the edge should go from A to B, indicating the direction of the regulation.

For the same biological process, we can construct different types of networks depending on specific research focuses. For instance, for metabolic networks, there are reactions, metabolites, and genes involved. If we take the reactions as the

sole node, then two reactions are connected if the product of the first reaction is the substrate of the second one. Such a network can be tailored for studying the structural organization of biochemical reactions. If we are more concerned about metabolites, a metabolite network can be constructed in which node represents metabolite and edge represents that two metabolites can be coupled by means of any reaction. In these two cases, only one kind of node is available in the network, causing the loss of some information. To avoid losing any structural information, a bipartite network can be constructed in which there are two sorts of nodes, metabolites and reactions. Each metabolite node can only connect to reaction nodes, where an edge starting from metabolite to reaction node indicating that the metabolite is one of the substrates of the reaction whereas an edge starting at the reaction node and finishing at the metabolite node indicating that the metabolite is one of the product of the reaction. In this bipartite network, the activity of reaction node is controlled by the availability of the corresponding enzyme-coding genes. If the gene is knocked out or deleted, then the corresponding biochemical reaction should be inactive.

1.3 Mathematical representation of biological network (adjacency matrix)

Formally, a graph (or network) G , consists of a set of vertices or nodes, $v(G)$ together with an edge set, $\varepsilon(G)$. All the neighbors of node k are given by $N(k)$.

$$v(G) = \{v_1, v_2, \dots, v_n\} \quad (1)$$

$$\varepsilon(G) \subseteq v(G) \times v(G) \quad (2)$$

$$N(k) = \{t \in v(G) : kt \in \varepsilon(G)\} \quad (3)$$

The adjacency matrix is the mathematical representation of the topological structure of a network, which is ready for the mathematical analysis and computational simulation of biological networks (for details of adjacency matrix for various types of networks, please refer to **Figure 1**). For a network with N nodes, the adjacency matrix is an N by N matrix where each entry in the matrix indicates the availability of linkage from one node to another. For unweighted network, the entry value is usually set to 1 or 0, where 1 indicates that the two nodes are connected somehow whereas 0 indicates that there is no edge linking these two nodes. The adjacency matrix of an undirected graph is symmetric (**Figure 1A**) while that of a directed graph is generally not (**Figure 1B**).

In weighted networks, different strengths may be assigned to each edge which is represented as the numerical values in adjacency matrix (**Figure 1C**). Many real complex systems can be mathematically represented by weighted networks. For example, in a traffic network, the weight on each edge can be used to represent the traffic load in between two cities. The larger the weight, the more congestion the traffic considered to be. In metabolic networks, the amount of mass flow through bio-chemical reactions can be represented as some specific weights on each edge.

1.4 Characterizing the topological features of complex networks

These topological characteristics of complex networks, which can be described by some statistical quantities, are quite different from ‘trivial’ networks such as regular lattice and random graphs. In the following sections, we will discuss some fundamental measures of a complex network (**Figure 2**):

1.4.1 Degree and degree distribution

Degree is one of the most basic topological characteristics of complex network, which indicates how many nodes are connected to the target node. For a directed network, the degree can be further distinguished as in- and out- degrees. In-degree denotes how many edges points to this node, whereas out- degree denotes how many edges start from this node. Nodes with high degree are called ‘hubs’. Mathematically, it is represented as: $\text{deg}(k) = |N(k)|$.

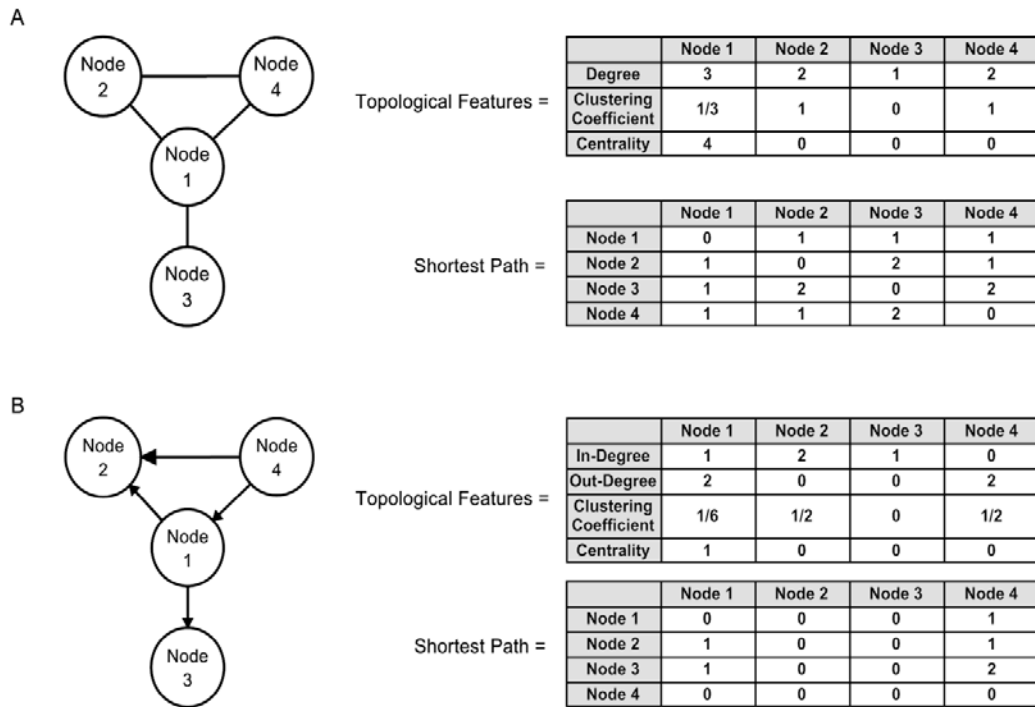


Figure 2 Topological features of complex network. For undirected network (A) and directed network (B), topological features such as in-degree, out-degree, clustering coefficient, centrality, and shortest path are illustrated.

The degree distribution $P(k)$ of a network is defined as the fraction of nodes within the network with degree k . For example, if in a network with N nodes, $N(k)$ of them have degree k , we define $P(k)$ as $N(k)/N$. The degree distribution is very important in understanding the structural organization of networks. It is found that the degree distribution follows a power law distribution in many real networks [37, 38], such as social networks, World-Wide Web networks [39], and biological networks [40], i.e. $P(k) \sim k^{-r}$ where k is the degree, the exponent r is typically in the range 2~3 (**Figure 3**). In such scale-free networks, the probability of having highly connected nodes (hubs) is much larger compared with random networks [41].

The structure of scale-free network warrants that it is robust to random attack. However, attack on certain high-degree nodes (hubs) can easily cause network failure [42]. In addition, such characteristic indicates that it is convenient for information transmission, and epidemic spreading in scale-free network. For example, the viruses can be easily spread across the whole network by targeting on the high-degree nodes [43]. Hence, in order to avoid epidemic spreading, we can isolate the infected nodes with large degree [44].

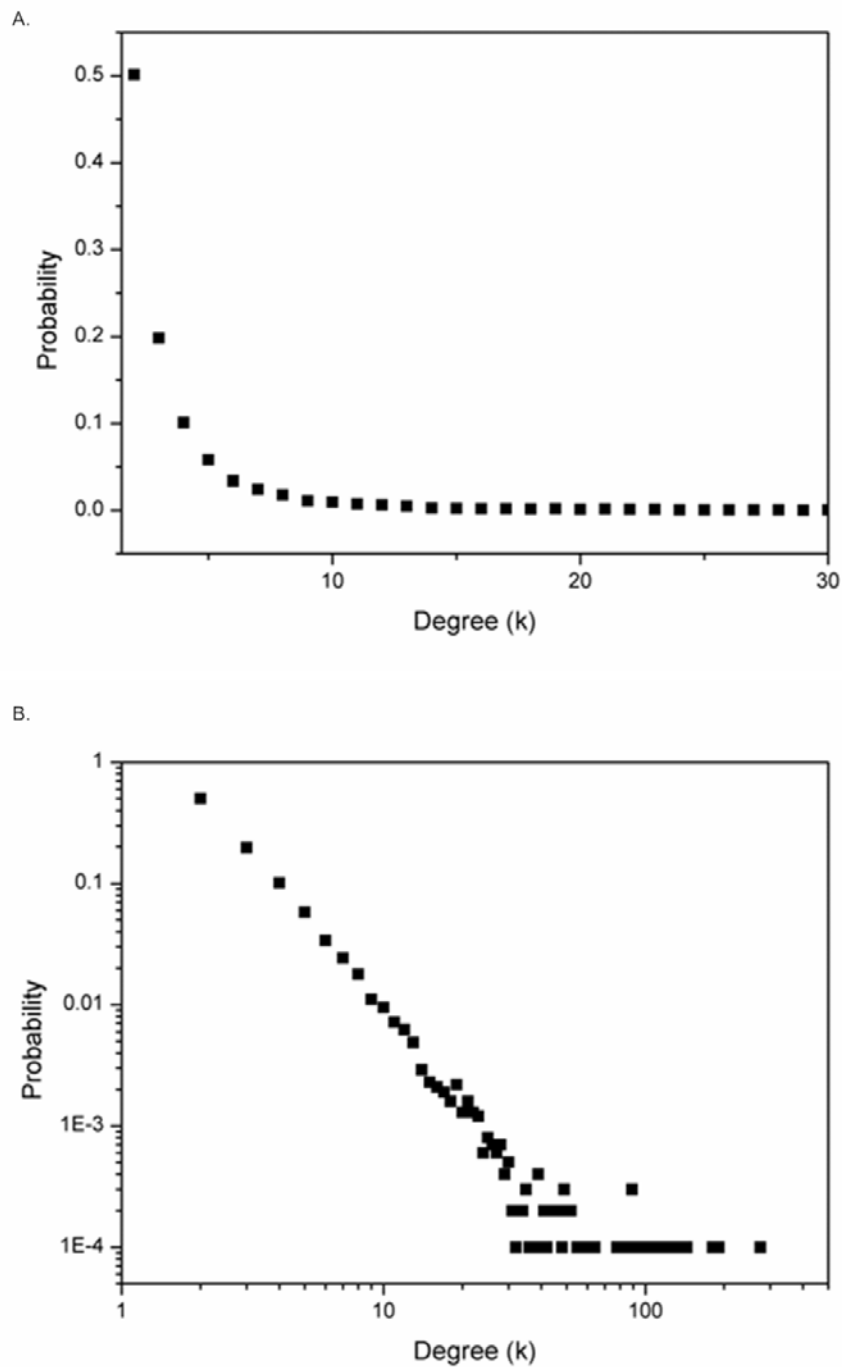


Figure 3 Degree distribution of scale-free network. (A) is the power law degree distribution in normal scale, while (B) is the same distribution in log-log scale. This figure is plotted from our scripts which are based on the Barabási–Albert (BA) model [41].

1.4.2 Path and shortest path

Let u, v be two vertices in a graph G . A sequence of vertices $\{u, v_1, v_2, \dots, v_i, \dots, v_k, v\}$ is said to be one of the path which starts at node u and ends at node v , of length k . Suppose there is more than one path from node u to node v , then the path covers the least nodes should be the shortest path. It is shown that the average path length (or distance) between any two vertices in complex network is much shorter than a random network [45].

1.4.3 Clustering coefficient

Clustering coefficient characterizes how likely the neighbors are connected to each other. The local coefficient is given by $C_u = \frac{2e}{k(k-1)}$ for undirected networks and $C_u = \frac{e}{k(k-1)}$ for directed network, where e represents the number of edges between the k neighbors of u in G . The average clustering coefficient for a network is the average of clustering coefficients all over the nodes. Study has showed that the average clustering coefficient of real network is usually much larger than random network [45]. Actually, the idea of clustering coefficient comes from the observation that if node A connects to node B , and node B connects to node C , it is likely that node A and node C also connect with each other. In other words, the local clustering coefficient can be interpreted as the number of triangles that can pass through a certain node.

1.4.4 Centrality

In large complex networks, it is not likely that all nodes are equivalent. Removal of a node may introduce distinct deletion effects depending on its relative topological location within the network. It is evident that removal of a cut-node (the analog of a bridge for edges) is more severe compared with removal of a dead-end node. As a consequence, centrality is introduced to describe the significance of a node within the complex network.

There are many ways to characterize centrality. Here we only discuss one of them - the betweenness centrality. For a given graph G , the betweenness centrality for vertex u is given by $g(u) = \sum_{s \neq u \neq t} \frac{\sigma_{st}(u)}{\sigma_{st}}$, where σ_{st} denotes the number of shortest paths from node s to node t and $\sigma_{st}(u)$ denotes the number of shortest paths from node s to node t that pass through node u . This formula needs to be normalized. The relative value of the betweenness centrality is a good indicator of the importance of the node. A larger value (close to 1) suggests that this particular node is inevitable to pass through for most of the shortest pathways.

Besides these mathematical measures, there are two other important terms we should know. First is 'giant component', the largest component in the network within which any two nodes can be connected regardless of the direction of the edges. In most cases, we are only concerned of the giant component, whereas

ignoring those isolated nodes. The second one is ‘strongly connected component’, which is similar to giant component except that it also considers the directions. In other words, node u and node v can reach each other mutually.

1.5 The application of network biology in drug development

The potential applications of network analysis include identifying drug targets [46-48], understanding cellular organization [49-51] and determining the regulation of biological processes [52]. In this section, we will elaborate its applications in drug development.

Network systems biology offers a novel way for drug discovery [53-55]. During the last century, drug discovery was driven by the assumption that ‘one gene, one target for one disease’ aiming to search for ‘disease-causing’ genes [56, 57]. Development of molecular biology and technologies in genomics [58] and high throughput screening facilitates the identification of genes responsible for diseases as candidates of drug targets [59, 60]. The target-based approach has successfully resulted in a variety of new drugs [61, 62]. Nevertheless, during the last few decades, there has been a significant decline in new drug candidates [63]. Researchers began to challenge the assumption of ‘one gene, one target for one disease’ [64, 65]. The growing field of network systems biology has also revealed the complexness underlying of molecular interactions. For example, genome-scale

knockout of genes in model organisms suggested that under laboratory conditions, many single-gene knockouts exhibit little impact on the phenotype. For example, in *Escherichia coli*, around 300 genes out of 4000 are vital for growth while all the others are dispensable [66]. Such robustness is considered to be caused by redundant genes or alternative pathways. In addition, research on drug-target networks [67] also indicates the common origins of many diseases and the distinct disease-specific functional modules. Furthermore, diseases like cancer are extraordinarily complicated with many genes and signaling pathways involved [63]. Therefore, it seems that the rational drug design based approach cannot guarantee success and may not be the best strategy. So in addition to searching for ‘disease-causing’ genes, topological and perturbation analysis of the disease-causing network or biological networks may offer valuable insights into diseases and approaches for drug discovery without losing any key information [53, 54].

One obstacle encountered by experimental-based drug discovery is the intensive resources required for large-scale drug target screening. Network based approach, a complementary to the traditional experimental based approaches, offers alternative way for identifying drug targets through topological analysis and predicting effects of attacks on the targets [53]. Topological analysis proposes a novel way of target identification, either single-gene or combination genes, by searching for those perturbations which can achieve a desired phenotype. In addition, understanding on genetic interactions in the context of metabolic network such as synthetic lethality [68], synthetic rescue [69] and genetic epistasis

helps us gain additional insight into drug target identification. Drug inefficiency and side effects are the two major reasons for the high failure rate in new drug development and such inefficiency may be reduced by integrating drug development with network biology [70]. For example, drug inefficiency may be circumvented by selecting drug targets whose functions cannot be easily replaced by other genes or pathways. Besides, network-based approaches can also be utilized to estimate the damage effects of normal cellular homeostasis from endogenous and exogenous perturbations [71]. The attacks are characterized by ‘knockout’ or ‘attenuation’ [72]. In a graph, ‘knockout’ removes target nodes and the relevant links from the network, while ‘attenuation’ weakens the strength of links. Algorithms such as flux balance analysis [73], and topological flux analysis [74] have been designed to characterize the damage effects from perturbations. In summary, network biology revolutionizes our view of disease and offers alternative ways for drug discovery.

1.6 Identification and analysis of essential genes

Large-scale use of antimicrobial agents such as penicillin, and streptomycin has saved millions of people from infectious diseases. However, in the last few years, the frequency of drug resistance was increasing [75, 76], indicating the necessity to explore more efficient drug targets.

Typically, drug targets are those elements involved in key biological processes, such as cell wall biosynthesis, cell membrane, protein biosynthesis, DNA replication and so on. Perturbation of these elements can lead to bacterial death or growth inhibition [77]. Hence, essential genes are considered as ideal candidates.

By definition, essential genes are those genes necessary for cellular growth in rich medium [78]. The disruption of essential genes can confer lethality phenotypes even in the presence of all the other genes. It indicates that these essential genes may participate in certain key processes, where the disruption prevents normal cell functions. Therefore, studies on essential genes not only provide a list of drug targets, but also deepen our understanding of cellular functions and organizations.

1.6.1 Molecular basis of gene mutation

All organisms can undergo genetic mutations which refer to the alteration in nucleotide sequence of genetic material (such as DNA, RNA, and extra-chromosomal genetic elements). Genetic mutations usually result from errors in the process of DNA replication, insertion or deletion of DNA segment to host genome by mobile genetic elements, or induced by exposing to chemicals mutagens and UV radiation [79-81].

Genetic mutations can occur on a small scale, such as point mutation where a specific nucleotide changes from Adenine (A) to Guanine (G) or from Cytosine (C) to Thymine (T) or vice versa. Because of genetic code redundancy, the effect of point mutation can be either silent mutation (i.e. a point mutation does not result in a change to the amino acid sequence), missense mutation (i.e. a point mutation results in a codon that codes for a different amino acid), or non-sense mutation (i.e. a point mutation results in a premature stop codon). Other types of small scale genetic mutations include insertion and deletion mutation which add or remove one or more extra nucleotides into DNA, respectively. In addition, it is not uncommon to observe large scale chromosomal mutations. For example, duplication of larger range of sequences or chromosomal segments may occur during the process of genetic recombination [82].

All the above stated molecular events can induce very distinct outcomes such as loss of function of genes (loss-of-function mutations), gaining new or abnormal functions (gain-of-function mutations), and so on. Some of those function changes are deleterious to the fitness of organisms which reduce fertility and even cause lethality. Other functional alterations can be either 'neutral' which have no biological impact on survival fitness or 'advantageous' which increase the fitness and the success of fertilization. Genetic mutations play a key role in species evolution. Under the pressure of natural selection, individual organisms with advantageous mutations can pass the good characteristics to offspring at a higher

rate of possibility, therefore, endows the offspring a better capability of adaptation to the changing external conditions [83, 84].

The genome of an organism normally carries a large number of mutations, whereas some deleterious mutations are tightly associated with diseases. Sometimes, even a very small amount of base pairs changing in the DNA sequence can result in serious physiological malfunctions. For example, sickle-cell anemia is caused by alteration of a single nucleotide in the gene coding beta chain of haemoglobin protein [85]. Other than that, many complex human diseases, such as cancer, are proved to be related to genetic mutations. One example is that women with germline mutations in gene BRCA1 and BRCA2 have a higher risk of developing breast and ovarian cancer [86].

Considering the important roles of genetic mutations in both normal and abnormal biological processes, it is valuable to identify those mutations and investigate their relation with physiological alterations. For example, identification of germline and somatic mutation can help to develop personalized therapeutics for cancer patients. Several genome-scale mutation projects on model animals (e.g. zebrafish) have been carried out [87]. However, distinguishing deleterious mutations from the large number of non-functional variants that occur within the whole genome is still a considerable challenge, and further endeavors are needed.

1.6.2 Experimental studies on essential genes

Genome-scale knockout experiments have been used to identify essential genes for different model organisms, including *E. coli* and *S. cerevisiae* [66, 88, 89]. In 2006, Baba *et al.* [66] systematically made a series of single-gene in-frame deletion for 4288 genes in *E. coli* K-12 by replacing the open-reading frame coding region with a kanamycin-cassette-flanked fragment and mutants for 3985 were obtained. The remaining 303 genes that cannot be disrupted are candidates for essential genes. This is one of the most reliable systematically generated essential gene lists, widely used in essential gene studies [13, 90]. Earlier, a genetic foot printing approach was used to assess gene essentiality for *E. coli* in the aerobic, rich media [91]. 620 genes were identified as essential genes while 3126 genes as non-essential under the given media. A systematic study on yeast [88] showed that around 1105 genes (18.7%) are essential for growth on rich glucose media. Techniques such as mutagenesis, transposon, or allelic replacement have also been used to disrupt target gene functions by replacing the corresponding open-reading frame with some non-functional fragment just like the Baba *et al.* experiments [66]. If the mutant strains can form clones or maintain a growth rate comparable with the wild type, the corresponding gene is considered as an essential gene, otherwise it is termed “non-essential”. Occasionally, researchers may be more concerned about identifying the conditionally essential genes which are the genes crucial in certain growth conditions [92], such as different carbon sources, aerobic, or anaerobic.

However, there are some limitations with these experimental techniques. First of all, the experiment is quite time-consuming, and resource intensive. Besides, it is not feasible to conduct experiments for all microorganisms, especially infectious ones. Moreover, the experimental results for the same species may vary a lot. For example, two experiments targeting on yeast generated essential gene lists of different sizes, one with around 600 genes [88], and the other with up to 900 genes [93]. Profiling of *E. coli* Chromosome (PEC) is a database for *E. coli* [94], in which all the essentiality information is obtained from the literatures, including 302 essential genes, 3136 non-essential genes and 5 unknown. Notably, a comparison with the Keio collection [66] indicates that they share 264 common essential genes while 50 genes specific to each database. This indicates that there is a demand to reconcile the experimental results.

1.6.3 Identification of essential genes via computational approaches

The development of high throughput technologies has made it possible to generate large-scale genome sequencing data, constructing large-scale computational models for a lot of organisms. Network based approaches either apply flux balance analysis (FBA), or network topological features for essentiality identifications. FBA [73] is a constraint-based approach which predicts the fluxes through each reaction with the knowledge of stoichiometric matrix, lower and upper constraints for each flux, and the biomass composition (for details, please refer to **Section 2.4, Chapter 2**). It has been widely used to predict the growth rate and viability of mutant strains, where a significant change in the growth rate

between mutant strain and wild type indicates the indispensable of target gene [95]. Later on, the assumption underlying flux balance analysis is challenged because researchers suggest that it does not make sense to assume that the mutant strains are also evolved optimally like wild type. As a result, a modified constraint based approach (MOMA) is proposed [96]. Though constraint-based approaches may have good prediction accuracy, clear and accurate nutrient availability information, biomass components, and a completeness network definition is required.

Besides FBA approach, topological based methods are also widely used in predicting essential genes. For example, network topology alone is considered enough for predicting viability of mutant strains by employing a proposed network measure ‘synthetic accessibility’ in *E. coli* and *S. cerevisiae* [97]. Another category is machine learning based approaches. The principle of this kind of approach is to identify some properties or features that may associate with essentiality, and then use them to train the network model and finally provide some predictions. Features used may include network topology (such as degree, closeness, betweenness centrality, or clustering coefficient, in-degree, and out-degree and so on) and genome sequence characteristics (such as codon adaption, GC content, localization signals) [98]. In some studies, more than 25 features are used to predict essentiality. However, one problem with this kind of approach is that there is no clear evidence indicating any causality between these features and essentiality. Although they may reveal some differences between essential and

non-essential genes in the structural organization, they are not the determinant factors to discriminate these two types gene. As a result, the prediction accuracy may not be that optimal. Instead of using these features, some causality features may achieve even better results. Therefore, it is necessary to find out the links between genotypes and phenotypes, and by what means. Studying on these internal interactions may deepen our insights in essential genes and shed light on the development of efficient essentiality prediction approaches.

1.6.4 Single gene deletion analysis

The advent of high throughput sequencing technologies makes it possible to reconstruct large-scale biological networks [1, 2, 5, 59]. Genome-scale single gene deletion analysis showed that most genes are dispensable, though a few of them are necessary for cellular growth [66, 88]. Such robustness is universal across species [99].

The concept of gene essentiality is also extended to metabolite/reaction/enzyme. A quantitative analysis of enzyme importance revealed that while a large fraction of enzymes can cause little damage when removed, there are a small fraction can cause serious damage [100]. Flux analysis on metabolic network showed similar findings. The global organization of metabolic fluxes [101] indicated that while most metabolic reactions have low fluxes, there are a few of them with quite high fluxes. Analysis on the flux distribution on central metabolism [102] suggested

that most reactions can be reduced greatly, for example, reduced to 19%, without having significant impact on the optimal growth rate. However, the fluxes via three-carbon glycolytic demonstrated limited robustness.

Previous study showed that under laboratory conditions, 80% of yeast genes seem not to be essential for viability [88]. It raises the question of the cause and evolution of mechanistic basis for essentiality and dispensability. Three explanations are proposed: (1) Conditional essentiality, i.e. some experimentally identified non-essential genes are indeed essential in conditions that not examined [103]. Several experimental studies [92, 104, 105] have been implemented to identify conditional essential genes. (2) Genetic compensation by duplicated genes or genes with overlap functions [103, 106]. (3) The presence of alternative pathways or redistribution of flux distribution [103, 107]. ‘k robustness’ analysis for multiple genetic knockout indicates the available of functional backup [108]. One study showed that the metabolite essentiality elucidates the metabolic network robustness [109]. They illustrated that essential metabolites have the ability to maintain a steady flux-sum even in response to severer perturbations by redistributing the fluxes. Another study [110] revealed that essential reactions are usually associated with low-degree metabolites. These reactions are essential because they are the only source available to consume/produce the respective metabolites.

In the context of metabolic network, it is intriguing to know how genetic perturbation leads to lethality. The mechanisms which link gene deletion to the final phenotype are seldom studied. As the structural organization of metabolic network determines key functions and regulations [111], we would like to explore on the structural and functional organizations of essential genes and how it lead to lethality.

1.6.5 Evolution of essential and non-essential genes

Studies showed that essential genes are more conserved compared with non-essential genes in bacteria [112]. As essential genes are highly participated in some key biological processes that are indispensable for an organism, it is expected that they are highly conserved across species [78]. This is further confirmed by some comparative genomic analysis. For example, sequencing of 36 clinical *Pseudomonas aeruginosa* isolates [113] revealed that essential genes evolve at a lower rate and moreover, no sequence variation is observed for 980 essential genes. Homologous analysis of *Staphylococcus aureus* and *Mycoplasma genitalium* genome [114] revealed 168 conserved genes. Interestingly, most of these genes are found to be essential in *M. genitalium* and other bacteria.

The conservation of essential genes in bacteria is widely accepted as the premise for identifying novel essential genes across species. For example, by deleting open reading frames (ORFs) from *E. coli* genome, the encoded gene products of which

are homologous to proteins from bacterial pathogens, researchers identified 4 novel essential genes, implying a new strategy for broad-spectrum antibiotics drug development [115]. Another down-selection approach is applied to DEG, a database of essential genes which includes a large number of putative essential genes [93], and identified 52 essential genes conserved across 7 (or more) out of 14 genomes [116]. Further experiments confirmed that 7 out of 8 mutants are shown to be essential for survival for a non-related species.

It is easy to conclude that the orthologous of essential genes in one species is essential in other species. However, this is not the case. Sometimes, orthologous of essential genes are missing or become non-essential genes in other species [91, 117, 118], implying that function of essential genes may be replaced by other non-essential genes.

In contrast to bacteria, no significant difference between the evolution rates of essential and non-essential genes is observed by analyzing the evolutionary distance between *S. cerevisiae* and *C. elegans* proteins [119]. An analysis of mouse and rat orthologous genes found that essential and non-essential genes evolved at similar rates if the biased assumption (i.e. genes thought to evolve under directional selection) was excluded from the analysis [120].

1.7 Genetic interactions

As indicated by the genetic buffering mechanisms [106], understanding on the genetic interactions (or epistasis) may help unveil a higher-order organization of complex biological networks and deepen our insights on the mechanisms regarding the cause and evolution of dispensability and essentiality.

By definition, epistasis is the ability of one mutant to ‘mask’ the phenotypes caused by other mutants, an indicator of the interactions between genetic mutants [121, 122]. Two types of interactions are especially of importance, namely, synthetic lethal and synthetic rescue. Synthetic lethal (or rescue) are gene pairs whose double deletion can lead to lethality (or survival) whereas their individual gene deletion cannot [123]. Their significance in anti-cancer drug development [68, 69, 124] motivated studies which targeting to identify/predict synthetic lethal/rescue pairs [90, 125].

Experimentally, synthetic genetic arrays (SGA) [126] and synthetic lethality analysis by microarray (SLAM) [127] are two approaches widely used for screening synthetic genetic interactions. A large-scale SGA screen for yeast identified a genetic network with ~1000 genes and ~4000 interactions [128]. Further network analysis revealed the small world properties, e.g. the length of the shortest path between pairs of genes is significantly shortened compared with random network [128].

As the experimental approaches are quite time-consuming and resource intensively, synthetic interactions are also extensively explored by computational approaches. In the framework of metabolic network, fitness may be computed to characterize the epistatic effects [129]. The epistasis analysis revealed that genetic interactions can be further classified as: buffering, aggravating, and non-interaction [50]. Positive epistasis where double deletion can enhance the fitness is found prevalent in both *E. coli* and *S. cerevisiae* [130]. The epistatic interactions are hierarchically organized, forming functional modules within which purely aggravating or buffering links are observed [50]. Synthetic lethal interactions are demonstrated capable of predicting genetic compensatory pathways [131]. While the genetic interactions identified in yeast are not highly conserved across animals [132], the synthetic lethal genetic interaction networks between distantly related eukaryotes are significantly conserved [133].

Although both experimental and computational analyses have revealed some properties associated genetic interactions, a mechanistic understanding on how the cellular phenotypes arise in response to double/multiple gene deletion is limited studied. A comprehensive understanding on the genetic interactions may help unveil the organization of biological networks and offer explanations for the observed genetic buffering or aggravating.

1.8 Objectives and outline of this thesis

1.8.1 Objectives of this thesis

One of our goals in this thesis is to reveal the nature of essential and non-essential genes by addressing problems like what causes the essentiality and dispensability from the scope of metabolic network and how it evolves across species. The mechanism studies regarding this problem are controversial and there are mainly three explanations: (1) Conditional essentiality, i.e. those non-essential genes are actually essential in the conditions not examined. (2) Overlapped gene function, which means that those non-essential genes are functionally overlapped by other genes. (3) Pathway compensation, in other words, alternative pathway is available to compensate the lost function. The study of conditional essential genes is not the focus of this thesis. Instead, we will study from the perspective of functional complementary to reveal how essential gene deletion can trigger the lethality phenotypes whereas non-essential gene deletion cannot. We will study this problem by characterizing damage caused by single gene deletions and addressing questions like what the functional and structural differences made the distinct phenotypes.

Another goal is to computationally predict essential genes from the reconstructed metabolic networks. Once we can gain new insights into how essential and non-essential genes can exert distinct deletion effects, some features tightly associated

with essential genes can be extracted, which are the basis for essential gene prediction. The more we gain from the mechanism studies, the more efficient prediction approaches we can propose.

Genetic interactions between pairs of genes is another goal of our study, which can help to reveal a higher order organization of biological networks and offer explanations for the observed ‘epistasis’. In this thesis, we study not only the characteristics of gene pairs with epistasis effects, but also how it evolves across species from the perspective of evolution.

1.8.2 Outline of this thesis

In **Chapter 1**, we first introduced network systems biology and reviewed some characteristics of network biology and its applications in drug discovery and development. Then, we illustrated the importance of essential gene analysis and its current progresses, including the development of essential gene prediction approaches, the mechanisms explanations for gene dispensability, evolution studies on essential genes, and genetic interactions between pairs of genes.

In **Chapter 2**, we summarized some approaches and databases used in this thesis. A brief illustration of how we can construct metabolic networks from some

available databases was also given. The algorithm we proposed to capture gene deletion effects was illustrated in detailed steps.

In **Chapter 3**, we systematically analyzed the differential deletion effects between essential and non-essential genes in the context of metabolic network. Initially, a new measure for single gene deletion effect was introduced. Then a mechanism was proposed to explain the distinct deletion effects between essential and non-essential genes based on some comparisons between these two kinds of gene deletion effects.

In **Chapter 4**, a newly developed method targeting for essential gene prediction was proposed. After a brief introduction of the algorithm, we then compared the performance of this approach with other available methods. Some suggestions on how to improve the prediction accuracy was also indicated.

In **Chapter 5**, we applied the double gene deletion analysis on the genome-scale *E. coli* and yeast metabolic networks to reveal the genetic interactions. Gene pairs with reduced deletion effects and enhanced deletion effects were further analyzed in details. Characteristics associated with gene pairs were further studied across species.

Chapter 6 summarized all the major findings and contributions of this work.

Limitations and some potential future works were also discussed in this chapter.

Chapter 2 Materials and methods

2.1 Biological databases

2.1.1 Kyoto Encyclopedia of Genes and Genomics (KEGG)

Kyoto Encyclopedia of Genes and Genomics (KEGG) [134] is a database resource available for metabolic network reconstruction and pathway analysis. It is hierarchically organized within which there are three sub-databases KEGG LIGAND, KEGG GENE and KEGG PATHWAY [134]. KEGG LIGAND is the collection of chemical compounds and reactions relating to the cellular processes, whereas KEGG GENE is the collection of genomic information such as gene annotation. KEGG PATHWAY represents a higher order of functional related information, such as metabolism, and signal transductions, among which the metabolism are the best organized part. Metabolism is classified into 13 categories (**Table 2**), which are further grouped into around 160 reference metabolic pathways.

Although these three sub-databases (i.e. KEGG LIGAND, KEGG GENES, and KEGG PATHWAYS) provide necessary knowledge for creating the draft of genome-scale metabolic network, lacking of the information of reaction directions for around 3000 biochemical reactions makes it insufficient for reconstructing a detailed directed metabolic network. Though previous studies [135] proposed some rules for determining the reaction directions, metabolic network

reconstructed from KEGG database is not the ideal candidate for our quantitative studies. However, the enzyme, gene, and pathway information are good resources for further pathway analysis and cross-species study.

Table 2 13 categories of metabolism classification in KEGG database

Label	Pathway
1.1	Carbohydrate metabolism
1.2	Energy metabolism
1.3	Lipid metabolism
1.4	Nucleotide metabolism
1.5	Amino acid metabolism
1.6	Metabolism of other amino acids
1.7	Glycan biosynthesis and metabolism
1.8	Metabolism of cofactors and vitamins
1.9	Metabolism of terpenoids and polyketides
1.10	Biosynthesis of other secondary metabolites
1.11	Xenobiotics biodegradation and metabolism
1.12	Reaction module maps
1.13	Chemical structure transformation maps

2.1.2 BiGG database

Biochemical Genetic and Genomic knowledgebase (BiGG) of large scale metabolic reconstructions is used for metabolic network reconstruction and further quantitative analysis in this thesis since the information contained in this database is of high confidence [136]. Different from KEGG database, BiGG is integrated from many available sources following by model refinement by approaches such as Flux Balance Analysis (FBA). Metabolic models available in BiGG is reconstructed by summarizing all the information from KEGG and other database,

followed by literature reviews and quantitative approaches such as FBA and gap analysis to fill in the ‘gaps’ by adding missing reactions into the network [137]. This procedure is iterated until the model is optimized and tested. Besides, each reaction is given a confidence level to indicate the reliability. As a consequence, network from BiGG is more reliable and suitable for modeling analysis.

2.2 Metabolic network reconstruction

2.2.1 Data source for network reconstruction

The metabolic networks of *E. coli* and *S. cerevisiae* are reconstructed from models *E. coli* iAF1260 [17], and *S. cerevisiae* iND750 [138] (for details, please refer to **Section 2.2.3, Chapter 2**), exported from BiGG database (<http://bigg.ucsd.edu/-bigg/main.pl>). The exported file includes information such as reaction abbreviation, subsystem, equation, gene association, confidence level and so on. In addition, BiGG provides the information of reaction directions by utilizing the symbols ‘-->’ and ‘<==>’ under the entry ‘Equation’. For the listed equation, ‘-->’ indicates that this reaction flows from the left side to the right side. If this is replaced by ‘<==>’, it suggests that the reaction is reversible (**Figure 4**).

2.2.2 Gene-protein-reaction (GPR) association

Biochemical reactions can be linked to enzyme-coding genes via gene-protein-reaction (GPR) association information. Basically, there are two kinds of relations, isozymes and protein complex, although others may be the mixture [137]. Isozymes refer to those gene products that function independently, namely, each gene product is sufficient for catalyzing the reactions. As for protein complex, multiple gene products form a complex in order to catalyze reactions. In this case, all of them are necessary for proper functioning. In the exported SBML (System Biology Markup Language) file, the isozymes and protein complex are indicated by the logic word 'AND' and 'OR' respectively. For the first example (**Figure 4A**), these genes are in the logic 'OR' relation, suggesting that either of them is enough to catalyze the reaction, which is the case of isozymes. For the second example (**Figure 4B**), these genes are in the logic 'AND' relation, indicating that all of them are required for the proper catalyzing, which is the case of protein complex. Sometimes the GPR association is a bit complex, involving both cases (**Figure 4C**). For this complicated case, proteins encoded by genes breakdown by logic 'OR' is enough for catalyzing the reaction (in this example, either gene pair b0077 and b0078, or gene pair b3670 and b3671 is enough for the reaction).

- A. `<html:p> Abbreviation: R_CYTDtex </html:p>`
`<html:p> Synonyms: _0 </html:p>`
`<html:p> SUBSYSTEM: Transport Outer Membrane Porin </html:p>`
`<html:p> Equation: cytd[e]<==>cytd[p] </html:p>`
`<html:p> ConfidenceLevel: </html:p>`
`<html:p> GENE ASSOCIATION: (b2215)or(b0241)or(b1377)or(b0929)`
`</html:p>`
- B. `<html:p> Abbreviation: R_AGt3 </html:p>`
`<html:p> Synonyms: _0 </html:p>`
`<html:p> SUBSYSTEM: Inorganic Ion Transport and Metabolism </html:p>`
`<html:p> Equation: ag[c]+h[e]-->ag[e]+h[c] </html:p>`
`<html:p> ConfidenceLevel: 3 </html:p>`
`<html:p> GENE ASSOCIATION: (b0574andb0572andb0575andb0573)`
`</html:p>`
- C. `<html:p> Abbreviation: R_ACHBS </html:p>`
`<html:p> Synonyms: _0 </html:p>`
`<html:p> Equation: [c]:2obut+h+pyr-->2ahbut+co2 </html:p>`
`<html:p> ConfidenceLevel: 0 </html:p>`
`<html:p> GENE ASSOCIATION: (b3670andb3671)or(b0077andb0078)`
`</html:p>`

Figure 4 Samples of exported SBML file from BiGG database. (A) Isozymes; (B) Protein complex; (C) Combination of both isozymes and protein complex. In the exported SBML file, each reaction is characterized by the abbreviation, equation, confidence level, gene association, or enzymes. The equation can either be irreversible, or reversible, where ‘>’ indicates its irreversible, and ‘<==>’ indicates its reversible.

2.2.3 Metabolic network representation and visualization

Network reconstruction is the process of converting and resembling biochemical reactions into computational data structure. In this thesis, we converted the model

into a directed, bipartite network, which consisting of two types of nodes, metabolites and reactions.

For the example reaction, $R: A + B \rightarrow C + D$, two types of nodes (or vertices) are defined in our graph structure, i.e. reaction node (e.g. R) and metabolite nodes (e.g. A, B, C, D). A directed edge connects a metabolic node to a reaction node if the metabolite participates in the reaction as reactant (directed towards the reaction node) or product (directed towards the metabolite node). In this example, four edges are added to the graph, i.e. $A \rightarrow R$, $B \rightarrow R$, $R \rightarrow C$ and $R \rightarrow D$. For reversible reaction, we treated it as two separately reactions. In this example, eight edges are added to this graph.

Currency metabolites, which typically involved in a variety of reactions, are generally excluded from metabolic network analysis as their presence may significantly decrease the path distance and add some meaningless links into the network [135, 139]. For example, in glycolysis metabolism, a sequence of biochemical reactions can convert glucose to pyruvate (**Figure 5**). However, as ATP is involved in both the first and last reaction, these two reactions will be linked if currency metabolites are included in the network reconstruction.

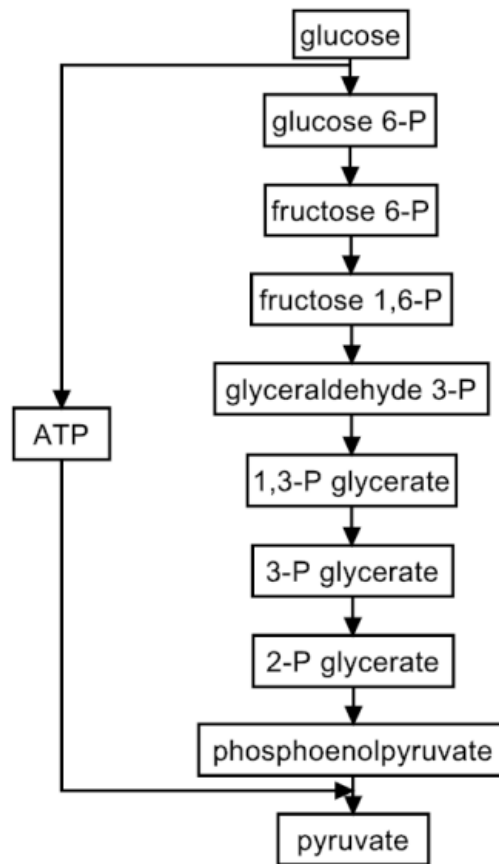


Figure 5 The glycolysis pathway which converts glucose to pyruvate [135]. Currency metabolite ATP introduces meaningless links that ‘short cut’ the distance from glucose to pyruvate.

As a consequence, 26 currency metabolites such as ATP, ADP, and NADPH are removed from our network (**Table 3**) [140]. Exchange reactions involving metabolites transporting between different cellular locations are also discarded as this is not our focus. The obtained *E. coli* iAF1260 network includes 2351 nodes and 4038 edges, while *S. cerevisiae* iND750 includes 1441 nodes and 2596 edges. A reconstructed bipartite metabolite from the cofactor and prosthetic group biosynthesis pathway of *E. coli* iAF1260 is demonstrated in **Figure 6**.

Table 3 26 currency metabolites.

26 currency metabolites			
ATP	UMP	CMP	propanoyl-CoA
ADP	GMP	CTP	L-glutamine
UTP	NAD	CDP	L-glutamate
UDP	NADH	H ₂ O	phosphate
GTP	NADP	CO ₂	2-oxogutarate
GDP	NADPH	NH ₂	
AMP	CoA	acetyl-CoA	

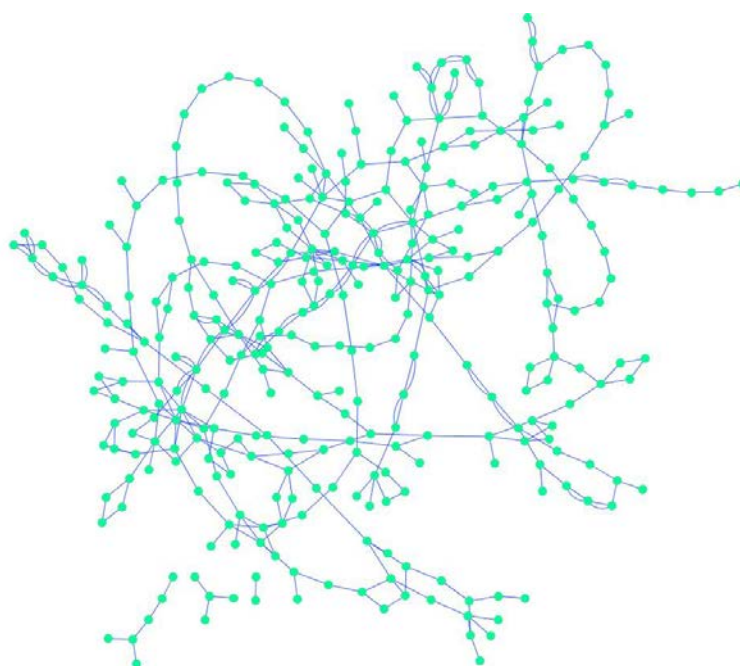


Figure 6 Reconstructed metabolic network. The cofactor and prosthetic group biosynthesis pathway of *E. coli* iAF1260 is reconstructed and visualized by open software Cytoscape [141], a powerful tool to customize the visualizations. In this figure, it is noticed that most nodes form a giant component whereas some reactions are isolated. Reversible reactions can also be detected.

2.3 Characterize gene deletion effects

In this thesis, we devised an algorithm which can characterize single and multiple gene deletion effects by a list of affected genes based on a previous topological flux balance analysis method [74].

2.3.1 Corresponding reactions

Before going further into our method, we need to clarify some terms used. Considering the GPR association (for details, please refer to **Section 2.2.2**), each gene may participate in several reactions. However, it is not likely that deletion of the gene can affect all of these reactions due to functional compensation. Hence, for each gene in the network, we can identify those reactions that cannot occur anymore due to the lack of a particular gene, which are termed as ‘corresponding reactions’.

As mentioned in previous section, there are basically two kinds of logic relations ‘AND’ and ‘OR’ in the GPRs associations (**Figure 7**). The logic ‘AND’ indicates that both genes are necessary for properly encoding the corresponding enzymes, which acting as the catalyst of reactions. The logic ‘OR’ suggests that either gene is enough for the reaction, implying the functional overlap between genes. Another complex one is the combination of two.

The ‘corresponding reactions’ of each reaction is determined case by case. For the first case (**Figure 7A**) where genes are in the logic ‘AND’ relations, all of these genes are necessary for the proper function of the reaction. Lacking any of them, the reaction cannot occur anymore since the lost function cannot be compensated. So in this scenario, the reaction is the ‘corresponding reaction’ for all of the associated genes.

For the second case (**Figure 7B**) where genes are in the logic ‘OR’ relations, any of the genes is capable of encoding the corresponding enzyme. Therefore, removing any of the single genes associated will not affect the reaction because of the existence other genes of similar functions. In this scenario, the reaction is not the ‘corresponding reaction’ for all of the associated genes.

Another case (**Figure 7C**) is a bit complex as it may involve both the logic ‘OR’ and ‘AND’. Since the logic ‘OR’ has high priority compared with the ‘AND’, we need to check the sub-gene groups. If one gene appears in all of the sub-gene groups, it indicates that the removal can affect all of sub-gene groups. As a result, the enzyme cannot be encoded and therefore no reaction occurs. So in this case, the reaction counts one ‘corresponding reaction’ for this gene. Hence, to judge whether one reaction belongs to the ‘corresponding reaction’ of a gene, we need to identify whether the removal of gene can affect the activity of the reaction.

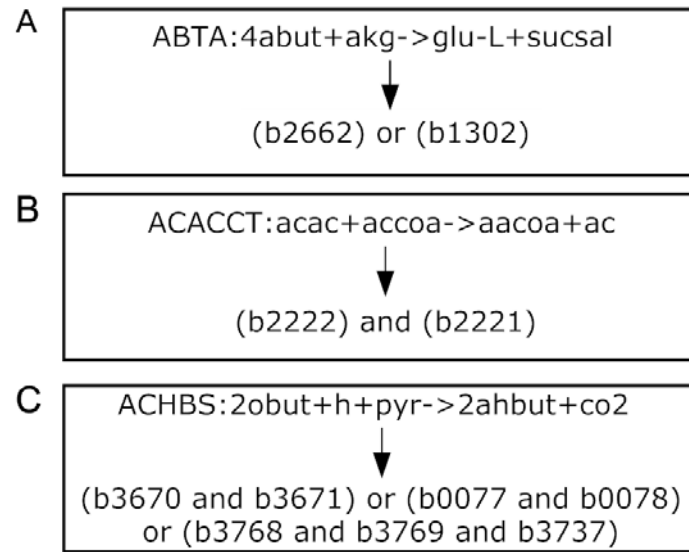


Figure 7 Gene-protein-reaction Association. (A) ABTA is a reaction, which is catalyzed by the encoded product of gene b2662 or b1302. Either gene is enough for the reaction. (B) Gene b2222 and b2221 work cooperatively to catalyze the reaction ACACCT. Both genes are required for the reaction. (C) For reaction ACHBS, sub-gene group (b3670 and b3671), or (b0077 and b0078), or (b3768 and b3769 and b3737) is enough to catalyze the reaction. All these reactions are obtained from BiGG database [136].

2.3.2 Cascading failure procedures

Our gene deletion algorithm is based on the assumption that at the steady state, each internal metabolite should have a non-zero in-degree (generated by upstream reactions) and out-degree (consumed by downstream reactions). To initiate the algorithm, we first identified the ‘corresponding reactions’ for each query gene according to the GPR association information. Only when all the enzyme subsets are non-functional, we assumed that a particular reaction cannot occur. Next, we started the cascading failure procedures by removing all the ‘corresponding

reaction' nodes for the query gene and their links from bipartite network simultaneously.

In the following step, we searched the bipartite metabolic network upwards and downwards, finding all metabolites with zero in-degree or out-degree. These metabolites and their links are removed in the next step, followed by searching for reaction nodes with incomplete substrates or products. The procedures are iterated until all the leftover metabolite nodes with non-zero in-degree and out-degree whereas reaction node with complete substrates and products. A list of affected reactions can be obtained in this process. The detailed network failure procedures were illustrated in **Figure 8**.

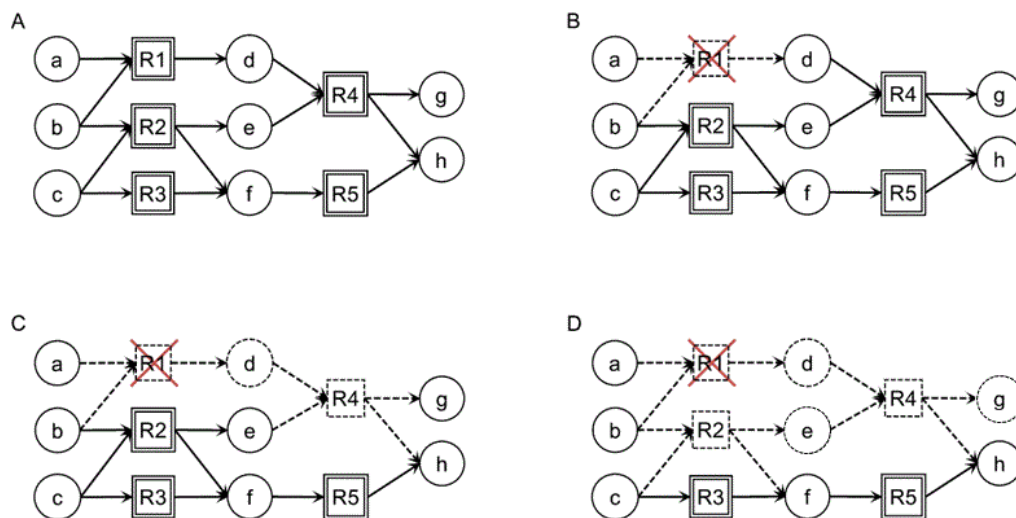


Figure 8 Schematic diagram for network cascading failure for single gene deletion. (A) original metabolic network (B) delete the ‘corresponding reaction’ R1 (cross, box with dashed border) for the query gene and its links (dash lines) (C) remove metabolites with zero in-degree or out-degree (metabolites: d, circle with dashed border) and reaction R4 (box with dashed border) which with incomplete substrates and products and the corresponding links (dash line) (D) iterating the deletion procedures upward and downwards in the network until all the sabotaged metabolite and reaction nodes are removed. Reactions R1, R2, and R4 will be removed in response to query gene deletion.

Finally, we mapped our obtained reaction lists to gene lists. For those affected reactions, as long as they belong to another gene’s corresponding reactions, we assumed that this particular gene will be affected in response to single gene knockout. The union of all the affected genes formed our damage lists. For the case that one gene is associated with reaction in the damaged reaction list, and also associated with reactions not in the list, we still put it in the damage list, because its partial function is affected. The overall procedures from query gene to damage list determination were summarized in **Figure 9**.

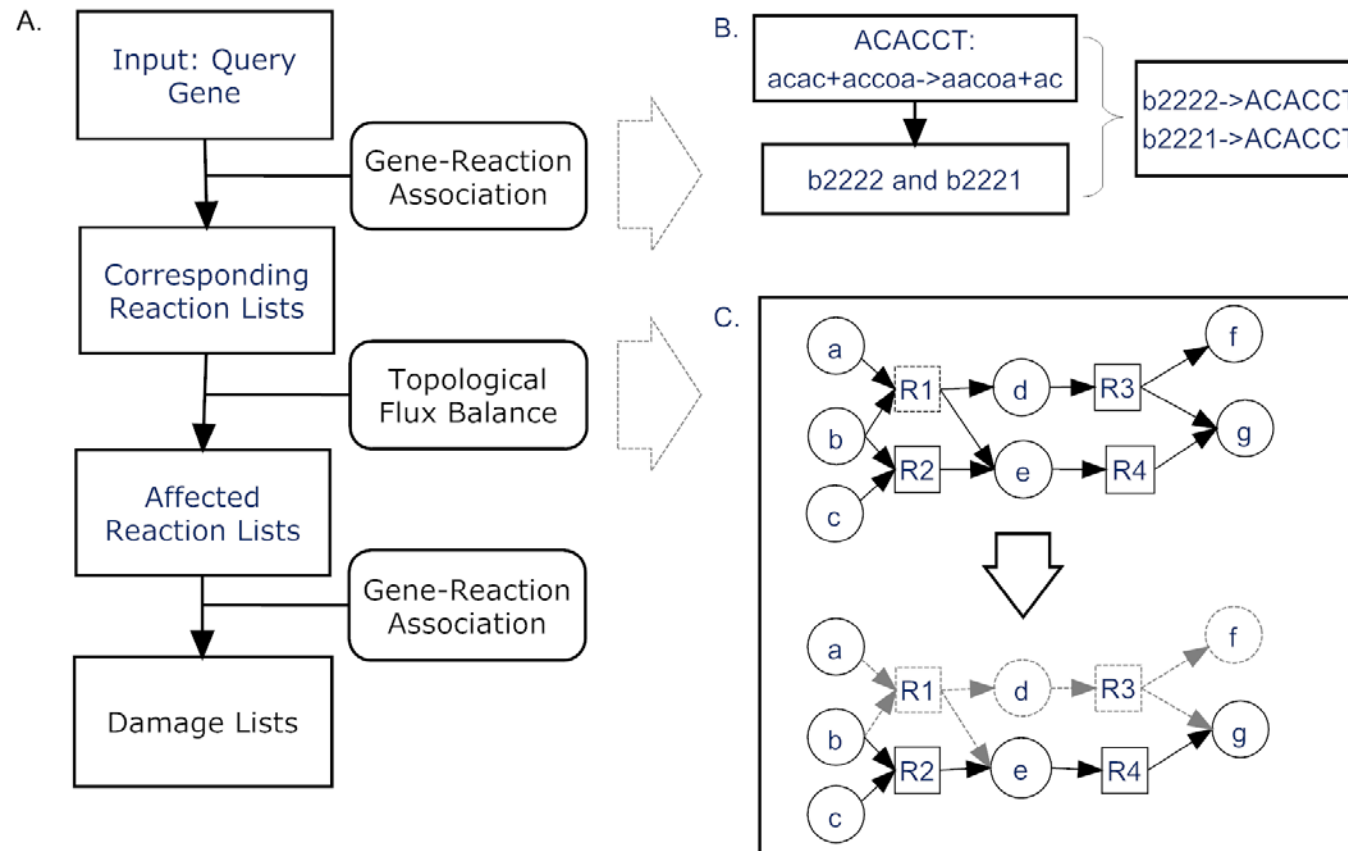


Figure 9 Overview of the self-devised algorithm. (A) The workflow of this algorithm. For any query gene, corresponding reactions can be identified through analyzing the gene-reaction association. The obtained corresponding reactions are the inputs for the topological flux balance, whose output is a list of affected reactions. These affected reactions are mapped to the corresponding genes based on gene-reaction association to find out genes whose function can be affected in response to the query gene deletion. (B) For reaction associated with genes in logic ‘AND’ relations, any gene is essential for the proper function. Therefore, this reaction is the corresponding reaction for either gene. For example, reaction ACACCT is catalyzed by the encoded protein of gene b2222 and b2221. This reaction is the corresponding reaction for either gene b2222 or b2221. (C) For reaction R1 deletion, initially all the edges connected to this reaction are removed, followed by metabolite nodes whose in-degree or out-degree becomes zero (such as metabolite node: d). Once metabolite d is removed from the network, reaction R3 cannot occur anymore since incompleteness of substrates. Such procedures are repeated until all the leftover nodes are satisfied with the criteria. In the end, metabolites d and f, and reaction R3 will be removed from the network. R3 is the affected reaction in response to R1 removal.

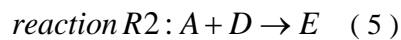
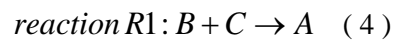
2.4 Flux balance analysis (FBA)

Flux balance analysis (FBA) is a widely used constraint-based approach in studying the genome-scale metabolic network [73, 142], which applying the linear optimization to determine the steady-state reaction flux distribution by maximizing or minimizing the objective functions. One of the key assumptions underlying FBA is that cells can perform optimally with respect to the different objective functions under the evolutionarily pressure [142]. The objective functions to be optimized may vary, depending on the purpose of the study. It may be maximization of the biomass production [143], or minimization of the nutrients requirement [144, 145]. Once the objective function is determined, the flux optimization problems can be solved to obtain the steady-state flux distribution by

adding the lower and upper boundary constraints to each flux. The obtained flux solutions may provide some insights into the metabolic network and some underlying biological processes [50, 146]. For example, FBA is applied to yeast metabolic network to study gene epistatic interactions, which helps to understand the functional organization of biological networks [50]. As the various forms of the FBA, it can be customized to solve different problems, such as predicting the phenotypes for single gene deletion strains [147, 148], the phenotypes under different carbon sources [145, 149]. In certain cases, FBA can even unveil interactions within the networks [150-152]. The FBA problems can be solved through two steps: mathematical representation and optimization.

2.4.1 Mathematical representation of FBA

The FBA problems can be converted to linear programming optimization problem. In the example below, metabolite A participates as a product in reaction R1 and substrate in reaction R2.



The concentration change with respect to time can be represented by the equation:

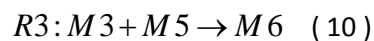
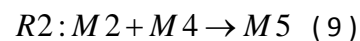
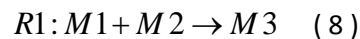
$$\frac{d[A]}{dt} = v(r1) - v(r2) \quad (6)$$

where $v(r1)$ and $v(r2)$ are the flux through reaction R1 and R2 respectively, $[A]$ is the concentration of metabolite A.

At steady state, the consumption rate of metabolites should equal to the production rate, in other words, the time derivative of the concentration equals to zero. For each metabolite, we can generate one such equation. Mathematically, it means for each metabolite M, it should satisfy the relation:

$$\frac{d[M]}{dt} = 0 \quad (7)$$

These equations can instead be represented as the product of a matrix and a vector, where the vector is the flux through each reaction and the matrix is the stoichiometric coefficient. Here is an example:



Six metabolites are involved in three reactions. As a consequence, the stoichiometric matrix is 6 by 3, where each column represents one reaction and each row represents one metabolite. Each entry is either 1 or 0, or -1, where 1 indicates it acts as product, -1 as substrate and the leftover entries are filled with zero.

$$S = \begin{bmatrix} -1 & 0 & 0 \\ -1 & -1 & 0 \\ 1 & 0 & -1 \\ 0 & -1 & 0 \\ 0 & 1 & -1 \\ 0 & 0 & 1 \end{bmatrix} \quad (11)$$

Here is a generalized way to obtain the matrix. For each single metabolite, the equation (6) can be rewritten as:

$$\frac{dA}{dt} = v(r1) - v(r2) = S * v \quad (12)$$

where $S = [1; -1]$ and $v = [v(r1), v(r2)]$.

For any metabolite, it has one such equation and to sum up, we obtained an S-matrix with dimension M by N, where M is the number of metabolites involved in the network and N is the number of reactions involved. Each row of the matrix represents one metabolite and each column represents one reaction. The entry of this matrix should be the negative (positive) stoichiometric coefficient if the metabolite is the substrate (product) of the reaction. In other cases, the entry value will be zero. The stoichiometric matrix is a sparse matrix as for most of the reactions, only a few compounds are involved. The dimension of the vector v should be N by 1. At steady state, the concentration change should be zero for any internal metabolites. Consequently, the equality constraints are obtained: $S * v = 0$.

As the number of metabolites is larger than the number of reactions, this equality constraint is under-determined. That means additional constraints are required to solve the problem. Commonly, the lower and upper boundaries are added for each flux.

One of the commonly used objective functions is the biomass production, which is frequently used as a measure of cellular growth, in terms of the biosynthetic requirements to produce a cell. The following is the biomass components in *E. coli* K-12 model (**Table 4**) [153], including some key biosynthetic precursors.

Table 4 Biomass components.

Coefficient	Metabolites	Coefficient	Metabolites
0.001	AMP	0.488	L-Alanine
0.0247	dATP	0.281	L-Arginine
0.0254	dCTP	0.229	L-Asparagine
0.0254	dGTP	0.229	L-Aspartate
0.0247	dTTP	0.087	L-Cysteine
45.7318	ATP	0.25	L-Glutamine
0.126	CTP	0.25	L-Glutamate
0.203	GTP	0.09	L-Histidine
0.136	UTP	0.276	L-Isoleucine
0.003	UDP-glucose	0.428	L-Leucine
0.035	Putrescine	0.326	L-Lysine
0.007	Spermidine	0.146	L-Methionine
0.582	Glycine	0.176	L-Phenylalanine
0.154	Glycogen	0.21	L-Proline
45.5608	H ₂ O	0.205	L-Serine
0.05	5-Methyltetrahydrofolate	0.241	L-Threonine

0.00005	Acetyl-CoA	0.054	L-Tryptophan
0.000006	Coenzyme A	0.131	L-Tyrosine
0.00001	FAD	0.402	L-Valine
0.00215	Nicotinamide adenine dinucleotide	0.0084	Lipopolysaccharide
0.00005	Nicotinamide adenine dinucleotide - reduced	0.001935	Phosphatidylethanolamine
0.00013	Nicotinamide adenine dinucleotide phosphate	0.0276	Peptidoglycan subunit of Escherichia coli
0.0004	Nicotinamide adenine dinucleotide phosphate - reduced	0.000464	Phosphatidylglycerol
0.000129	Cardiolipin	0.000052	Phosphatidylserine
0.000003	Succinyl-CoA		

Different objective functions are used for diverse purposes. To maximize the growth rate, it is preferred to set the biomass equation as our object. Hence the flux balance analysis problem can be represented at follows:

$$\text{objective: } c * \text{biomass} \quad (13)$$

$$\text{subject to: } S * v = 0 \quad (14)$$

$$lb_i \leq v_i \leq ub_i \quad (15)$$

where c is the coefficient of metabolites in biomass equations, S is the stoichiometric matrix, v is the vector of flux through each reaction, and equation (15) is the lower and upper boundary constraint on each individual flux.

2.5 Essential gene lists

Essential genes are defined as those genes necessary for cellular functions under the given medium, which is usually the rich medium. It is usually assessed through single gene knockout experiment. A series of single gene deletion strains were constructed and the final lethality phenotype suggested the essential genes. Experimental results for *E. coli* are quite abundant. For example, a genetic footprinting technology was applied to identify the gene essentiality in the *E. coli* K-12 model under the uniform growth conditions, i.e. logarithmic aerobic growth of strain MG1655 in enriched LB medium [91]. Another list is the Profiling of *E. coli* Chromosome database (PEC database) [94], where the information of essentiality was concluded from a systematic review of experimental literature. In

this database, 302 genes are classified as essential genes, 4432 as non-essential while the left 5 are unknown. However, the most complete one should be the Keio collection [66], where a genome-scale single gene deletion in the *E. coli* K-12 was conducted. For the 4288 target genes, 3985 of them can obtain mutant strains. For the leftover 303 genes that cannot have the mutant strains, they are considered as essential genes. 3 of them have no corresponding gene name or the b-name, leaving 300 genes. In our studies, all the calculations are based on the Keio collection, whereas in the discussion part, PEC database is also used for cross-reference. For yeast, the essential gene list is obtained based on Saccharomyces Genome Deletion Project ([http://wwwsequence.stanford.edu/group/-yeast_deletion_project/deletions3.html](http://www.sequence.stanford.edu/group/-yeast_deletion_project/deletions3.html)).

Chapter 3 Single gene deletion analysis

3.1 Introduction

Living organisms are dramatically robust, resistant to various genetic perturbations and environmental stress [154]. It is reported that more than 80% of *S. cerevisiae* gene mutants are viable under rich glucose medium [88]. Genome-scale single gene knockout study in *E. coli* also showed similar results [66]. Yet, some specific knockouts of single gene, the so-called 'essential genes' can be lethal to organisms [78], by halting cell growth due to their essential roles in crucial biological functions. Thus, essential genes have been conventionally considered as promising targets for antimicrobial drug development [155-158].

Typically, essential genes were determined by systematic single gene knockout experiments [66]. Genes whose disruption mutants cannot grow under the given medium (generally rich medium) were considered as essential genes. Previous studies [159, 160] on essential genes mainly concern the final metabolic phenotype (lethal/survival) following the genetic perturbation. Such approaches were widely used to predict gene essentiality [142, 161-163].

As indicated by genome-scale gene deletion experiments [66, 88], most genes are dispensable without causing any significant impact on the cellular fitness. It raises the attention to study the cause and evolution of mechanistic basis for

dispensability. Three kinds of explanations are proposed: (1) Conditional essentiality, i.e. those non-essential genes are indeed essential in other conditions that not examined [103]. Studies [92, 104, 105] aiming to identify conditional essential genes have been implemented. (2) Genetic redundancy [103, 106], in other words, a gene's function may be buffered by duplicates or functional overlapping genes. (3) Alternative pathways or redistribution of flux distribution [103, 107]. The latter two explanations involve functional back-up mechanisms.

Previous studies showed that around 74% of yeast *S. cerevisiae* metabolic genes can contribute to some functional essential processes, although only 13% found to be essential [108]. It indicated that non-essential genes may be backed up through certain mechanisms. The intermediate processes, which connect the genetic perturbation (input) to their corresponding phenotypes (output), were less studied. Also, it is unclear to what extent an essential gene is correlated to other genes. Thus, a clear picture of these internal changes occurred in the intermediate processes may help us to bridge the gap from genotypic events to phenotypic consequences and enrich our understanding of the causes of gene essentiality. In this study, we focused on understanding the altered downstream processes following essential/non-essential gene deletion by conducting a thorough investigation of the cascade failure processes in genome-scale metabolic network. Knowledge obtained may unveil the topological/functional basis for the observed discrepancy between essential and non-essential gene perturbations.

In the context of metabolic network, the functions of enzyme-coding genes can be reflected by the activities of the corresponding biochemical reactions (reactions that are directly catalyzed by enzymes encoded by the gene). Hence, we characterize the effect of gene knockout on the whole systems by deleting the corresponding reactions in the metabolic network. However, at the steady state achieved following gene perturbation, the knockout effects are not limited to its own functions, but can spread to other functional related genes [164, 165], analogous to the ripple effects produced when a pebble thrown into the pond. As a result, the observed phenotype should be the cooperative or emergent effects produced by both the deleted genes and perturbed downstream processes, which we termed as ‘damage list’ in this study. Analysis of this ‘damage list’ may enrich our understanding of the differential deletion effects between essential and non-essential genes.

In this study, we modified and repurposed a previous method [74] to capture the deletion effects for any given gene knockout using metabolic network of *E. coli* [17], one of the most well studied models available. By incorporating essential gene information, we revealed the cooperative nature of metabolic essential genes and it suggested a possible mechanism concerning how essential genes can lead to lethal phenotypes. The obtained properties were quite conserved across species, found also in *S. cerevisiae*, indicating evolutionarily conserved features. We further showed that findings obtained from analyzing the structural organization of essential genes in metabolic network provided the structural basis for the

observed cooperative effects among essential genes, implying close relations between gene essentiality and network structure.

3.2 Materials and methods

3.2.1 Statistical hypothesis test

Statistical hypothesis testing is a method widely used in scientific researches. A result is considered statistically significant if it is not likely to occur by chance, according to a pre-determined threshold probability (significance level), which is usually set to 0.05. Typically, research question is expressed in terms of there being no difference between two groups, which is known as null hypothesis. A probability ' p -value' stands for the probability that how likely that any observed differences between groups is due to chance. It is the determinant whether the null hypothesis is to be accepted or rejected. A smaller p -value, generally smaller than the pre-determined 'significance level', indicates that the observed difference is not due to chance. As a result, the null hypothesis is rejected. On the other hand, a p -value larger than the 'significance level' suggests that the observed difference is due to random variations, therefore the null hypothesis is accepted.

3.2.2 Null model

The null model used was constructed via rewiring the edges in the metabolic network. We randomly chose 2 different edges from the network, for example, edge 1 (a → b) and edge 2 (c → d). Here, a, b, c, and d represents nodes in the network. Then we swapped these two edges, forming the new ones: edge 1' (a → d) and edge 2' (c → b). Such procedures were repeated 2 times the total number of edges. The generated null model preserved the degree distribution of the real network. In our study, totally 1,000 null models were generated.

3.2.3 Gene essentiality information

Two *E. coli* essential gene lists are used in this study. One is from Keio collection [66] whereas the other is from PEC database [94]. Essential genes identified by Keio collection are used as the input whose deletion effects are compared with non-essential genes. Essential genes from PEC are used for cross-reference. For yeast, the list is based on Saccharomyces Genome Deletion Project (http://www.sequence.stanford.edu/group/yeast_deletion_project/deletions3.html).

3.2.4 Metabolic network reconstruction

As described in **Chapter 2**, networks are reconstructed from *E. coli* iAF1260 and *S. cerevisiae* iND1260 model. In the bipartite network, two kinds of nodes are available, metabolite node and reaction node. Enzyme-coding gene information is

stored for each reaction node. Our *E. coli* metabolic network includes 2351 nodes, and 4038 edges whereas *S. cerevisiae* with 1441 nodes and 2596 edges.

3.2.5 Computational algorithm of cascading failure in metabolic network

Single gene deletion is initiated by removing the corresponding reaction nodes and their links from the reconstructed bipartite metabolic network, followed by iterated procedures which search upwards and downwards the network until no further reaction or metabolite nodes to be removed, i.e. the leftover reactions are with complete substrates and products whereas metabolite nodes with non-zero in-degree or out-degree. Detailed procedures are described in **Section 2.3, Chapter 2**. A pseudo-code can be found in the **Appendix**.

3.3 Single gene deletion

3.3.1 Single gene deletion in metabolic network

The aim of this study is to investigate the mechanistic explanations why some genes (essential genes) are more important than others. To investigate the underlying reason, we studied the differential single gene deletion effects and then identified why some genes are more significant than others.

A bipartite metabolic network of *E. coli* was reconstructed from model *E. coli* iAF1260 [17] consisting of 2351 nodes (1354 reaction nodes and 987 metabolite nodes) and 4038 edges. There are totally 1261 genes in the *E. coli* metabolic networks, and among them 132 are classified as essential genes according to the Keio collection [66]. We initiated the gene failure by removing single gene's corresponding reactions from the bipartite network and then propagated the failure upwards and downwards until the leftover nodes were satisfied with the criteria, which require each internal metabolite with non-zero in-degree and out-degree and reactions with complete substrates and products. The consequential effects were characterized by a set of genes whose corresponding reactions would be removed according to the previous iterations. We defined damage list as a set of genes whose corresponding reactions can be impaired and the number of genes within the damage list called damage size, is denoted as d . Essential and non-essential genes are labeled as E and N , respectively, and will be subsequently used hereafter.

3.3.2 Essential gene deletion induces large damage list

Based on the observation that single gene deletion can produce distinct phenotypes (lethal/growth), one may expect that their discrepancy can be reflected from the damage lists, i.e. those genes affected due to target gene deletion. So we first studied the relation between damage size and gene essentiality type. A comparison of the damage size of these two types of genes (essential and non-essential genes) revealed that essential genes generally have a larger damage size,

in other words, the essential gene deletion causes wider spreading of perturbations in metabolic network. While more than 10% essential genes had a damage size bigger than 15, only 2% for the non-essential genes. Besides, around 50% non-essential genes had a damage size of zero (**Figure 10**). However, large damage size does not necessarily imply essentiality as we can still observe a fraction of non-essentials with large damage size. Hence, it is rather intriguing to decipher the underlying properties that distinguish essentials from non-essentials.

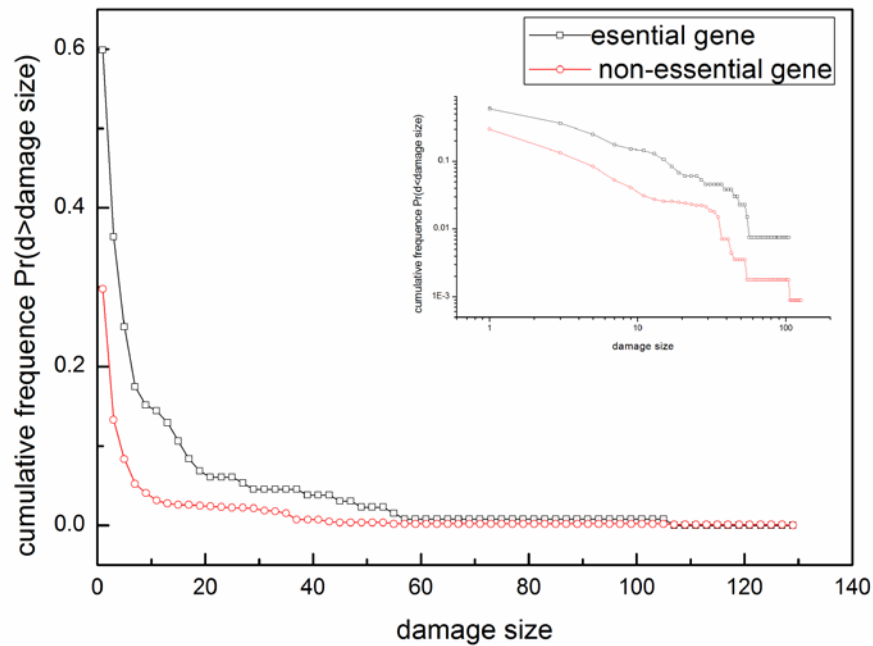


Figure 10 Cumulative distribution of damage size. The cumulative distribution for essential and non-essential genes is shown, where rectangle (above) represents essential genes and circle (below) node represents non-essential genes. The inset is in the log-log scale (above is essential gene, and below is non-essential gene).

3.3.3 Analysis of components within damage list

An in-depth examination revealed some significant patterns in the damage lists. We used variables e and n to denote the number of essential and non-essential genes in the damage list of each gene, respectively. Genes with zero damage size were excluded from our studies. As a result, there are 300 E->E pairs, 186 E->N pairs, 103 N->E pairs and 1676 N->N pairs obtained with 116 essential genes and 533 non-essential genes participated. Here, an E->E pair can be interpreted as the removal of the former essential gene can lead to the functional loss of the latter

essential gene. In other words, the latter essential gene is included in the former essential gene's damage list. Our result suggested that genetic perturbation of one essential gene can always induce additional functional failure of other essential and non-essential genes, whereas majority of non-essential genes mainly exert their impact on non-essential genes. We validated the observed discrepancies between essential and non-essential genes using the χ^2 tests (<http://faculty.vassar.edu/lowry/odds2x2.html>), with p -value smaller than 0.0001 (**Table 5**).

Table 5 Damage list composition. Each entry represents the observed (or expected) number of corresponding gene pairs.

#gene pairs	Essential	Non-Essential
Essential	300(86)	186(399)
Non-Essential	103(317)	1676(1462)

Further, we investigated the composition of each damage list by using the ratio of essential and non-essential genes in the damage lists (**Figure 11**). The range of $(e/(e+n))$ varies from 0 to 1, where 1 indicates the damage list is exclusively composed of essential genes (non-essential genes) and 0 denotes the damage list is solely made up of non-essential genes (essential genes). For a well-sorted gene list based on essentiality, a clear and strong pattern showed up that is consistent with our previous observations.

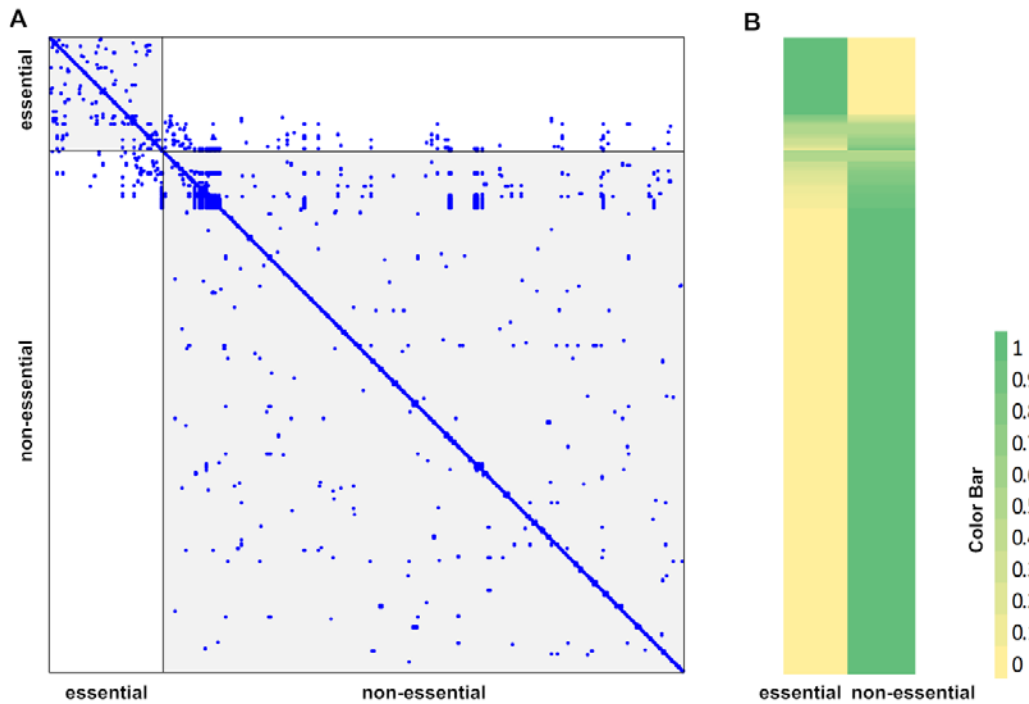


Figure 11 Distinct deletion effects between essential and non-essential genes. (A) Each row and column corresponds to one gene, sorted according to essentiality and the gene order in (B). A blue spot indicates that gene knockout (the corresponded row gene) can perturb the corresponding column gene function. Knockout of essential genes can cause a large range of essential gene function failure (as shown in the upper grey area) while knockout of non-essential genes mainly perturb non-essential genes function (as shown in the lower grey area). A quantitative representation is shown in (B), where each row is the same as in (A) while column is the percentage of essential genes and non-essential genes in the damage lists. The color is illustrated based on the color bar, where 1 is labeled as green and 0 is labeled as yellow.

Briefly, analysis on the damage list composition revealed a distinctive difference between these two types of genes, implying that the cooperative effects of essential genes may have some connections with essentiality.

3.3.4 Genes with similar damage lists share the same essentiality

Since the obtained damage list can capture the effects of single gene deletion on the whole metabolic network, we therefore considered whether two genes with highly similar damage list tend to function via similar mechanisms. To test the hypothesis, we used Jaccard similarity coefficient [166], one of the most widely used similarity coefficients in the literatures, which is defined as the size of intersection divided by the size of the union of the genes' corresponding damage lists in this given context (**Figure 12**). We linked those gene pairs (node) with a similarity coefficient bigger than a threshold (in this study, we chose 0.6). It is intriguing that many sub-networks were formed, within which nodes were tightly associated (**Figure 13**). In most cases, genes within the same subnetwork shared the same property of essentiality (either essentials or non-essentials) and participated in the same or related pathways.

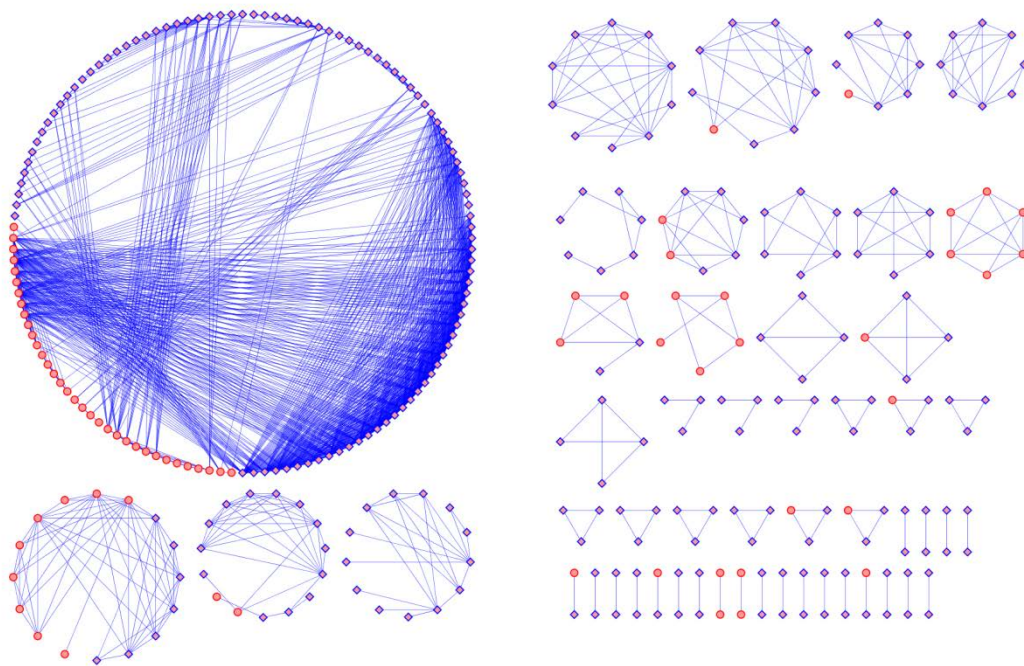


Figure 12 Overview of damage list similarity. Gene pairs with their damage list similarity bigger than 0 are connected. Nodes with red circle border and blue diamond border represent essential and non-essential genes respectively.

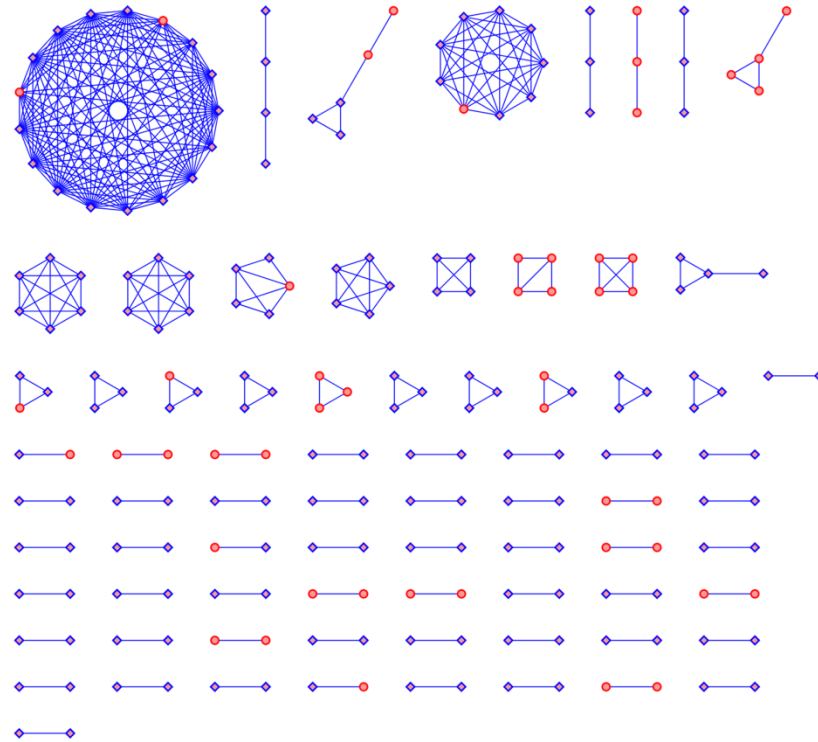


Figure 13 Essentiality consistency within gene subnetworks sharing similar damage lists. Gene pairs with large overlapping damage lists (bigger than 0.6) are connected. Genes within the same subnetwork tend to have the same essentiality. Nodes with red circle border represent essential genes while nodes with blue diamond border represent non-essential genes.

Quantitatively, we designed a score function $|N_N^i - N_E^i| / (N_N^i + N_E^i)$ to evaluate the essentiality consistence within the same subnetwork, where N_E^i and N_N^i denote the number of essential and non-essential genes within the *ith* subnetwork respectively. The score ranges from 0 to 1, where 0 represents there is no essentiality preference within the subnetwork whereas 1 represents that genes within the subnetwork are biased to have the same essentiality property. The averaged score across all the subnetworks denoted the overall essentiality consistence, which in our case is up to 0.9. Our results indicated that the genes within each subnetwork were of high possibility to share the same essentiality property with others genes from the same subnetwork. Besides, our results were robust and invariant with respect to the threshold. (**Table 6**)

Similar results were obtained when choosing other similarity coefficients (e.g. overlap coefficient $(D1 \cap D2) / \min(D1, D2)$, or dice coefficient $2(D1 \cap D2) / (|D1| + |D2|)$). Furthermore, analogous observations can also be found in other species, such as *S. cerevisiae*, suggesting its universality (**Table 7**).

Table 6 Essentiality consistency within genes sharing highly similar damage lists in *E. coli*. #group, # essential gene, and #non-essential gene represents the number of subnetworks obtained, essential genes, and non-essential genes for the given threshold.

threshold	#group	#essential	#non-essential	score
0.4	78	61	192	0.87
0.5	80	55	186	0.90
0.6	76	49	167	0.91
0.7	69	39	146	0.90
0.8	64	34	132	0.91
0.9	59	29	120	0.89

Table 7 Essentiality consistency within genes sharing highly similar damage lists in *S. Cerevisiae*. #group, #essential gene, and #non-essential gene represents the number of subnetworks obtained, essential genes, and non-essential genes for the given threshold.

threshold	#group	#essential	#non-essential	score
0.4	39	29	94	0.86
0.5	39	27	87	0.87
0.6	37	24	73	0.87
0.7	32	24	60	0.84
0.8	29	18	52	0.91
0.9	28	13	49	0.86

Several subnetworks were extracted for further analysis. In **Figure 13**, the largest subnetwork is composed of 15 non-essential genes and 2 essential genes for the given threshold 0.6 (b3624, b3631, b3632, b0200, b3052, b3628, b3629, b3630, b3198, b3627, b1855, b2040, b3623, b0918, b3619, b3625, and b3626). As we increased the threshold, the subnetwork is totally composed of non-essential genes (as shown in **Figure 14**). It suggests that the most robustness part in this subnetwork is made up of those non-essential genes.

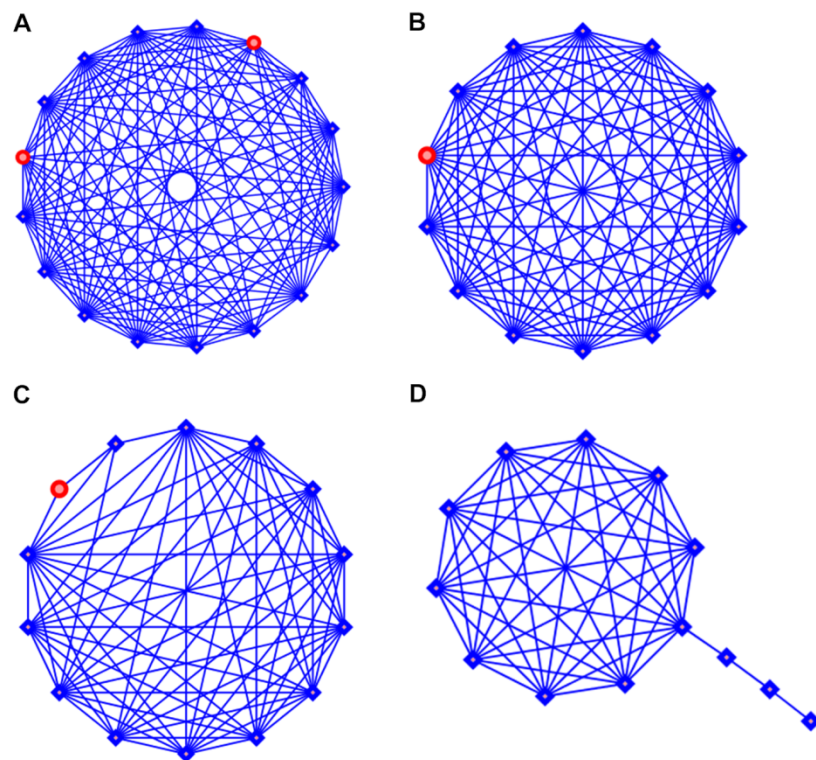


Figure 14 Evolution of the largest subnetwork. (A) The largest subnetwork for the threshold 0.6 as shown in **Figure 13**. (B), (C), and (D) are at threshold 0.7, 0.8 and 0.9, respectively. Increasing the threshold removes the essential genes from the subnetwork, indicating non-essential genes are the most robust components.

In this subnetwork, 16 out of 17 genes are from the subsystem Lipopolysaccharide Biosynthesis/Recycling, involved in LPS core biosynthesis. Here are the top five GO functions: GO: 0009103 (Lipopolysaccharide biosynthetic process) GO: 0016740 (transferase activity) GO: 0005515 (protein binding) GO: 0016757(transferring glycosyl groups) and GO: 0016020 (membrane) (**Table 8**). We also analyzed the common damaged genes of this subnetwork: b0622 (*pagP*), b2027 (*cld*), b2034 (*wbbl*), b2035 (*rfc*), b2040 (*rfaB*), b2254 (*arnC*), b2255 (*arnA*), b2257 (*arnT*), b3622 (*rfaL*), b3623 (*waaU*), b3624 (*rfaZ*), b3626 (*rfaJ*), b3627 (*rfaI*), b3628 (*rfaB*), b3620 (*rfaF*), b3631 (*rfaG*), b3785 (*wzzE*), b3790 (*rffC*), b3793 (*wzyE*), b3794 (*rffM*), and b4481 (*rffT*). These common genes are non-essential genes except the gene b3623 (*waaU*), and b3793 (*wzyE*) according to the essential gene list identified through the Keio collection [66]. However, these two genes were considered as non-essential genes in other studies [94]. So according to our studies, these two genes were of high possibility to be non-essential. They may have some important roles, but the lack of them may not lead to cell lethality.

Table 8 Top five GO functions in the largest subnetwork.

Accession	Ontology	Synonyms
GO:0009103	Biological Process	Lipopolysaccharide biosynthesis
GO:0016740	Molecular Function	Transferase activity
GO:0005515	Molecular Function	Protein amino acid binding
GO:0016757	Molecular Function	Glycosyltransferase activity
GO:0016020	Cellular Component	Membrane

3.3.5 Associated gene sets are necessary for survival

Taking the previous findings relating the composition of damage lists into consideration, we noticed that essential genes within the same subnetwork share some common damage lists, where the loss can lead to lethality. Subnetwork consisting of genes *murE* (b0085), *murD* (b0088), *murC* (b0091), and *murI* (b3967) was extracted for further analysis. It is revealed that their damage lists share two common genes: *murE* (b0085) and *murD* (b0088). Interestingly, the corresponding reactions of these genes can disturb the production of UDP-N-acetylmuramoyl-L-alanyl-D-gamma-glutamyl-meso-2, 6-diaminopimelate, a key metabolite involved in the cell wall biosynthesis [109]. Another example was the subnetwork: *lpxH* (b0524), *lpxB* (b0182), *lpxK* (b0915), and *kdtA* (b3633). The common damage lists of these genes were made up of *lpxC* (b0096), *lpxD* (b0179), *lpxB* (b0182), *lpxH* (b0524), and *lpxK* (b0915), which can perturb the production of lipid A, a principle and essential component for bacteria growth [167]. In this regard, we proposed that single gene deletion propagated its effects via these common damage lists, which were termed as ‘associated gene sets’ in our studies.

Combined with previous findings that essential genes were distinguished from non-essential genes in respect of their capability to impact on other essential genes, it was estimated that the common damage lists required for survival should be composed mainly of essential genes rather than non-essential genes. For each essential gene, the affected essential genes in its damage lists form one candidate

for ‘associated gene sets’. In case one set can cover another completely, the minimal one was kept. Redundant modules were excluded. As a consequence, 72 associated gene sets were identified with size ranging from 1 to 5, among which 50 associated gene sets were of size 1, 17 were of size 2, while the leftover were of size more than 2. It is expected that most gene sets are of size 1 as it is consistent with our common understanding of essential genes which are defined as those genes whose deletion can lead to lethality [78]. It is intriguing to identify some associated gene sets with size more than 1 as it implies the significance of genetic interactions in the context of essentiality and gains new insights on the drug development.

Genes within these sets mainly participated in the cell envelope biosynthesis, tRNA charging, nucleotide salvage pathway, or cofactors and prosthetic group biosynthesis subsystems. Besides, we found that the corresponding reactions of genes from each set were generally involved in the production of key metabolites, which were considered necessary for cell survival in other studies [109]. For instance, 14 sets of size 1 were composed of genes that encode aminoacyl-tRNA synthetase, which catalyze the attachment of a specific amino acid to its compatible cognate tRNA to form an aminoacyl tRNA. The produced aminoacyl tRNA plays crucial roles in translating during protein synthesis [168]. Therefore, we proposed that knockout of essential genes can spread its effects to such kind of associated gene sets, whose failure may lead to a lack of some key metabolites,

which are necessary for cell growth. **Table 9** summarized how the loss of these sets can result in lethality.

Table 9 Associated gene sets

Subsystem	Gene Set	Metaoblite	Ref
tRNA charging	b0194	L-Prolyl-tRNA(Pro)	[168]
tRNA charging	b0526	L-Cysteinyl-tRNA(Cys)	[168]
tRNA charging	b0642	L-Leucyl-tRNA(Leu)	[168]
tRNA charging	b0680	L-Glutaminyl-tRNA(Gln)	[168]
tRNA charging	b0893	L-Seryl-tRNA(Ser)	[168]
tRNA charging	b0930	L-Asparaginylyl-tRNA(Asn)	[168]
tRNA charging	b1637	L-Tyrosyl-tRNA(Tyr)	[168]
tRNA charging	b1719	L-Threonyl-tRNA(Thr)	[168]
tRNA charging	b1866	L-Aspartyl-tRNA(Asp)	[168]
tRNA charging	b1876	L-Arginyl-tRNA(Arg)	[168]
tRNA charging	b2514	L-Histidyl-tRNA(His)	[168]
tRNA charging	b3384	L-Tryptophanyl-tRNA(Trp)	[168]
tRNA charging	b3560	Glycyl-tRNA(Gly)	[168]
tRNA charging	b4258	L-Valyl-tRNA(Val)	[168]
tRNA charging	b1713,b1714	L-Phenylalanyl-tRNA(Phe)	[168]
tRNA charging	b2114,b3288	L-Methionyl-tRNA (Met); N-Formylmethionyl-tRNA	[168]
Cofactor and Prosthetic Group Biosynthesis	b0025	FMN; FAD	[109]
Cofactor and Prosthetic Group Biosynthesis	b0174	Undecaprenyl diphosphate	[109]
Cofactor and Prosthetic Group Biosynthesis	b0417	Thiamine diphosphate	[169]

Cofactor and Prosthetic Group Biosynthesis	b0420	Glyceraldehyde 3-phosphate	[109]
Cofactor and Prosthetic Group Biosynthesis	b0583	(2,3-Dihydroxybenzoyl)adenylate	[125]
Cofactor and Prosthetic Group Biosynthesis	b1740	Nicotinamide adenine dinucleotide	[109]
Cofactor and Prosthetic Group Biosynthesis	b2153	2-Amino-4-hydroxy-6-(erythro-1,2,3-trihydroxypropyl) dihydropteridine triphosphate	[109]
Cofactor and Prosthetic Group Biosynthesis	b2315	Dihydropteroate	[109]
Cofactor and Prosthetic Group Biosynthesis	b2615	Nicotinamide adenine dinucleotide phosphate	[109]
Cofactor and Prosthetic Group Biosynthesis	b3650	Guanosine 3',5'-bis(diphosphate)	[170]
Cofactor and Prosthetic Group Biosynthesis	b3835	2-Octaprenylphenol	[171]
Cofactor and Prosthetic Group Biosynthesis	b0029,b2515	Isopentenyl diphosphate; Dimethylallyl diphosphate	[172]
Cofactor and Prosthetic Group Biosynthesis	b0173,b2747	2-C-methyl-D-erythritol 4-phosphate	[173]

Prosthetic Group Biosynthesis			
Cofactor and Prosthetic Group Biosynthesis	b0369,b3805	Uroporphyrinogen III	[174]
Cofactor and Prosthetic Group Biosynthesis	b0415,b1662	6,7-Dimethyl-8-(1-D-ribityl)lumazine; 3,4-dihydroxy-2-butanone 4-phosphate	[109]
Cofactor and Prosthetic Group Biosynthesis	b0414,b1277	4-(1-D-Ribitylamino)-5-aminouracil	[109]
Cofactor and Prosthetic Group Biosynthesis	b1210,b2400	L-Glutamyl-tRNA(Glu)	[168]
Cofactor and Prosthetic Group Biosynthesis	b3634,b3639	Pantetheine 4'-phosphate; Dephospho-CoA	[109]
Cofactor and Prosthetic Group Biosynthesis	b0475,b3850	Heme	[175]
Nucleotide Salavage Pathway	b0474	AMP; ADP; IDP; etc	[109]
Nucleotide Salavage Pathway	b0171	UMP; UDP; dUMP; dUDP	[109]
Nucleotide Salavage	b1098	dTMP;dTDP	[109]

Pathway			
Nucleotide Salavage Pathway	b3648	GMP; ATP; dGMP	[109]
Nucleotide Salavage Pathway	b2234,b2235	Reduced thioredoxin; Oxidized thioredoxin	[109]
Cell Envelope Biosynthesis	b2533	dsRNA	[176]
Cell Envelope Biosynthesis	b3729	D-Glucosamine 6-phosphate	[109]
Cell Envelope Biosynthesis	b3730	N-Acetyl-D-glucosamine 1-phosphate; UDP-N-acetyl-D-glucosamine; D-Glucosamine 1-phosphate	[109]
Cell Envelope Biosynthesis	b0085,b0088	UDP-N-acetylmuramoyl-L-alanyl-D-gamma-glutamyl-meso-2,6-diaminopimelate; UDP-N-acetylmuramoyl-L-alanyl-D-glutamate	[109]
Cell Envelope Biosynthesis	b0087,b0090	Undecaprenyl-diphospho-N-acetylmuramoyl-L-alanyl-D-glutamyl-meso-2,6-diaminopimeloyl-D-alanyl-D-alanine; Undecaprenyl-diphospho-N-acetylmuramoyl-(N-acetylglucosamine)-L-ala-D-glu-meso-2,6-diaminopimeloyl-D-ala-D-ala	[109]
Cell Envelope Biosynthesis	b3189,b3972	UDP-N-acetyl-D-glucosamine; UDP-N-acetyl-3-O-(1-carboxyvinyl)-D-glucosamine	[109]
Cell Envelope Biosynthesis	b0954,b1093,b2323	short-chain unsaturated acyl-ACP; short and long chain saturated and unsaturated β -ketoacyl-ACPs	[109, 177]
Cell Envelope Biosynthesis	b1093,b1288,b2323	short-chain unsaturated acyl-ACP; short and long chain saturated and unsaturated β -ketoacyl-ACPs	[109]
Threonine and	b2472	LL-2,6-Diaminoheptanedioate	[109]

Lysine Metabolism			
Threonine and Lysine Metabolism	b3433	4-Phospho-L-aspartate; L-Aspartate 4-semialdehyde	[109]
Threonine and Lysine Metabolism	b0031,b0166	2,3-Dihydrodipicolinate; 2,3,4,5-Tetrahydrodipicolinate	[109]
Threonine and Lysine Metabolism	b0031,b2478	2,3-Dihydrodipicolinate; L-Aspartate 4-semialdehyde	[109]
Lipopolysaccharide biosynthesis/recycling	b1215	3-Deoxy-D-manno-octulosonate 8-phosphate	[109]
Lipopolysaccharide biosynthesis/recycling	b3793	Undecaprenyl diphosphate	[109]
Lipopolysaccharide biosynthesis/recycling	b0096, b0181,b1094	UDP-3-O-(3-hydroxytetradecanoyl)-D-glucosamine	[109]
Lipopolysaccharide biosynthesis/recycling	b0096,b0179,b0182,b0524,b0915	LipidA	[167]
Glycerophospho	b1912	Phosphatidylglycerophosphate	[109]

lipid			
Metabolism			
Glycerophospho	b3018	acyl carrier protein	[109]
lipid			
Metabolism			
Glycerophospho	b4041	Glycerol 3-phosphate	[109]
lipid			
Metabolism			
Glycerophospho	b2585,b4160	phosphatidylserine; phosphatidylethanolamine	[109]
lipid			
Metabolism			
Glycolysis/Gluc	b1779	Glyceraldehyde 3-phosphate; 3-Phospho-D-glyceroyl phosphate	[109]
oneogenesis			
Glycolysis/Gluc	b2779	D-Glycerate 2-phosphate; Phosphoenolpyruvate	[109]
oneogenesis			
Glycolysis/Gluc	b2926	3-Phospho-D-glycerate; 3-Phospho-D-glyceroyl phosphate	[109]
oneogenesis			
Purine and	b1131	N6-(1,2-Dicarboxyethyl)-AMP	[109]
Pyrimidine			
Biosynthesis			
Purine and	b2780	CTP	[109]
Pyrimidine			
Biosynthesis			
Histidine	b1207	5-Phospho-alpha-D-ribose 1-diphosphate; alpha-D-Ribose 5-phosphate	[109]
Methionine	b2942	S-Adenosyl-L-methionine	[109]
Metabolism			
Folate	b0529	5,10-Methenyltetrahydrofolate	[109]

Metabolism			
Alternate	b4084	D-Allose	[178]
Carbon			
Metabolism			
Alternate	b3608	Dihydroxyacetone phosphate;Glycerol 3-phosphate	[109]
Carbon			
Metabolism			
Membrane Lipid	b0185,b2316,b3255,b3256	Malonyl-CoA	[109]
Metabolism			
Unassigned	b0657	lipoprotein	[179]

Previous analysis suggested that although the majority of non-essential genes cannot induce the functional loss of other essential genes, there were still 103 N->E pairs involving 58 non-essential genes inconsistent with our findings. A thorough investigation revealed that the majority of these inconsistent pairs involve at least one essentiality disputable gene compared with the PEC (Profiling of *E. coli* Chromosome) database [94]. For example, *entD* (b0583), an essential gene in the Keio collection, was identified as non-essential in the PEC database (**Table 10**). Accordingly, we removed these inconsistent genes, leaving 45 N->E pairs with 28 non-essential genes and 29 essential genes participated. Among them, 3 non-essential genes showed their ability to affect those essential genes that do not belong to any associated gene sets. Another 5 non-essential gene can induce the functional loss of those essential genes that were part of certain associated gene sets. For the leftover 20 non-essential genes, 7 of them were essential for growth on glycerol minimal medium [92]. These findings favored our proposed concept of functional core module and its relation to cell survival.

Table 10 Comparison between Keio Collection and PEC database. There are 264 common genes between these two databases (not shown in this table). The left side is genes specific to Keio collection, where the right side is genes specific to PEC database.

Keio Collection				PEC			
b0057	b1175	b2783	b3595	b0026	b2496	b3176	b2651
b4503	b1356	b2941	b3623	b0103	b2695	b3198	b3935
b0270	b1570	b3021	b3709	b0416	b2697	b3201	b3974
b0583	b1572	b3113	b3793	b0455	b2891	b3559	b3976
b0733	b1689	b3146	b3834	b0536	b3019	b3761	b3978
b0886	b2017	b3159	b3835	b0672	b3058	b3783	b4143
b1131	b2559	b3191	b3843	b0971	b3066	b3796	b4147
b1145	b2567	b3471	b4084	b1133	b3067	b3797	b4161
b1174	b2573	b3532	b4224	b1909	b3123	b3799	b4362
				b1910	b3174		

3.3.6 Structural organization of essential and non-essential genes in metabolic network

The functional properties can always be related to the structural properties. We next studied how the structural organizations of essential and non-essential genes within the metabolic network contribute to essentiality. Two genes were considered linked together via certain metabolite as long as one gene's corresponding reactions consume this particular metabolite whereas the other's corresponding reactions produce it. According to essentiality, all these pairs can be classified into three groups, adjacent essential-essential gene pairs (EE), adjacent essential-non-essential gene pairs (EN), and adjacent non-essential gene pairs (NN). Next, we examined the degree features associated with these metabolites. Similarly, metabolites can be distinguished as: uniquely generated (UG) (in-degree equals to 1), uniquely consumed (UC) (out-degree equals to 1), and branched nodes (BN) (with in-degree and out-degree both bigger than 1) (**Figure 15**). The former two (i.e. UG and UC nodes) were combined together since many metabolites were both uniquely consumed and uniquely generated. Therefore, for each group (adjacent EE, adjacent EN, and adjacent NN) we calculated the percentage of gene pairs linked via UG/UC (either UG or UC) and BN nodes. Our results indicated that more than 45% of gene pairs were associated via UG/UC metabolite nodes in adjacent EE group, which was quite distinct from 8.3% in adjacent EN group and 12.3% in NN group.

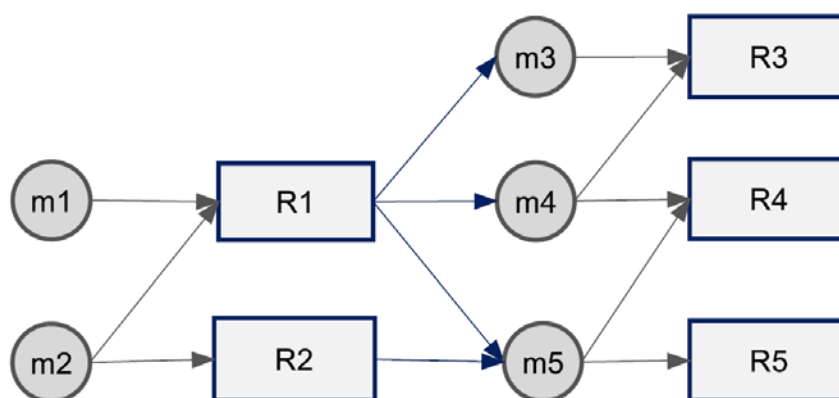


Figure 15 Different types of metabolites. In this figure, circle node represents metabolite, where rectangle node represents reaction. Metabolite m3 is an example of one in-degree and one out-degree. Metabolite m4 is an instance of one in-degree and multi out-degree. Metabolite m5 is an example of multi in-degree and multi out-degree. Metabolites m1 and m2 are external metabolites, which are the input of the whole network.

Besides, such structural organization observed in the metabolic network was quite distinct from null model. The null model was constructed by switching pairs of randomly selected edges. 1,000 random networks were generated, with each switched 2 times the total number of edges (details refer to **Section 3.2.2**). For the adjacent EE group, we found that the percentage of UG/UC intermediate metabolites in the real network was significantly larger than that obtained for random networks. On the other hand, for the adjacent EN group, it was much less in the real network than in the null model. Furthermore, the average percentage of UG/UC nodes in these three groups (adjacent EE, adjacent EN, and adjacent NN) in the null model was 25.5%, 18.2%, and 13.3% respectively, suggesting the distinct organizations of the adjacent EE and adjacent EN groups in the real metabolic networks (**Figure 16**).

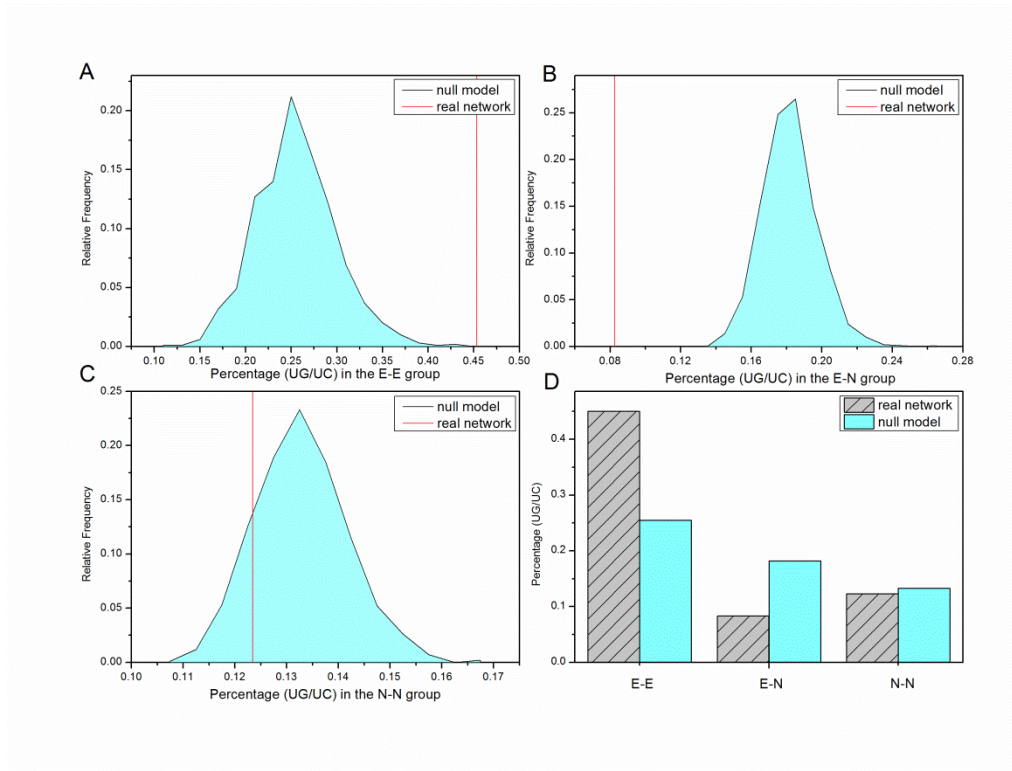


Figure 16 Structural organizations of essential and non-essential genes. (A) Distribution of the percentage of UG/UC metabolites linking the adjacent EE gene pairs (curve) in the 1,000 null models compared with in the real network (vertical line). (B) Distribution of the percentage of UG/UC metabolites linking the adjacent EN gene pairs (curve) in the 1,000 null models compared with real network (vertical line). (C) Distribution of the percentage of UG/UC metabolites linking the adjacent NN gene pairs (curve) in the 1,000 null models compared with real network (vertical line). (D) The percentage of UG/UC metabolites linking adjacent EE, EN, and NN gene pairs in *E. coli* metabolic network and null models, where bar with oblique lines represents metabolic network and bar without oblique lines represents the average percentage in the null model.

Since essential genes were prone to be interconnected through UG/UC nodes, it is highly possible that removal of any one can easily spread its effects to the counterparts. On the other hand, as adjacent EN pairs were bridged via branched nodes, loss of one gene may not impose its effect directly on the other since

alternative pathway can compensate the functional loss. Thus, the topological organization of essential genes and non-essential genes provides a structural and mechanistic basis for the previously observed results.

3.4 Discussions

Essential genes have been widely studied considering their significance role in antimicrobial drug development. Mechanistic studies on gene dispensability indicate that there are potentially three kinds of explanations: (1) conditional essential genes; (2) genetic buffering by duplicate genes or gene with overlapping functions and (3) alternative pathway or redistribution of fluxes. The mechanisms are still controversial and no consensus has been reached. In this study, we will study from the perspective of functional back-up mechanisms, i.e. the latter two explanations. It is intriguing to know how the deletion of essential genes can be buffered while non-essential genes cannot. Current knowledge about essentiality cannot offer an explanation why some genes are much more significant than others although they all can be participated in some crucial biological processes. In our studies, we attempted to unravel the properties associated with essentiality in the context of metabolic networks from the perspective of functional and structural studies to show how these properties contribute to gene essentiality.

In typical genetic studies, one gene is considered indispensable if knockout can lead to a lethal phenotype. However, single gene knockout disturbs not only that

particular gene function, but also a wide range of other genes, as suggested by gene expression profiles [180-182]. Furthermore, researchers argued that gene interaction was always neglected in the studies of gene essentiality [183]. Therefore, we modified and repurposed a previous algorithm to capture the gene knockout effects on the whole network using a set of genes whose function may be interrupted in response to the single gene deletion, which was totally determined by the metabolic network structure and the corresponding reactions.

Damage list is a collective of genes whose functions may be impaired in response to single gene deletion and it is identified via initiating the cascading failure procedures. Integrating essentiality information into the damage lists revealed some remarkable differences between essential and non-essential genes lie in their impact on other essential genes. While knockout of essential genes can cause a large range of other essential gene function failure, knockout of non-essential genes can only interrupt other non-essential genes' functions. This suggests that essential and non-essential genes propagated their deletion effects via distinct routes. Phenotypic differences observed between essential and non-essential gene knockout can be explained by the cooperative effects of essential genes. Such observed interactions were consistent with the previous epistasis studies, in which essential genes were found prevalingly positive epistatic [130] and interaction between essential genes were much more intense compared with non-essential genes [126, 128]. In addition, function of essential genes can be buffered by both non-essential genes and other essential genes [126, 128, 156, 184, 185].

Our studies also revealed that genes with highly similar damage lists tend to have the same essentiality and participate in the same or related pathways. When genes with similarity coefficient larger than a threshold were correlated, genes with the same essentiality were modularly organized, forming many small subnetworks. These findings were consistent with the previous studies that for both essential and non-essential genes, their synthetic interactions were highly biased towards genes that shared related functions [184]. Particularly, functional related essential genes had a strong tendency to show similar spectrum of interactions [184].

A further investigation revealed that subnetworks consisting of essential genes generally have an ‘associated gene set’, which can disrupt the production or consumption of certain essential components. Failure in the production or consumption of these indispensable components eventually resulted in lethal phenotype. Accordingly, we proposed the existence of associated gene sets, which were vital for cell survival. We identified 72 associated gene sets where mapping these sets to pathway we found their involvement in crucial pathways or metabolites. Our identified associated gene sets thus reflected the way how a gene exerted its impact over the whole cell, suggesting a possible mechanism relating the internal changes following gene knockout.

Structural analysis revealed distinct organizational principles of essential and non-essential genes. Adjacent EE pairs were overwhelmingly linked through UG/UC metabolites, whereas adjacent EN pairs were rare. Our findings offered an

explanation for the observed discrepancies underlying these two types of damage lists. The UG/UC metabolites were relatively fragile and perturbation on any gene can easily spread the damaging to the counterparts for lacking alternative pathways. Therefore, perturbation in adjacent EE pairs can easily spread to their respective counterparts. On the contrary, two genes connected via branched metabolites were irrelevant in their damage effects as alternative pathways can compensate the functional loss. Synthetic lethal pairs can be used to further validate our findings in these structural principles. Double knockout, especially those genes within the same pathway [128], may lead to the consequences that some gene pairs once linked through branched metabolites were replaced by UG/UC nodes, with increasing vulnerability. In brief, structural organization of essential and non-essential genes dictated their distinct impacts on the whole networks where perturbing the associated gene sets lead to lethal phenotypes.

Hence, according to our results, we proposed that essential and non-essential genes spread their deletion effects via different routes (**Figure 17**). Following genetic perturbations, essential gene can spread its deletion effects to other functional related essential genes. On the other hand, non-essential genes seldom affected other essential genes. The associated gene sets, which involved in the production or consumption of the key metabolites required for cell survival, were affected in response to essential gene knockout. Genes with highly similar damage lists tend to share the same essentiality, since they may function via similar mechanisms. Our proposed scheme relating the internal changes following genetic

perturbations were supported by the structural organization of essential and non-essential genes.

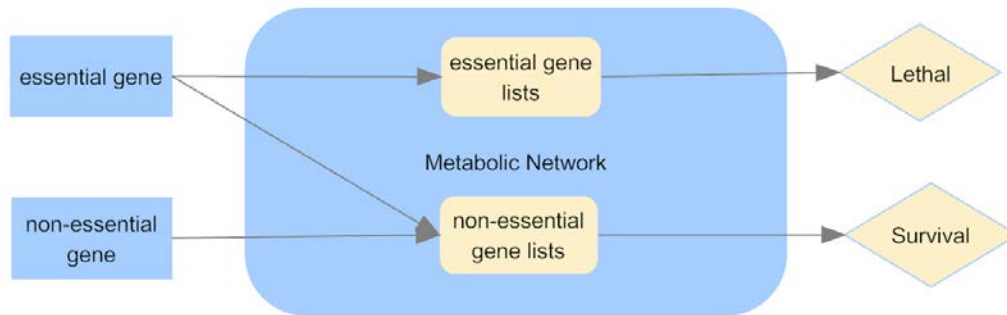


Figure 17 Proposed mechanisms for the differential single gene deletion effects. For essential genes, they can affect other essential genes, where the cooperative effects can lead to lethal phenotypes. For non-essential genes, they can only affect other non-essential genes.

Chapter 4 Essential gene prediction

In this chapter, we discussed how to predict essential genes based on previous findings in the context of metabolic network. A variety of computational based methods have been implemented to predict essential genes from the scope of biological networks and most of them require network with high quality, or a large number of network topological features that associated with essentiality in order to have a high prediction accuracy. In this study, we aimed to propose an approach that can predict essentiality with high accuracy while using limited network information.

4.1 Overview of essential gene predictions

Essential genes are usually defined as those genes necessary for cell growth in a rich medium, i.e. medium containing all nutrients required for growth. Deletion of any essential gene is sufficient to confer a lethal phenotype even in the presence of all the other genes. It is widely believed these genes involved in crucial cellular processes whose impairment can disrupt normal cellular growth [186, 187]. Identification of such genes is rather significant, not only for understanding of the minimal requirement of cellular life [188], but also for antimicrobial drug development [189].

Experimental approaches such as transposon mutagenesis have been widely used in determining essential genes across a variety of species, generating large information [104, 190-192]. For example, a systematic transposon mutagenesis experiment is conducted in *E. coli* K-12 to replace the open reading frame of each target gene with a non-functional fragment, obtaining 3985 mutant strains out of 4288 genes targeted [66]. The leftover genes that cannot get mutant strains are considered as essential genes. Similarly, in yeast *S. cerevisiae*, a systematic gene disruption experiment produced a collection of gene-deletion mutants for 96% of annotated open reading frames [88]. Although these methods have accumulated abundant resources concerning essentiality, the procedures are rather time consuming and resource intensive since we need to construct the mutant strains for each individual gene. Besides, such approach is not feasible for all organisms, especially those infectious microorganisms. Experimental results for the same species also vary, sometimes to a large extent. For example, for *E. coli*, the Keio collection [66] and PEC database [94] share 264 common essential genes, where 36 genes are specific to Keio collection while 38 genes specific to PEC database (**Table 10**).

The development of high-throughput technologies and whole genome sequencing make it possible to reconstruct large-scale computational models. The accurate metabolic networks of some model organisms have been reconstructed [17, 138]. Computational approaches such as Flux Balance Analysis (FBA) [95], and Minimization of Metabolic Adjustment (MOMA) [96] have been developed and

widely used to assess the essentiality of genes *in silico*. For example, FBA is a constraint-based approach used to analyze whether the growth rate is interrupted in the absence of genes (for details, please refer to **Section 2.3, Chapter2**). Although the results are of high confidential, it requires a clear definition of the metabolic network, nutrient availability and biomass components.

Another category of computational approaches are based on machine learning [14, 160, 193-195], which utilize a variety of network based topological features, sequence characteristics that potentially associated with essentiality to predict gene's essentiality. However, one common problem regarding this kind of approaches is that they employed a large number of features but in most cases these features didn't show clear cause and effect relations with gene essentiality. For example, some basic network features including degree, closeness, betweenness centrality, and clustering coefficient are considered potentially associated with essential genes. Yet, no observations or researches indicate that these basic network features are the cause of gene essentiality. Therefore, the exploration of some essentiality related features may be crucial, able to increase the prediction accuracy.

In the previous chapter, we discussed why some genes are more important than others in details, from the perspective of both functional and structural organization in metabolic network. It is suggested that essential and non-essential

genes exhibit different deletion effects in the context of metabolic network. While essential genes can spread its deletion effects to both essential and non-essential genes, nearly all the non-essential genes can only affect themselves or other non-essential genes. In this chapter, we used the previous findings as a start point to explore some biological network specific features that significantly associated with essentiality and then use them to predict essential genes in *E. coli* and other species. Some functional and structural related features turned out to have strong association with essentiality and hence we incorporated them into our approach. Given a small set of gene's essentiality information, our approach can predict the majority of the leftover genes' essentiality type based on the highly correlated topological features. Our approach showed a high level of prediction accuracy compared with other computational approaches. This is validated by yeast *S. Cerevisiae*, indicating the robustness across species, which also suggests its potential application in essential gene identification.

4.2 Materials and Methods

4.2.1 Identification of functional and network topological features associated with gene essentiality

In the previous chapter, we investigated why some genes are more important than others in the context of metabolic networks. Some functional and network topological properties are revealed to be key factors in discriminating essential

and non-essential genes. Therefore, we inherited from the previous chapter to explore features strongly correlated with essential genes.

First of all, we examined the relation between damage size and essentiality. It is noticed that although the size of the damage list cannot be a determinant feature in discriminating essential and non-essential genes, essential genes generally have a larger damage size. Besides, it is found out that among the 1261 genes 611 of them have a damage size of zero, where 595 (97.38%) of them are non-essential genes. Therefore, the gene damage list with size zero is considered as a promising feature in classifying essential and non-essential genes.

Besides the damage size, the damage list similarity coefficient is also a good indicator. It is noticed that when we set a threshold to link gene pairs with similarity coefficient higher than the threshold, many subnetworks are formed. Interestingly, genes within each subnetwork nearly all share the same essentiality type (**Table 11**). For example, when the threshold is set to 0.9, around 95% gene pairs linked together have the same essentiality type. It is suggested that if one gene's essentiality type is known, we can estimate that other genes from the same subnetwork should have the same essentiality type.

Table 11 The number of gene pairs with similarity coefficient at different level threshold. For different threshold, total gene pair refers to the number of gene pairs obtained if we link two genes with their damage list similarity coefficient bigger than the threshold, whereas rate refers to the percentage of gene pairs sharing the same essentiality, and genes refer to the number of genes involved.

threshold	total gene pairs	rate	genes
0.1	2584	0.8	457
0.2	1764	0.82	438
0.3	1350	0.85	406
0.4	1118	0.87	368
0.5	1024	0.88	363
0.6	826	0.88	280
0.7	672	0.92	245
0.8	546	0.95	235
0.9	476	0.95	219

4.2.2 Self-devised algorithm for predicting gene essentiality

Here is our proposed method used to predict gene essentiality (**Figure 18**). Starting from metabolic network, we randomly chose certain number of genes and labeled them as known-type genes, where the essentiality information of these genes is inherited from Keio collection [66], which can be either be essential or non-essential genes. The leftover genes are labeled as unknown-type genes, where the essentiality type for these genes is to be predicted from the known-type genes.

First of all, we calculated the damage lists for each gene in the metabolic network as we described in the previous chapter. If the size of the damage list is zero, we

assumed the corresponding gene to be non-essential gene. Otherwise, we calculated the Jaccard coefficient between the damage lists of the unknown-type genes and known-type genes. For each unknown gene, we picked out 5 genes from the known-type genes which have the highest damage list similarity coefficients with this query gene, where the similarity coefficient is also required to meet certain threshold (e.g. similarity coefficient > 0.4). If the number of genes satisfying this criterion is less than 5, we just keep them without further modifications. For these selected top 5 or less genes, as long as any of them is essential, we consider this unknown query gene to be essential, otherwise non-essential or unknown (the case that we cannot find any known genes with similarity bigger than the threshold for the query gene). Such procedures are iterated until either all the unknown genes are labeled with an essentiality type, or a pre-defined iteration maximal number is reached. In the end, a parameter called prediction accuracy is introduced to measure the efficiency of our approach, which is defined as the number of corrected predicted genes over the total number of genes to be predicted ($N(\text{correct_predicted}) / N(\text{total_to_predicted})$). Since some genes couldn't be predicted through our approach as they are not associated with the features we adopted in this study, it is not appropriate to include them when computing the prediction accuracy. As a consequence, we also defined another parameter, called modified prediction accuracy, which is defined as the number of corrected predicted genes over the total number of predicted genes ($N(\text{correct_predicted}) / N(\text{total_predicted})$).

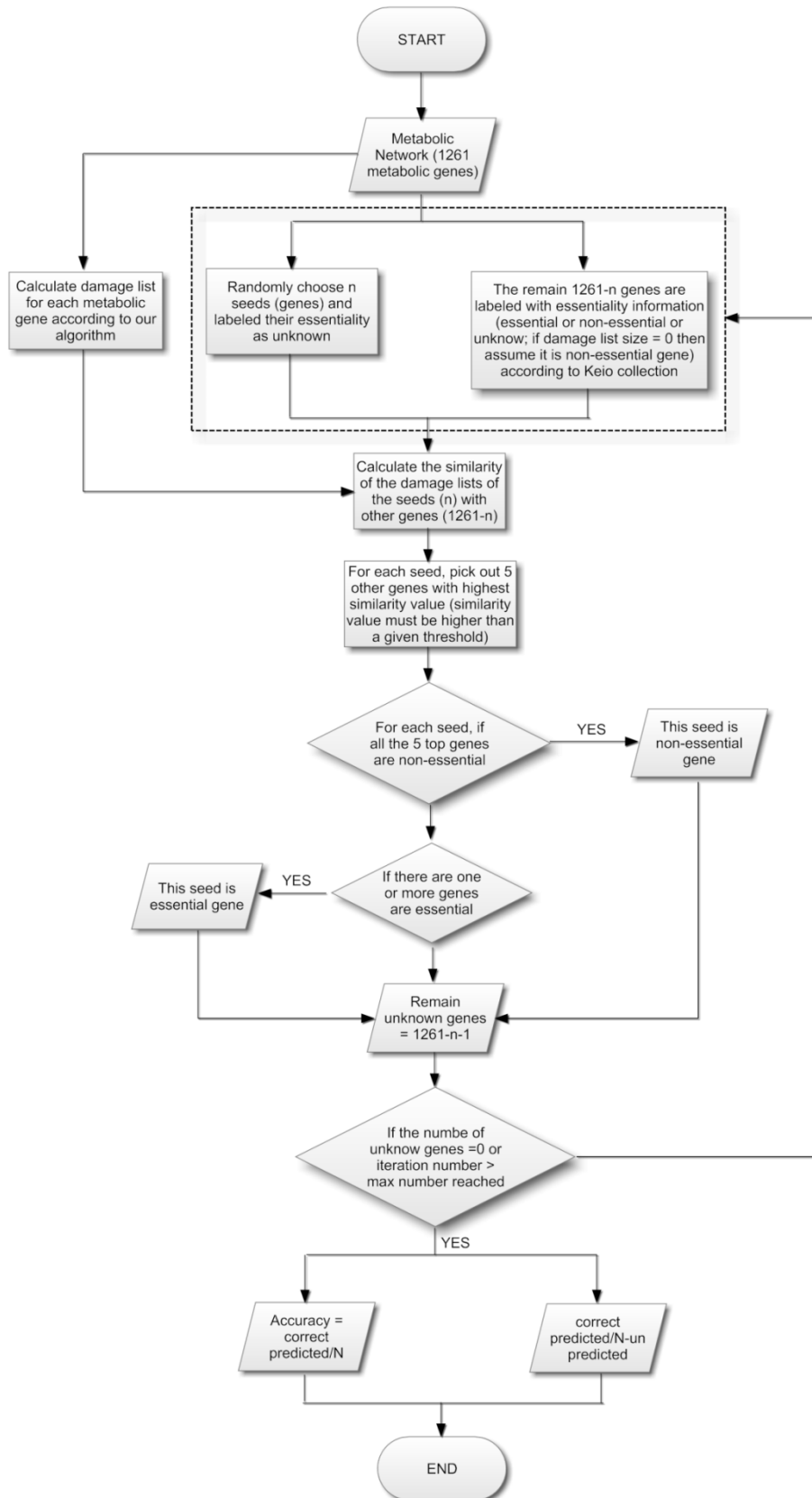


Figure 18 Overview of the algorithm. Starting from *E. coli* metabolic network, in each computational cycle, N genes are randomly chosen, which are labeled as unknown-type genes, where the essentiality is to be predicted. The leftover genes are labeled as known-type genes, where the essentiality information is inherited from Keio collection. The damage list for each metabolic gene is calculated. For the unknown gene, if the size of the damage list is zero, we updated its essentiality to be non-essential. Otherwise, its damage list is compared with other known-type genes and a similarity coefficient is computed to measure their damage list similarity. The top five similar genes (if exists) are selected for each unknown-type gene. The essentiality of the unknown-type gene is updated to be essential, non-essential, or unknown depending on the components of the selected top five similar genes. Details are described in the Method part. Such procedures are iterated until all the unknown genes are identified, or a pre-defined maximal iteration number is reached. Finally the prediction accuracy is computed as we discussed in the method part. This is one computational cycle. Such cycle is repeated using different unknown gene sets. An average is taken over all these cycles to get averaged prediction accuracy.

4.3 Gene essentiality prediction

4.3.1 Results of gene essentiality prediction

Before we made any predictions, we initialized single gene deletion according to the procedures detailed in the previous chapter for each gene in the reconstructed *E. coli* bipartite metabolic network. For each query gene, we can obtain a corresponding damage list, which is composed of genes whose functions may be impaired in response to the target gene deletion.

In our study, all the genes are classified as unknown-type and known-type, where unknown-type refers to genes with unknown essentiality type and known-type refers to genes with known essentiality type. The details regarding how to predict these unknown-type genes from the known-type genes are illustrated in the previous section. The procedure from randomly choosing unknown-type genes to the prediction of these genes is called one calculation cycle. Such procedure is iterated multiple times and average was taken over all these cycles to obtain averaged prediction accuracy.

In our method, there are some parameters involved, such as the number of unknown-type genes, pre-defined maximal iteration number, similarity coefficient threshold, cycle number and so on. The impact of each parameter is tested in the following sections and an optimal value for each parameter is also determined.

Besides the prediction accuracy, sensitivity and specificity are also calculated for each scenario, where the sensitivity is a measure of the ability to identify positive results and specificity is a measure of the ability to identify negative results. They are defined as:

$$\text{sensitivity} = \frac{\text{number_true_positive}}{\text{number_true_positive} + \text{number_false_negative}}$$

$$\text{specificity} = \frac{\text{number_true_negative}}{\text{number_true_negative} + \text{number_false_positive}}$$

Here, true (or false) positive refers to essential genes (or non-essential genes) correctly (or incorrectly) identified as essential, whereas true (or false) negative refers to non-essential genes (or essential genes) correctly (or incorrectly) identified as non-essential.

4.3.2 The effect of iteration number on prediction accuracy

In each calculation cycle, we iterated our prediction procedures until all the unknown genes are predicted, or a pre-defined maximal iteration number is reached. Based on our observation, it is always the iteration number reached first. The iteration number only impact on whether the iteration balance is reached. If no other gene can be predicted based on available knowledge, increment of the iteration number cannot further increase the prediction accuracy. A smaller iteration number sometimes may lead to lower prediction accuracy. Changing the iteration number while keeping other parameter the same, we found out that the

modified prediction accuracy stays robustness, fluctuates around 0.95. Besides, the number of non-predicted genes also shows a high level of consistence. Therefore, this parameter is considered not have much impact on the prediction accuracy. In the following sections, we use 500 as the maximal iteration number (**Table 12**).

Table 12 The effect of iteration number on prediction accuracy. For different maximal iteration number (Iteration), prediction accuracy (prediction acc.), modified prediction accuracy (modified prediction acc.), sensitivity, specificity, and the number of unpredicted genes (unpredicted) are computed.

Iteration	prediction acc.	modified prediction acc.	sensitivity	specificity	unpredicted
500	0.58244	0.95273	0.50792	0.98446	388
1000	0.58714	0.95314	0.47018	0.98563	383
1500	0.58778	0.95396	0.50166	0.9857	383
2000	0.58512	0.95345	0.47562	0.98705	386
2500	0.58466	0.95231	0.48612	0.98412	386
3000	0.58396	0.95365	0.47223	0.98682	387

4.3.3 The effect of threshold on prediction accuracy

Threshold is another important parameter used in our study. As mentioned, gene pairs with high similarity coefficients tend to share the same essentiality. Thus, a moderate threshold is required since it may guarantee the correlation between similarity coefficient and gene essentiality. The cost of a high threshold is the number of unpredicted genes since fewer genes are involved when increasing the

threshold. On the other hand, a relative low threshold may introduce some false positive (or false negative) predictions. As a result, we need to balance the tradeoff between prediction accuracy and the number of unpredicted genes. To find an optimal threshold, we studied two scenarios, one with 800 unknown genes (**Table 13**) and the other with 100 unknown genes (**Table 14**). The leftover genes for each individual scenario are known-type genes, whose information is used to predict the unknown genes. In the first scenario, the overall prediction accuracy goes up when the threshold goes up. So does the overall unpredicted genes. Similar observations can be found in the second scenario. Considering the tradeoff between sensitivity and specificity, we choose the threshold 0.4, because it may correspond to a moderate sensitivity and specificity compared to other threshold.

Table 13 The effect of threshold on the prediction accuracy (800 unknown). For each threshold, prediction accuracy (prediction acc.), modified prediction accuracy (modified prediction acc.), sensitivity, specificity, and the number of unpredicted genes are computed (unpredicted gene). The maximal iteration number, number of top genes, cycle number, and the number of unknown-type genes are: 2000, 5, 50, and 800 respectively.

Threshold	prediction acc.	modified prediction acc.	sensitivity	specificity	unpredicted gene
0.1	0.710	0.911	0.614	0.940	176
0.2	0.696	0.926	0.615	0.957	198
0.3	0.665	0.936	0.586	0.968	231
0.4	0.635	0.948	0.569	0.980	264
0.5	0.622	0.951	0.529	0.985	277
0.6	0.590	0.958	0.521	0.989	307
0.7	0.573	0.963	0.476	0.991	324
0.8	0.566	0.965	0.445	0.993	330
0.9	0.557	0.966	0.390	0.993	338

Table 14 The effect of threshold on the prediction accuracy (100 unknown). For each threshold, prediction accuracy (prediction acc.), modified prediction accuracy (modified prediction acc.), sensitivity, specificity, and the number of unpredicted genes are computed (unpredicted gene). The maximal iteration number, number of top genes, cycle number, and the number of unknown-type genes are: 2000, 5, 50, and 100 respectively.

Threshold	prediction acc.	modified prediction acc.	sensitivity	specificity	unpredicted genes
0.1	0.766	0.908	0.665	0.935	15
0.2	0.757	0.920	0.683	0.944	17
0.3	0.739	0.929	0.677	0.953	20
0.4	0.730	0.943	0.703	0.965	22
0.5	0.717	0.944	0.679	0.969	24
0.6	0.668	0.954	0.644	0.981	29
0.7	0.644	0.958	0.649	0.984	32
0.8	0.639	0.953	0.570	0.986	32
0.9	0.617	0.961	0.558	0.988	35

4.3.4 The effect of top N similar genes on prediction accuracy

Although we set a threshold for the damage list similarity coefficient to identify those correlated gene pairs, occasionally the number of qualified genes is too many. Therefore, we added a second filter, the top N similar genes, which are defined as those top N genes in the list of genes satisfying the first threshold criterion sorted according to the similarity coefficient. For each query unknown gene, we first identified those genes have a similarity coefficient bigger than the defined threshold, and then selected the top N genes. We examined the essentiality type of these N genes. If any one of them is essential, we predicted that this query gene is also an essential gene, otherwise non-essential. If the total number of genes satisfying this requirement is less than N, we used all these genes. By changing the value of N, we found out this parameter may affect our prediction results (**Table 15**).

Table 15 The effect of Top N similar genes on the prediction accuracy. For different N, predicted accuracy (prediction acc.), modified prediction accuracy (modified prediction acc.), sensitivity, specificity, and the number of unpredicted genes (unpredicted gene) are computed. The maximal iteration number, threshold, cycle number, and the number of unknown-type genes are: 2000, 0.4, 50, and 100 respectively.

N	prediction acc.	modified prediction acc.	sensitivity	specificity	unpredicted gene
2	0.727	0.941	0.647	0.973	22
5	0.722	0.941	0.661	0.97	23
8	0.713	0.937	0.669	0.963	23
10	0.7176	0.933	0.63	0.965	23
15	0.702	0.925	0.674	0.948	24
20	0.7064	0.923	0.681	0.948	23

Our results indicated that the prediction accuracy is quite sensitive with respect to the number of similar genes selected. A lower number of similar genes (e.g. 2) always correspond to high prediction accuracy and a relative low sensitivity, whereas a higher number of similar genes (e.g. 20) always associated with low prediction accuracy and high sensitivity. In our work, we chose 5 as our optimal parameter because at this level, we not only have high prediction accuracy, but also reasonable sensitivity and specificity.

4.3.5 The effect of cycle number on prediction accuracy

In each simulation cycle, we randomly selected one set of unknown genes. To reduce the effects of such randomness, we repeated the simulation cycle multiple times, which is termed as cycle number. The prediction accuracy for each unknown gene set is computed and then averaged. It is observed that the effect of this parameter is quite similar to the iteration number. As long as this number is large enough, the obtained result is robust. For the cycle number ranging from 10 to 80, we didn't observe any significant difference between them. In our work, we chose 50 for this parameter (**Table 16**).

Table 16 The effect of cycle number on prediction accuracy. For different cycle numbers, prediction accuracy (prediction acc.) and modified prediction accuracy (modified prediction acc.) are computed. The maximal iteration number, threshold, similar genes, and the number of unknown-type genes are: 2000, 0.4, 5 and 1000 respectively.

cycle	threshold	top	unknown	iteration	prediction	modified
		gene	gene		acc.	prediction acc.
10	0.4	5	1000	2000	0.576	0.951
20	0.4	5	1000	2000	0.58545	0.951
30	0.4	5	1000	2000	0.5865	0.955
40	0.4	5	1000	2000	0.58538	0.953
50	0.4	5	1000	2000	0.58778	0.954
80	0.4	5	1000	2000	0.5851	0.953

4.3.6 Prediction accuracy for different number of unknown genes

Using all the optimized parameters we identified in the previous section, we performed simulations for different number of unknown genes, varying from 100 to 1200. The summarized results are listed in the following table (**Table 17**).

We found out that the modified prediction accuracy can be up to 0.97 and is rather robustness, which is even higher than FBA approaches, which is around 0.87 (**Figure 19**). However, there are a number of genes that cannot be predicted using our approach, which may due to the incompleteness and the quality of the reconstructed metabolic network. The ratio of unpredicted genes also increases associated with increasing number of unknown genes. So if we take these genes

into account and considered them as misclassified, the prediction accuracy may drop to around 0.7 (when the number of unknown genes is less than 600), which is still comparable with some machine learning based prediction methods. One limit is that when the number of genes to be predicted is too many (e.g. 1200), our prediction accuracy (without modified) is relative low (e.g. 0.505) as there is a large number of genes that cannot be predicted (e.g. 573). However, those predicted are with high accuracy (e.g. 0.967).

Table 17 Summary of the prediction results for *E. coli*. For simulations with different number of unknown genes (ranging from 50 to 1200), prediction accuracy (prediction acc.), modified prediction accuracy (modified prediction acc.), sensitivity, specificity, the number of unpredicted genes (unpredicted), and the ratio of unpredicted genes (unpredicted ratio) are computed.

unknown genes	prediction acc.	modified prediction acc.	sensitivity	specificity	unpredicted	unpredicted ratio
50	0.733	0.943	0.646	0.971	11	0.220
100	0.722	0.941	0.657	0.970	23	0.200
150	0.707	0.942	0.654	0.971	37	0.280
200	0.718	0.944	0.642	0.975	47	0.220
250	0.708	0.945	0.647	0.972	62	0.240
300	0.698	0.946	0.655	0.973	78	0.257
350	0.702	0.944	0.633	0.974	89	0.233
400	0.696	0.944	0.639	0.972	105	0.217
450	0.691	0.945	0.635	0.974	120	0.277
500	0.684	0.946	0.639	0.975	138	0.267
550	0.676	0.945	0.605	0.976	156	0.285
600	0.666	0.945	0.618	0.974	177	0.273
650	0.662	0.946	0.599	0.976	195	0.284
700	0.653	0.947	0.585	0.978	217	0.269
750	0.645	0.949	0.585	0.979	240	0.282
800	0.629	0.949	0.564	0.980	269	0.324
850	0.627	0.952	0.546	0.983	289	0.355
900	0.613	0.952	0.540	0.983	319	0.351

950	0.592	0.952	0.525	0.984	359	0.340
1000	0.582	0.953	0.507	0.984	388	0.324
1050	0.567	0.957	0.453	0.988	427	0.407
1100	0.552	0.958	0.384	0.990	466	0.414
1150	0.525	0.961	0.291	0.992	521	0.407
1200	0.505	0.967	0.248	0.996	573	0.420
1250	0.478	0.972	0.051	0.999	635	0.406

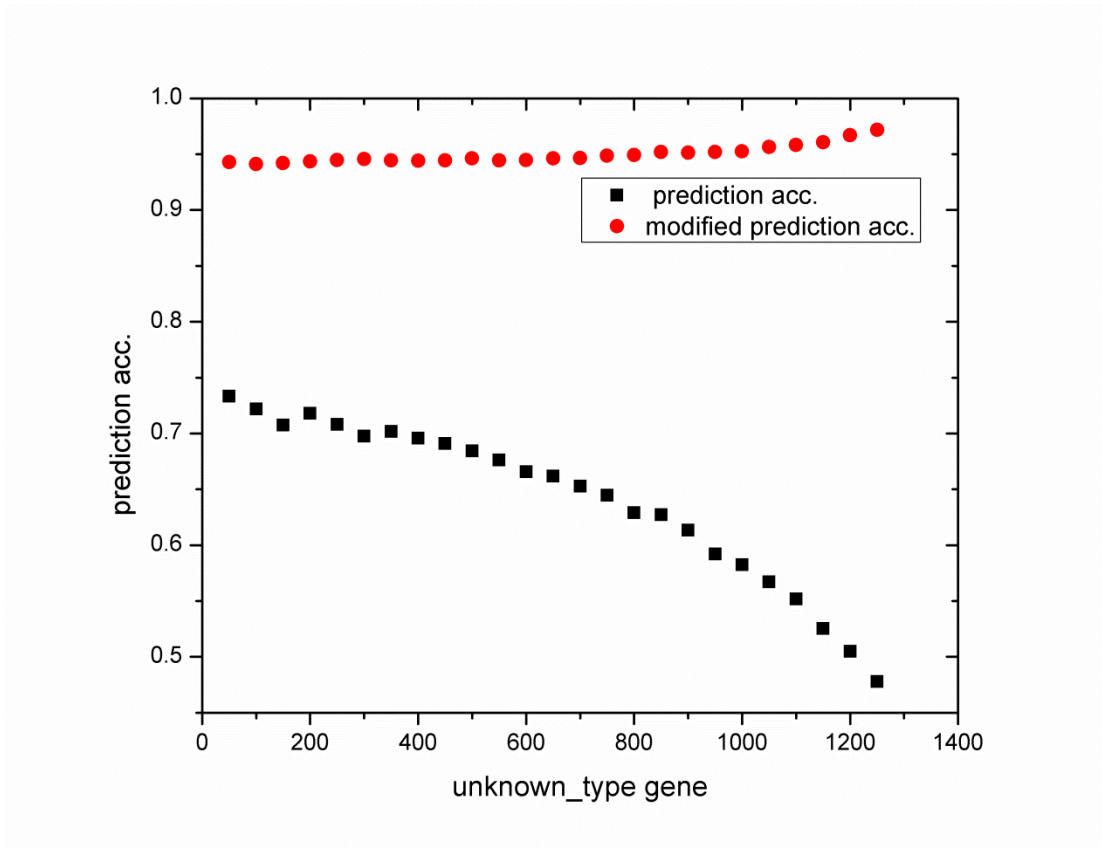


Figure 19 Prediction results for different number of unknown genes. In this figure, the horizon is the number of unknown type genes to be predicted in our study, and the vertical is the prediction accuracy, where dot is the modified prediction accuracy and rectangle is the prediction accuracy.

4.3.7 Strategies to increase the prediction accuracy

4.3.7.1 Combination with FBA

Since some genes cannot be predicted using our approach, we proposed the idea of combining our approach with other available prediction methods, such as FBA.

In other words, we first predicted the essentiality type with our own method, and

for those unpredicted genes, we applied FBA method. The overall prediction accuracy is computed. It turned out that the combined method could maintain a high level of prediction accuracy (**Figure 20**). This is validated even in the presence of limited information (the number of known gene less than 100).

4.3.7.2 Increased number of known essential genes

In practice, researchers are more concerned about essential genes. Experimental results have accumulated abundant information regarding these genes [91, 92, 191, 196]. Besides, essential genes are considered more evolutionary conserved compared with non-essential genes in bacteria [112, 117]. Mapping from one species to another may also provide some information for essential genes. Therefore, it is reasonable to adequately increase the weight for essential genes in the known type genes. It is noticed that when the number of unknown essential genes decreases (corresponding to an increased number of known essential genes), the prediction accuracy increases. This is another strategy we can use to increase the prediction accuracy (**Figure 21**).

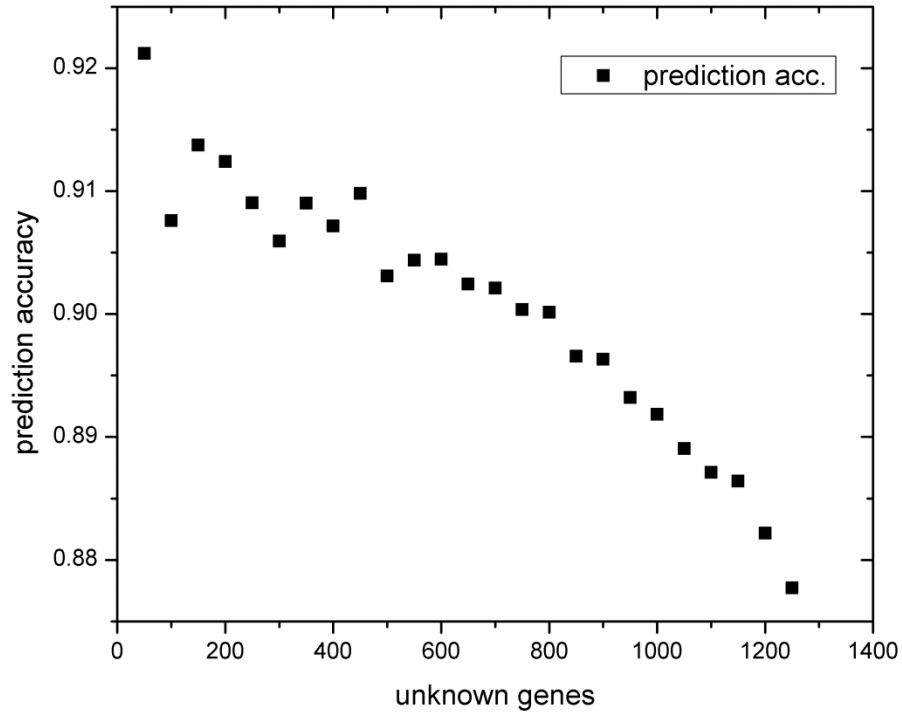


Figure 20 Prediction accuracy combined with FBA. In this figure, the horizon is the number of unknown genes to be predicted, whereas the vertical is the modified prediction accuracy. The starting point of the vertical is the prediction accuracy for FBA solely. It is evident that combination with FBA, our prediction accuracy can still maintain at a high level. The iteration number, cycle number, threshold, and the number of similar genes are: 500, 50, 0.4, and 5 respectively.

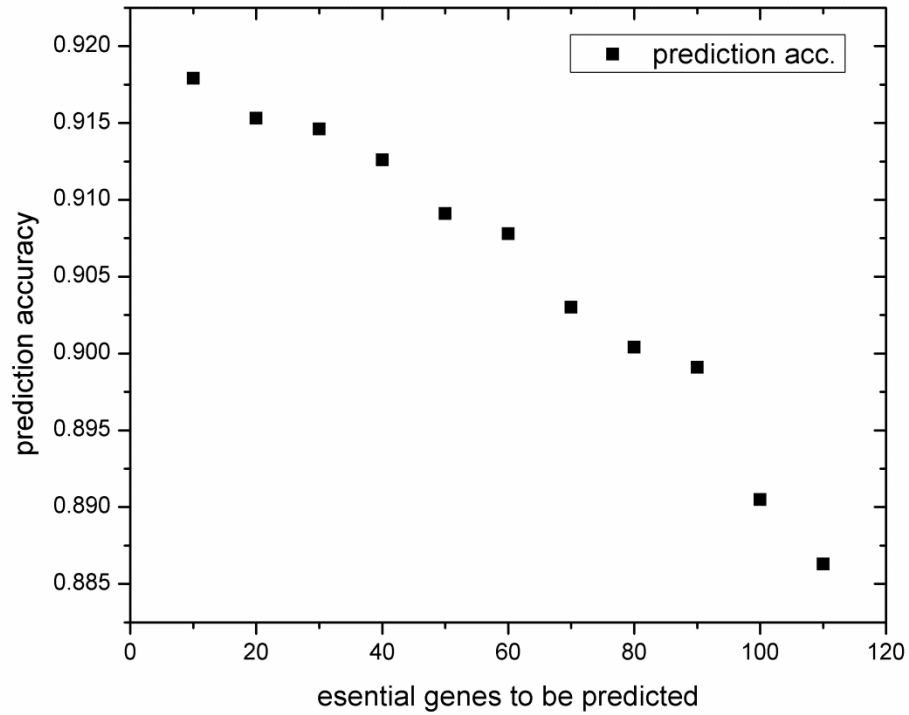


Figure 21 The prediction accuracy corresponding to different level of essential genes to be predicted. When we decrease the number of essential genes to be predicted (corresponding to an increased number of known essential genes), the overall prediction accuracy will increase. The iteration number, cycle number, threshold, the number of similar genes, and the number of unknown genes are: 500, 50, 0.4, 5 and 800 respectively.

4.3.8 Cross-species validation

Here we extended our work to another model organism, yeast *S. cerevisiae*. In yeast there are around 760 genes involved in the metabolic network. Using the optimized parameters we determined for *E. coli* as a reference, we obtained the prediction accuracy for yeast. Since the parameters used in *E. coli* may not be the optimized parameters for yeast, it is expected that the prediction accuracy for yeast should be higher than we actually obtained.

It is obvious that the prediction results are quite similar to those obtained for *E. coli*. When the unknown gene set is relative small, our prediction accuracy is high and the ratio of unpredicted genes is also small. However, when the number of unknown genes increase, sensitivity decreases and so does the prediction accuracy (**Table 18, Figure 22**). This may be explained by the fact that our unknown genes are generated randomly. Since the total number of essential genes is fewer compared with non-essential genes, a relatively small quantity of essential genes would be included as known genes. On the other hand, as we see in the previous results, some essential genes are only connected to 1 essential gene, or even zero. In this sense, it is not strange that our sensitivity is quite low when the unknown gene is large.

Table 18 Summary of the prediction results for yeast *S. cerevisiae*. For predictions with different number of unknown genes (ranging from 50 to 700), prediction accuracy (prediction acc.), modified prediction accuracy (modified prediction acc.), sensitivity, specificity, the number of unpredicted genes (unpredicted), and the ratio of unpredicted genes (unpredicted ratio) are computed. The iteration number, number of similar genes, threshold, and cycle number are: 2000, 5, 0.4, and 10 respectively.

unknown genes	prediction acc.	modified prediction acc.	sensitivity	specificity	unpredicted	unpredicted ratio
50	0.628	0.883	0.468	0.931	14	0.29
100	0.634	0.893	0.489	0.941	28	0.29
150	0.622	0.891	0.463	0.944	45	0.30
200	0.625	0.899	0.508	0.947	61	0.31
250	0.612	0.893	0.448	0.946	78	0.32
300	0.605	0.896	0.459	0.949	97	0.33
350	0.599	0.898	0.431	0.953	116	0.33
400	0.591	0.899	0.399	0.958	137	0.34
450	0.575	0.902	0.403	0.959	163	0.36
500	0.566	0.906	0.378	0.966	187	0.38
550	0.553	0.910	0.370	0.966	215	0.39
600	0.534	0.914	0.300	0.977	249	0.42
650	0.517	0.925	0.260	0.984	286	0.44
700	0.489	0.931	0.198	0.987	332	0.47

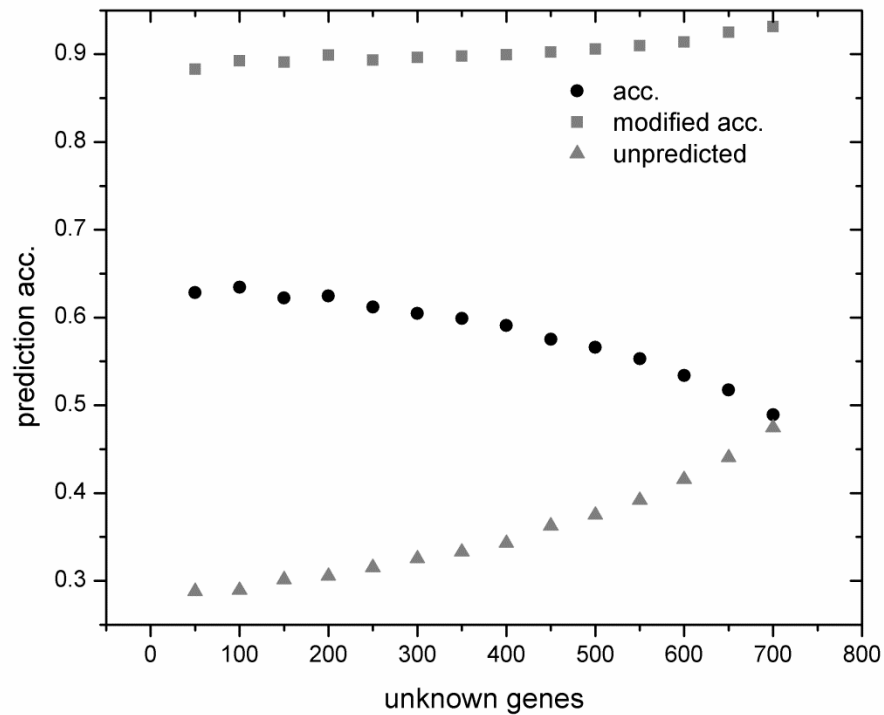


Figure 22 Prediction results for yeast *S. cerevisiae*. The horizon is the number of unknown genes to be predicted, and the vertical is prediction accuracy (circle), modified prediction accuracy (rectangle), or the ratio of unpredicted genes (triangle). The iteration number, number of similar genes, threshold, and cycle number are: 2000, 5, 0.4, and 10 respectively.

4.4 Discussions

Essential genes are considered as promising drug targets considering their ability to cause lethal phenotypes [189]. Two categories of computational based approaches are widely used to assess gene essentiality. One is constraint-based approaches, such as FBA [95] and its variants [96, 197], while the other is machine learning based approaches, which utilizing a variety of network topological features, sequencing characteristics that potentially associated with essentiality to predict gene essentiality [160, 193, 194, 198].

In **Chapter 3**, we revealed some functional and structural features that tightly associated with essential genes, which may confer to the differential deletion effects between essential and non-essential genes. Our studies indicated that essential genes generally have a larger damage list whereas majority of non-essential genes with zero damage list. Gene pairs with a higher damage list similarity coefficient tend to share the same essentiality. From the perspective of structural organization, essential genes tend to be interconnected through low-degree metabolites. Based on these findings, we derived some features that significantly associated with essential and non-essential genes, which are then used to predict unknown-type genes. Two features are used in our work. One is that genes with zero damage lists are significantly to be non-essential genes. Another is that gene pairs with a higher damage list similarity

coefficient tend to share the same essentiality type, which in other words, we can predict unknown genes' essentiality using known type genes based on their damage list profiles.

Since there are several parameters, such as similarity threshold, number of similar genes in our proposed approach, we first examined their individual effects on the prediction accuracy and then optimized these parameters. Using the optimized parameters, we examined the prediction accuracy for different number of unknown genes. Our results are quite consistent and robust, where the prediction accuracy can be up to 0.97, much better than some available approaches. For example, the prediction accuracy for FBA is around 0.87 and 0.83 for *E. coli* and yeast respectively [17, 138]. Other machine learning based approaches have worse prediction accuracy even though a lot of structural and functional features are employed [14]. Similar prediction is performed in yeast, which also showed high accuracy.

Here we made a simple comparison between our approach and machine learning based approaches, since both are based on structural and functional features. First of all, our features are based on previous studies on how essential and non-essential genes confer to different deletion effects. We unveiled some structural and functional features that potentially associated with it and also proposed mechanisms explaining how the differential deletion effects are caused, which is supported by literatures. In

most machine learning based approaches, there is no clear causality between essentiality and these features, although with high statistical significance.

One major limit of our approach is that there are some genes cannot be predicted solely using our approach. If we consider these genes as misclassified, the prediction accuracy will drop to around 0.7, changing depending on other parameters. This may due to the incompleteness of our network or some critical links are missing. We analyzed the genes that cannot be predicted and classified them into two categories: the largest category is the genes that cannot be captured using these features. For these genes, they can only affect themselves. As a result, when we calculated similarity coefficients for gene pairs, these genes are always neglected. Such phenomena may due to the incompleteness of metabolic network and some critical links may be missing in the network. Another kind of gene is those that can only affect other unknown-type genes. As a consequence, these pairs are neglected together as no other genes in the known-type list can be used to give a proper prediction. This happens when the number of unknown genes is large.

We proposed two strategies to improve the prediction accuracy. First of all, combine with other methods. For example, combined with FBA showed improved prediction accuracy compared with FBA alone, with all the genes predicted. Another strategy is to slightly increase the ratio of essential genes in the known-type list. Studies showed

that increasing the ratio of known essential genes can increase prediction accuracy. This can be achieved via mapping essential genes between different species since it is more conserved. The most important way to increase prediction accuracy is to explore more essentiality related features to increase the coverage.

In summary, although we used limited number of functional and structural features, the prediction accuracy is comparable with other methods which may use up to 20 features. It is suggested that the most important factors in essential gene identification is how to unveil some useful and essentiality related features. Effects in understanding gene organization may shed light on essential gene identification.

Chapter 5 Large-scale epistasis analysis in *E. coli* and *S. cerevisiae*

In **Chapter 3**, we have extensively discussed single gene deletion effects and how essentiality is arising from the perspective of functional and topological organization. To reveal a higher order functional and structural organization of complex metabolic network (such as genetic buffering), it is necessary to study the genetic interactions in the context of metabolic networks. Here, we studied the systems-level double genetic interactions by computing the double gene deletion effects in both *E. coli* and *S. cerevisiae* using the proposed cascade failure procedures as described in **Chapter 2**. A comparison with single gene deletion effect revealed pairs of genes with reduced, enhanced, and unchanged deletion effects. Our analysis indicated that gene pairs with reduced deletion effects tend to be from the same pathway whereas gene pairs with enhanced deletion effects tend to be from diverse pathways, which is consistent with findings from the shortest path distance between gene pairs. Detailed investigations on these gene pairs offered some evolutionary clues and also the mechanisms regarding how double gene deletion can lead to lethality or survival.

5.1 Introduction

Robustness is an intrinsic property of metabolic networks [154], which enables the living organisms maintaining its functions in response to various genetic perturbations and environmental stresses. As most genes are dispensable for growth

[66, 88], implying the presence of compensatory mechanisms [199], an in-depth understanding on the genetic interactions (or epistasis) in the framework of metabolic network would gain insights into the complex organization of metabolic networks [200]. Though single gene deletion has been extensively studied [66, 88], which also revealed some valuable functional and structural organization principles of metabolic networks [49, 74, 102, 109, 201], a comprehensive understanding on the higher order organization requires large-scale double- or even multiple- gene deletion analysis.

Epistatic interaction, defined as the ability of one gene to mask the phenotypic impact caused by other mutants [121], may help us to elucidate functional associations between genes and the whole organizations of metabolic network. For example, a system-level single and double gene deletion analysis of 890 yeast genes revealed that the epistatic interaction networks can be organized hierarchically into function-enriched modules [50]. Positive epistasis (i.e. interactions between genes that can enhance their functionality) was prevalent observed in both *E. coli* and *S. cerevisiae* metabolic network [130]. Analysis of epistatic interactions also serves as the fundamental basis for understanding the synthetic lethal/rescue mutants [90, 125], which have attracted wide attention due to their pivotal role in drug development [68, 70, 124]. Despite the significance of epistatic interaction, the extent and nature of epistasis are less studied. As we have explored the nature of essentiality via

computing damage list for single gene deletion in **Chapter 3**, investigations on the damage lists for double gene deletion may shed light on the understanding of epistasis.

Mathematically, the way to characterize the deletion effects of two independent genes is not consistent between studies. One may define it as the sum of individual deletion effects [122], i.e. $W(xy) = W(x) + W(y)$, whereas other may use the multiplicative model, defining it as the production of individual mutational effects [122], i.e. $W(xy) = W(x) * W(y)$. The epistatic interaction is determined by the relative difference between the actual double deletions effects and the expected effects assuming they are independent. The different definitions of independence are partially due to the various ways to measure the deletion effects. For example, in the fitness analysis [50, 130], the growth rate is computed and compared to determine whether one gene is essential or not. For such traits, it is convenient to use the production rule. On the other hand, in our study, the damage list is adopted to capture the deletion effects for both single/double mutants, the additivity rule would be more appropriate.

In this study, we extended the single gene deletion procedures to capture the double gene deletion effects for any given gene pairs using metabolic network of *E. coli* and *S. cerevisiae*. Gene pairs with enhanced deletion effects, reduced deletion effects were identified and investigated. Pathway analysis suggested that gene pairs with

reduced deletion effects have a tendency of arising from the same pathway, whereas gene pairs with enhanced deletion effects tend to be from diverse pathways. The obtained properties are quite conserved, also found in yeast, suggesting evolutionarily conserved features. We further showed that lethality caused by non-essential gene pairs with enhanced deletion effects are due to the fact that their deletion effects can disrupt some key metabolites/reactions/genes, the loss of which lead to lethality. Essential gene pairs with reduced deletion effects are quite conserved across species. Our analysis implied the potential way to identify the candidates for synthetic lethal/synthetic rescue pairs and some clues to drug design.

5.2 Materials and methods

5.2.1 Metabolic network

The *E. coli* and *S. cerevisiae* metabolic networks used in double-gene deletion analysis are reconstructed from *E. coli* iAF1260 and *S. cerevisiae* iND1260 model which are available in BiGG database [136]. The reconstructed network is a directed bipartite graph which contains two types of nodes (metabolites and reactions nodes) (for details, please refer to **Section 2.2, Chapter 2**).

5.2.2 Deletion effects in response to double gene removal

In **Chapter 3**, we have introduced an approach which is used to characterize single gene deletion effects in the context of metabolic networks. This approach is extended to capture double gene deletion effects in this chapter, where the deletion effects are also characterized by a list of affected genes. The major difference between single gene deletion and double gene deletions is how to determine the corresponding reactions. For single gene deletion, it is defined as reactions that cannot proceed further due to the single gene removal. We cannot simply union the corresponding reactions for the gene pairs for double gene deletion. Instead, it is necessary to take the gene interactions into consideration. For instance, the removal of two isozymes can lead to the failure of reactions for lacking compensating pathways, which is not the case in single gene deletion. The non-linear effect introduced by the topological characteristics of metabolic network determines that the corresponding reactions of double genes are not simply equal to the additive of their individual.

In this algorithm, we first identify the ‘corresponding reactions’ for each pair of genes according to the gene-protein-reaction (GPR) association information (for details, please refer to **Section 2.2.2, Chapter 2**). Only when all the enzyme subsets are non-functional, we assume that a particular reaction cannot occur. Next, we initiate the cascading failure procedures by removing all the ‘corresponding reactions’ and their

links from the network simultaneously. Metabolite nodes with either zero in-degree or out-degree and their links are to be removed in the next step. Following, reactions nodes with incomplete substrates or products and their associated links are removed from the metabolic networks. Such procedures are repeated until all the leftover metabolites with non-zero in-degree and out-degree and reactions with complete substrates and products. Consequently, double gene deletion can be characterized by a list of affected reactions.

The removed reactions from previous step are then mapped to the corresponding gene lists. For each reaction identified, as long as they belong to another gene's corresponding reactions, we assume that this particular gene will be affected in response to double gene deletions. The union of all the affected genes is the damage lists for double gene deletion.

5.2.3 Shortest path distance

In a directed graph, for any gene pair $\{G_1, G_2\}$, a path is defined as a series of nodes $P = \{v_1, v_2, \dots, v_n\}$ that can connect these two nodes. The path with minimal nodes is called the shortest path and the number of nodes in this path is termed as the shortest distance between these two genes.

In our studies, a directed bipartite metabolic network is used, which is composed of metabolite and reaction nodes. Firstly, the reactions catalyzed by each gene are identified. For instance, $\{R_1, R_2, \dots, R_m\}$ are the reactions for gene G_1 and $\{r_1, r_2, \dots, r_n\}$ are the reactions for G_2 . In the next step, we search the graph for the shortest paths between any pair of reactions, i.e. $\{R_1, R_2, \dots, R_m\} \times \{r_1, r_2, \dots, r_n\}$ and among all these pairs, the one with the shortest distance is defined as the shortest path for gene pairs.

5.3 Double gene deletions in *E. coli*

5.3.1 Comparisons between damage lists for single- and double- gene deletions

The algorithm to characterize single gene deletion is tailored to capture the double gene deletion effects. The cascade failure procedures are initiated by removing the corresponding reactions of the target genes from the reconstructed metabolic network *iAF1260* [17], followed by the iterated failure procedures, which goes upwards and downwards until all the leftover metabolite nodes are of non-zero in-degree and out-degree and reaction nodes with complete substrates and products. To distinguish with single gene deletion, the damage list obtained for double gene deletion is termed as ‘damage list for double gene deletion’. Analysis in previous chapter reveals that essential genes spread their deletion effects via other essential genes. Thus, in this

study, instead of studying the whole damage lists, we only concern those affected essential genes (i.e. the essential genes in the damage list).

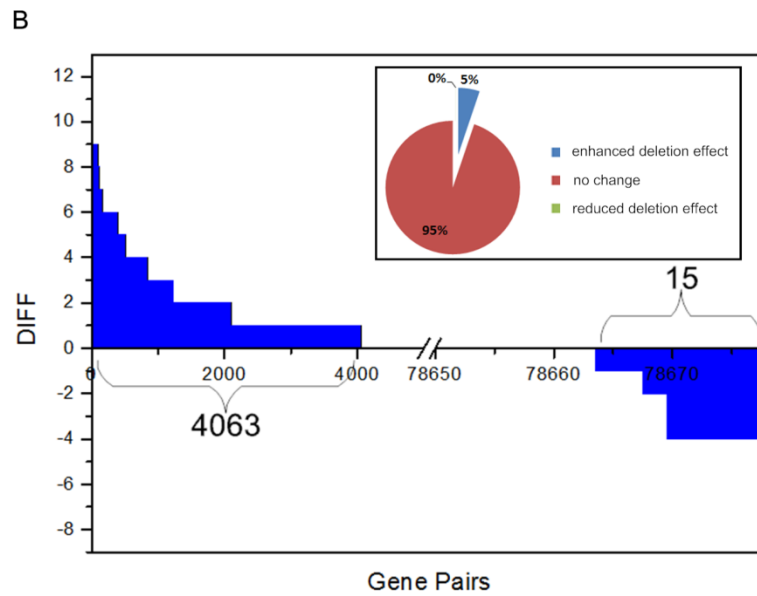
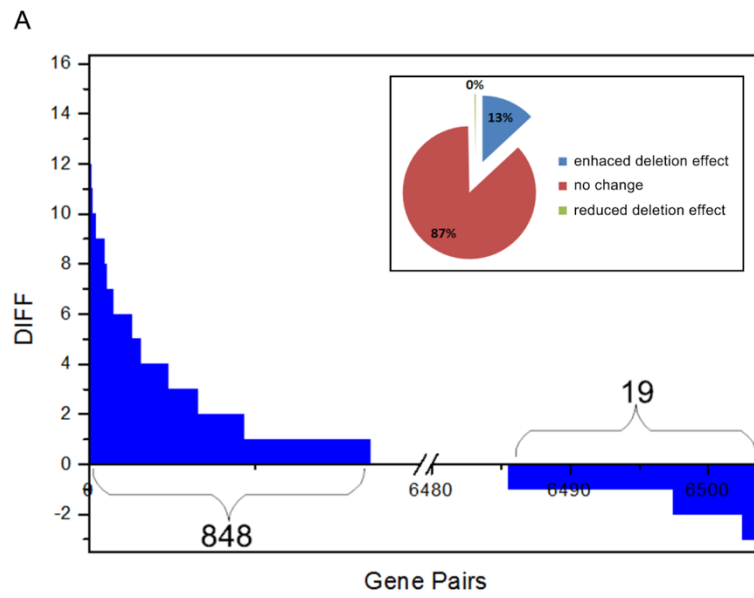
A comparison between the damage lists of single- and double- gene deletion unveiled gene-gene epistatic interactions. In this study, gene pairs are further grouped into three categories: gene pairs with enhanced deletion effects, reduced deletion effects and unchanged effects based on the relative relations between the damage list size of double gene deletions and single gene deletion. Suppose one gene pair $\{g_1, g_2\}$ where the damage list of individual gene is represented as $d(g_1), d(g_2)$, whereas the damage list of double deletion is denoted as $d(g_{12})$. It is considered as gene pair with enhanced deletion effects if $d(g_{12}) > d(g_1) \cup d(g_2)$. Similarly, it is classified as gene pair with reduced deletion effect if $d(g_{12}) < d(g_1) \cup d(g_2)$. The leftover gene pairs are those pairs with unchanged effects. From the perspective of genetic interactions, gene pairs with unchanged effects are of less information since it indicates the irrelevance of genes in terms of lethality or survival.

Among the 608075 gene pairs studied (those gene pairs with damage size larger than 0), 7628 of them showed enhanced deletion effects, whereas 47 with reduced deletion effects. As there are a large number of gene pairs with both zero double gene deletion (i.e. $d(g_{12}) = 0$) and single gene deletion effects (i.e. $d(g_1) = 0$ and $d(g_2) = 0$) (result

totally 158047 gene pairs), we excluded these gene pairs from our analysis. As a consequence, around 5% (7628 out of 158047) gene pairs with enhanced deletion effects, and less than 1% (47 out of 158047) gene pairs with reduced deletion effects are identified.

Further quantitative analysis of the difference between damage list sizes of double gene deletion and single gene deletion (**Figure 23**) indicates that although majority of the gene pairs have an absolute difference around 2 or 3, a small fraction of them can have a difference up to 10. The larger the difference, the tighter the interactions are considered to be.

19 essential/essential gene pairs (E-E) (**Figure 23A**) and 15 essential/non-essential (E-N) gene pairs (**Figure 23B**) are identified to have reduced deletion effects, while 2716 non-essential/non-essential gene pairs (N-N) show enhanced deletion effects. While the former are the ideal candidates for synthetic rescue, the latter pairs are considered to be ideal candidates for synthetic lethal. We will discuss these pairs in details. In the following sections, we will use E-E, E-N, and N-N to denote the corresponding gene pairs.



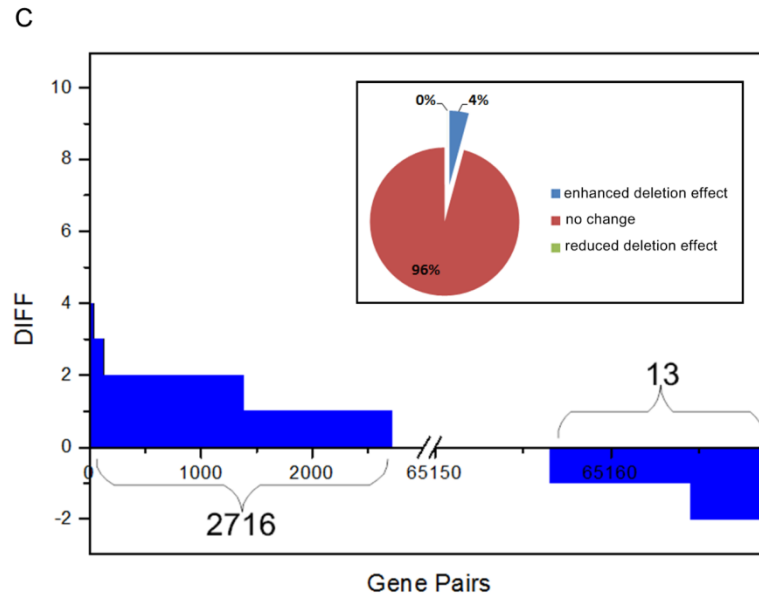


Figure 23 Epistatic interactions in *E. coli* between (A) E-E pairs, (B) E-N pairs, and (C) N-N pairs. The vertical is the epistasis between genes while the horizon represents the corresponding number of gene pairs. A positive epistasis indicates the enhanced deletion effects whereas a negative epistasis suggests the reduced deletion effects. Genes with non-epistatic interactions are excluded. The inset is the percentage of gene pairs with enhanced deletion effects, reduced deletion effects, and unchanged deletion effects in each individual group.

5.3.2 Pathway analysis for gene pairs with enhanced and reduced deletion effects

Two genes are participated in the same pathway if reactions catalyzed by the enzymes encoded by the genes belong to the same pathway. Statistical results of gene pairs

from different categories (total gene pairs, E-E, and N-N) breakdown by the sign of epistasis and pathway indicate that gene pairs with enhanced deletion effects tend to be from different pathways whereas gene pairs with reduced deletion effects are mainly from the same pathway (**Table 19**). Interestingly, such observations are strengthened in E-E pairs with reduced deletion effects (100% in E-E pairs versus 94% in Total gene pairs), and N-N pairs with enhanced deletion effects (85% in N-N pairs versus 83% in Total gene pairs).

Table 19 Summary of pathway analysis in *E. coli*

		Enhanced Deletion Effect		Unchanged		Reduced Deletion Effect	
		Same Pathway	Different Pathway	Same Pathway	Different Pathway	Same Pathway	Different Pathway
Total	No.	1319	6309	32970	567430	44	3
	%	17%	83%	5%	95%	94%	6%
E-E	No.	198	651	1039	6619	19	0
	%	23%	77%	14%	86%	100%	0%
N-N	No.	399	2317	26215	431097	12	1
	%	15%	85%	6%	94%	92%	8%

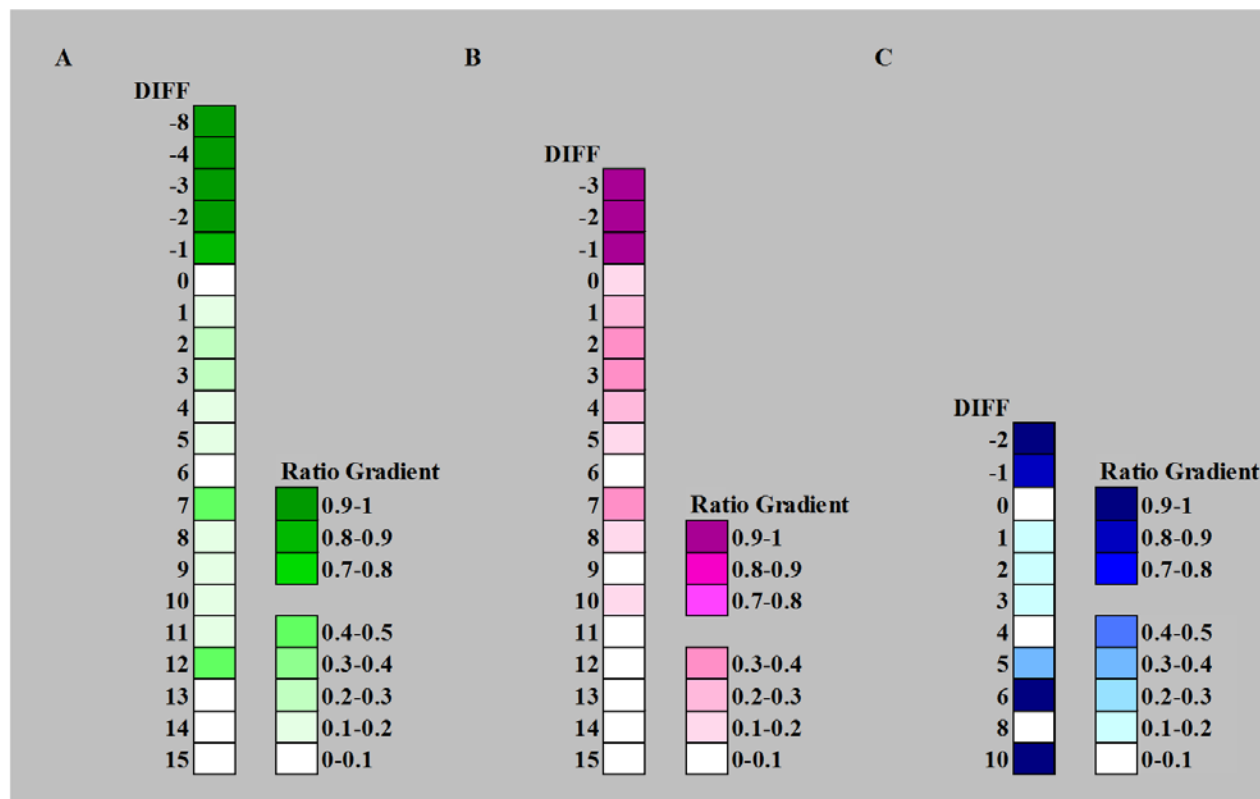


Figure 24 Pathway analysis of gene pairs in *E. coli*. The ratio of gene pairs from the same pathway for each epistasis level (labeled as DIFF) is computed for (A) Total gene pairs (B) E-E gene pairs (C) N-N gene pairs and represented as color gradient.

The detailed pathway analysis by considering gene pairs with different levels of epistasis is presented in **Figure 24** and the results are consistent with **Table 19**. In addition, the path distances between pairs of essential genes with reduced deletion effects are significantly shortened compared with the general E-E pool (**Figure 25**). On the other hand, the shortest distances between pairs of non-essential genes with enhanced deletion effects are lengthened compared with the general N-N pool (**Figure 26**).

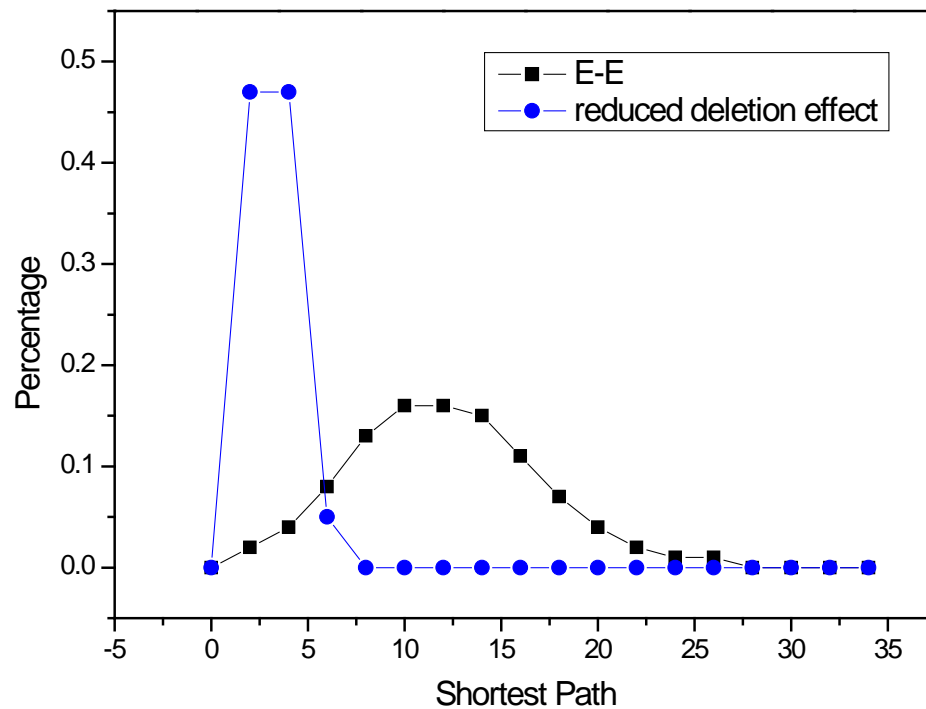


Figure 25 Distribution of shortest path distance between E-E pairs. The percentage of gene pairs with the corresponding shortest path distance (horizon) is plotted. Nearly 50% of E-E gene pairs with reduced deletion effect have a distance smaller than 5 (circle), while the average distance for E-E group is around 10 (rectangle).

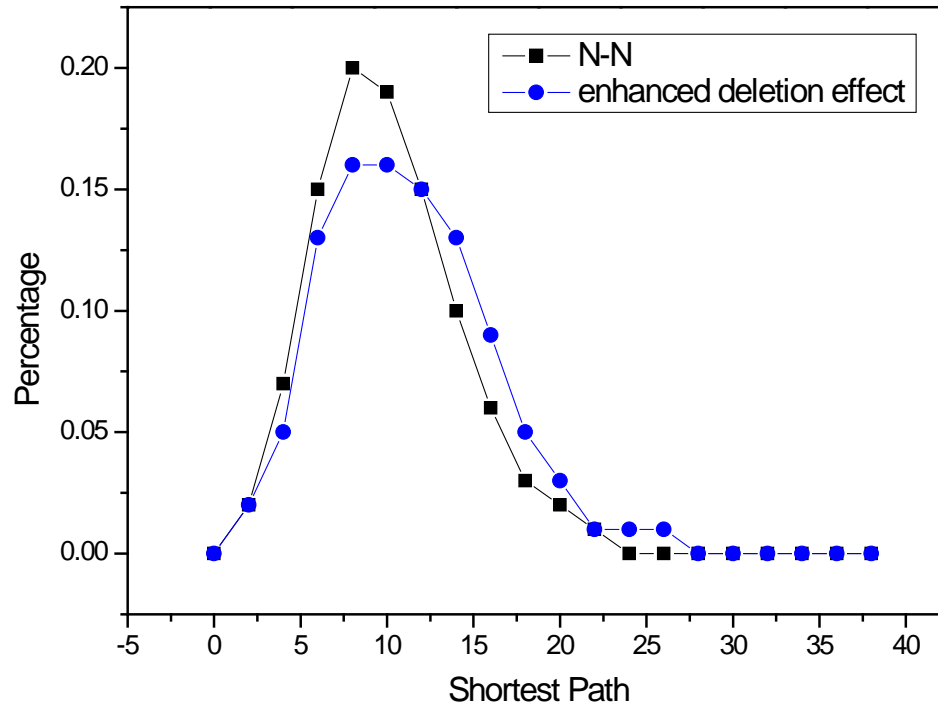
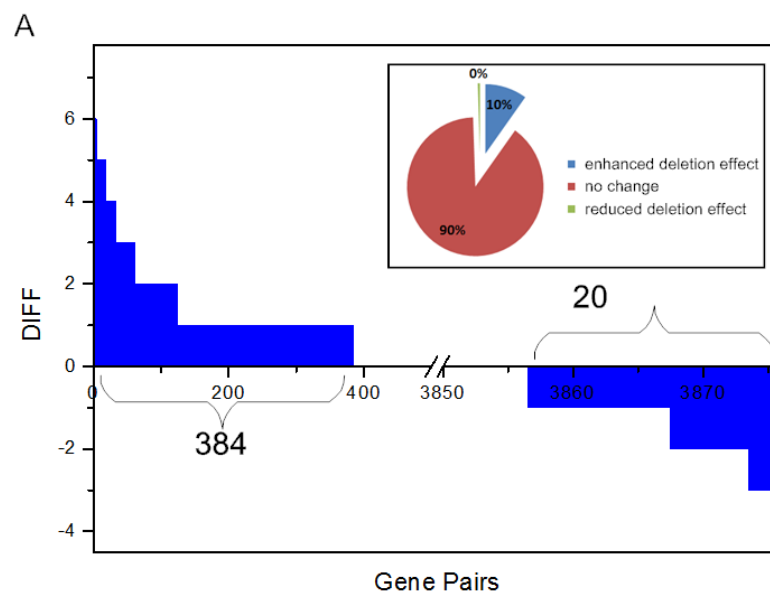


Figure 26 Distribution of shortest path distance between N-N pairs. The percentage of gene pairs with the corresponding shortest path distance (horizon) is plotted. The distribution for N-N pairs with enhanced deletion effect (circle) is left-skewed compared with whole N-N gene pairs (rectangle).

5.4 Double gene deletions in *S. cerevisiae*

To compare double gene deletion effects between organisms, a genome-scale double gene deletion analysis is conducted to *S. cerevisiae* *iND750* model [138]. Among the 212979 gene pairs studied, 2239 of them show enhanced deletion effects, whereas 45 with reduced deletion effects. Further quantitative analysis of the epistasis

interactions indicates that although majority of the gene pairs have an absolute difference around 1, a small fraction of them can have a difference up to 6 (**Figure 27**).



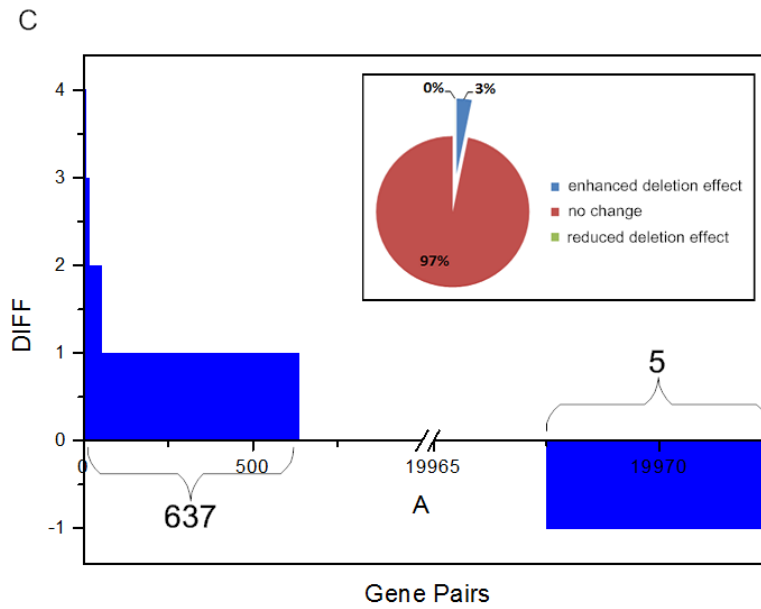
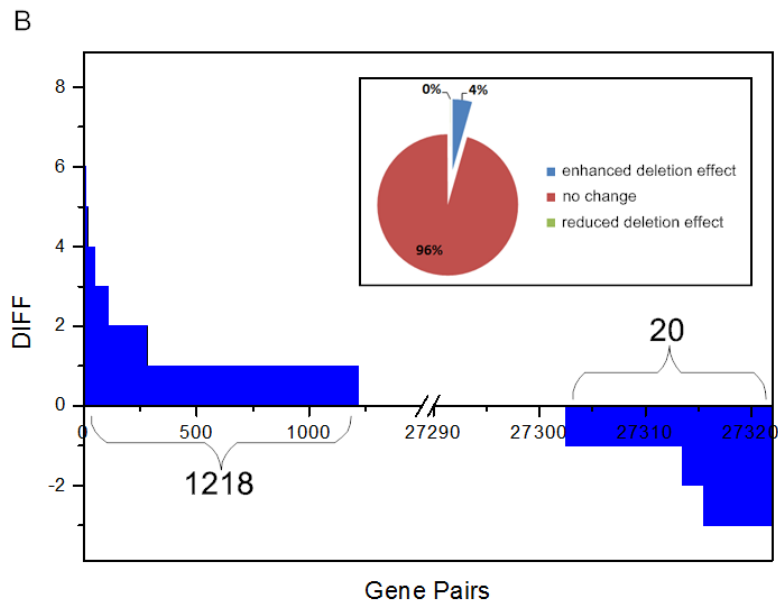


Figure 27 Epistatic interactions in *S. cerevisiae* between (A) E-E pairs, (B) E-N pairs, and (C) N-N pairs. The vertical is the epistasis between genes while the horizon represents the corresponding number of gene pairs. A positive epistasis indicates the enhanced deletion effects whereas a negative epistasis suggests the reduced deletion effects. Genes with non-epistasis interactions are excluded.

20 E-E pairs and 20 E-N pairs show reduced deletion effects while 637 N-N pairs show enhanced deletion effects. Findings from pathway analysis (**Table 20, Figure 28**) are also consistent with *E. coli*, in other words, E-E pairs with reduced deletion effects tend to be from the same pathway, whereas N-N pairs with enhanced deletion effects are mainly from different pathways.

Table 20 Summary of pathway analysis for *S. cerevisiae*

		Enhanced Deletion Effect		Unchanged		Reduced Deletion Effect	
		Same Pathway	Different Pathway	Same Pathway	Different Pathway	Same Pathway	Different Pathway
Total	No.	98	2141	5751	204944	41	4
	%	4%	96%	3%	97%	91%	9%
E-E	No.	25	359	201	6116	19	1
	%	7%	93%	3%	97%	95%	5%
N-N	No.	41	596	3946	135181	5	0
	%	6%	94%	3%	97%	100%	0%

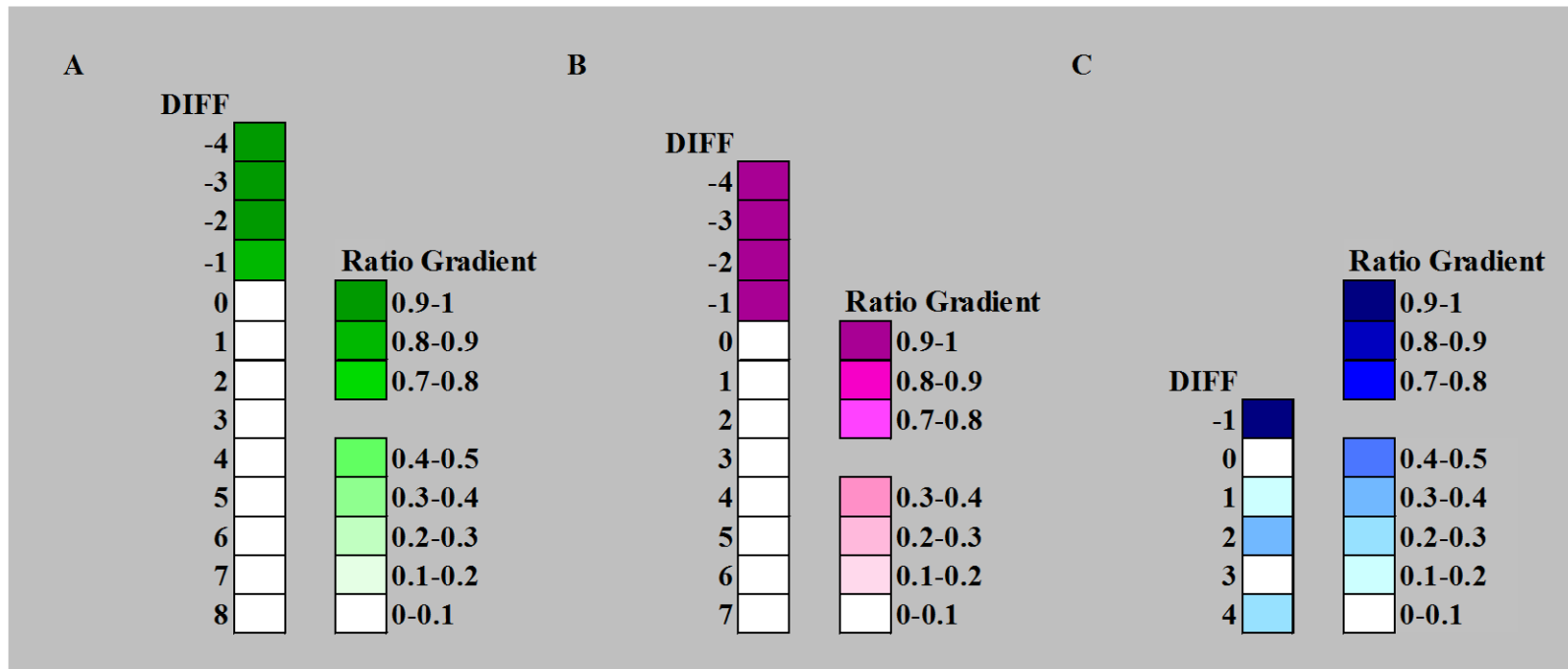


Figure 28 Pathway analysis of gene pairs in *S. cerevisiae*. The ratio of gene pairs from the same pathway for each epistasis level is computed for (A) Total gene pairs (B) E-E gene pairs (C) N-N gene pairs and represented by color gradient.

5.5 Gene pairs with reduced deletion effects arise from the same pathway and are conserved between species

Investigations on gene pairs with reduced double gene deletion effects indicate a tendency of genes arising from the same pathway, or functional related pathways. This is easy to be understood since the deletion effects are more likely to be interrupted shortly after the cascade failure procedures are initiated in order to have the reduced effects. Mapping the gene pairs to KEGG PATHWAY database shows that these gene pairs from *E. coli* are mainly located in the following pathways: Glycan Biosynthesis and Metabolism, Metabolism of Terpenoids and Polyketides, Metabolism of Cofactors and Vitamins and Metabolism of other Amino Acids (**Table 21**). On the other hand, those genes from *S. cerevisiae* are found to be from these pathways: Metabolism of Terpenoids and Polyketides, Metabolism of Cofactors and Vitamins, Lipid Metabolism, and Amino Acid Metabolism (**Table 22**).

Two common pathways are shared by these two *E. coli* and *S. cerevisiae*. One is Metabolism of Cofactors and Vitamins, the genes from which pathways are further identified to be orthologous genes and encode the same enzymes (**Table 23**). Another is Metabolism of Terpenoids and Polyketides, genes of *E. coli* and *S. cerevisiae* are found belong to two alternative pathways of Isopentenyl-diphosphate production. For the biosynthesis of Isopentenyl-diphosphate, there are usually two ways: one is the classical mevalonate pathway, which is available in higher eukaryotes and many

bacteria, whereas the other is the non-mevalonate pathway [172]. Our analysis suggests that those genes of yeast are actively present in mevalonate pathway, whereas their counterparts from *E. coli* are available in non-mevalonate isoprenoid pathway. In addition to these common pathways, there are some species-specific pathways, e.g. Glycan Biosynthesis and Metabolism, the function of which is to maintain cell wall integrity.

Table 21 E-E gene pairs with reduced deletion effects (*E. coli*)

Gene1	Gene2	Symbol1	Symbol2	Pathway	Pathway Category
b0029	b2515	<i>ispH</i>	<i>ispG</i>	Terpenoid backbone biosynthesis	Metabolism of terpenoids and polyketides
b0029	b2746	<i>ispH</i>	<i>ispF</i>	Terpenoid backbone biosynthesis	Metabolism of terpenoids and polyketides
b0085	b3967	<i>murE</i>	<i>murI</i>	Peptidoglycan biosynthesis	Glycan biosynthesis and metabolism
b0085	b3972	<i>murE</i>	<i>murB</i>	Peptidoglycan biosynthesis	Glycan biosynthesis and metabolism
b0086	b0087	<i>murF</i>	<i>mraY</i>	Peptidoglycan biosynthesis	Glycan biosynthesis and metabolism
b0086	b0090	<i>murF</i>	<i>murG</i>	Peptidoglycan biosynthesis	Glycan biosynthesis and metabolism
b0087	b0090	<i>mraY</i>	<i>murG</i>	Peptidoglycan biosynthesis	Glycan biosynthesis and metabolism
b0088	b3972	<i>murD</i>	<i>murB</i>	Peptidoglycan biosynthesis	Glycan biosynthesis and metabolism
b0091	b3967	<i>murC</i>	<i>murI</i>	D-Glutamine and D-glutamate metabolism	Metabolism of other amino acids
b0096	b0524	<i>lpxC</i>	<i>lpxH</i>	Lipopolysaccharide biosynthesis	Glycan biosynthesis and metabolism
b0154	b1210	<i>hemL</i>	<i>hemA</i>	Porphyrin and chlorophyll metabolism	Metabolism of cofactors and vitamins
b0154	b2400	<i>hemL</i>	<i>gltX</i>	Porphyrin and chlorophyll metabolism	Metabolism of cofactors and vitamins
b0173	b1208	<i>dxr</i>	<i>ispE</i>	Terpenoid backbone biosynthesis	Metabolism of terpenoids and polyketides
b0173	b2747	<i>dxr</i>	<i>ispD</i>	Terpenoid backbone biosynthesis	Metabolism of terpenoids and polyketides
b0179	b0524	<i>lpxD</i>	<i>lpxH</i>	Lipopolysaccharide biosynthesis	Glycan biosynthesis and metabolism
b0369	b3805	<i>hemB</i>	<i>hemC</i>	Porphyrin and chlorophyll metabolism	Metabolism of cofactors and vitamins
b0415	b3041	<i>ribE</i>	<i>ribB</i>	Riboflavin metabolism	Metabolism of cofactors and vitamins
b1208	b2747	<i>ispE</i>	<i>ispD</i>	Terpenoid backbone biosynthesis	Metabolism of terpenoids and polyketides
b2515	b2746	<i>ispG</i>	<i>ispF</i>	Terpenoid backbone biosynthesis	Metabolism of terpenoids and polyketides

Table 22 E-E gene pairs with reduced deletion effects (*S. cerevisiae*)

Gene1	Gene2	Symbol 1	Symbol 2	Pathway	Pathway Category
YDL205C	YDR232W	<i>hem3</i>	<i>hem1</i>	Porphyrin and chlorophyll metabolism	Metabolism of cofactors and vitamins
				Glycine, serine and threonine metabolism	Amino acid metabolism
YDL205C	YOR278W	<i>hem3</i>	<i>hem4</i>	Porphyrin and chlorophyll metabolism	Metabolism of cofactors and vitamins
YDR044W	YDR047W	<i>hem13</i>	<i>hem12</i>	Porphyrin and chlorophyll metabolism	Metabolism of cofactors and vitamins
YDR232W	YGL040C	<i>hem1</i>	<i>hem2</i>	Glycine, serine and threonine metabolism	Amino acid metabolism
				Porphyrin and chlorophyll metabolism	Metabolism of cofactors and vitamins
YDR232W	YOR278W	<i>hem1</i>	<i>hem4</i>	Glycine, serine and threonine metabolism	Amino acid metabolism
				Porphyrin and chlorophyll metabolism	Metabolism of cofactors and vitamins
YER043C	YGL001C	<i>sah1</i>	<i>erg26</i>	Cysteine and methionine metabolism	Amino acid metabolism
				Steroid biosynthesis	Lipid metabolism
YGL001C	YLR100W	<i>erg26</i>	<i>erg27</i>	Steroid biosynthesis	Lipid metabolism
YGL040C	YOR278W	<i>hem2</i>	<i>hem4</i>	Porphyrin and chlorophyll metabolism	Metabolism of cofactors and vitamins
YGR175C	YHR072W	<i>erg1</i>	<i>erg7</i>	Steroid biosynthesis	Lipid metabolism
				Sesquiterpenoid and triterpenoid biosynthesis	Metabolism of terpenoids and polyketides
YGR175C	YHR190W	<i>erg1</i>	<i>erg9</i>	Sesquiterpenoid and triterpenoid biosynthesis	Metabolism of terpenoids and polyketides

YGR175C	YJL167W	<i>erg1</i>	<i>erg20</i>	Sesquiterpenoid and triterpenoid biosynthesis Terpenoid backbone biosynthesis	Metabolism of terpenoids and polyketides Metabolism of terpenoids and polyketides
YHR007C	YHR190W	<i>erg11</i>	<i>erg9</i>	Steroid biosynthesis	Lipid metabolism
YHR007C	YJL167W	<i>erg11</i>	<i>erg20</i>	Sesquiterpenoid and triterpenoid biosynthesis Steroid biosynthesis Terpenoid backbone biosynthesis	Metabolism of terpenoids and polyketides Lipid metabolism Metabolism of terpenoids and polyketides
YHR072W	YHR190W	<i>erg7</i>	<i>erg9</i>	Steroid biosynthesis	Lipid metabolism
YHR072W	YJL167W	<i>erg7</i>	<i>erg20</i>	Sesquiterpenoid and triterpenoid biosynthesis Steroid biosynthesis Terpenoid backbone biosynthesis	Metabolism of terpenoids and polyketides Lipid metabolism Metabolism of terpenoids and polyketides
YHR190W	YJL167W	<i>erg9</i>	<i>erg20</i>	Sesquiterpenoid and triterpenoid biosynthesis	Metabolism of terpenoids and polyketides
YMR208W	YPL117C	<i>erg12</i>	<i>idi1</i>	Terpenoid backbone biosynthesis	Metabolism of terpenoids and polyketides
YMR220W	YPL117C	<i>erg8</i>	<i>idi1</i>	Terpenoid backbone biosynthesis	Metabolism of terpenoids and polyketides
YNR043W	YPL117C	<i>mvd1</i>	<i>idi1</i>	Terpenoid backbone biosynthesis	Metabolism of terpenoids and polyketides
YOR176W	YOR278W	<i>hem15</i>	<i>hem4</i>	Porphyrin and chlorophyll metabolism	Metabolism of cofactors and vitamins

Table 23 Gene pairs from Metabolism of Terpenoids and Polyketides

Yeast Gene	Yeast Gene Name	<i>E. coli</i> Gene	<i>E. coli</i> Gene Name	Enzyme
YGL040C	<i>hem2</i>	b0369	<i>hemB</i>	EC4.2.1.24
YDL205C	<i>hem3</i>	b3805	<i>hemC</i>	EC2.5.1.61
YOR278W	<i>hem4</i>	b3804	<i>hemD</i>	EC4.2.1.75
YDR047W	<i>hem12, hem6</i>	b3997	<i>hemE</i>	EC4.1.1.37
YDR044W	<i>hem13</i>	b2436	<i>hemF</i>	EC1.3.3.3
YER014W	<i>hem14</i>	--	--	EC1.3.3.4
--	--	b3850	<i>hemG</i>	EC1.3.5.3
YOR176W	<i>hem15</i>	b0475	<i>hemH</i>	EC4.99.1.1
--	--	b3867	<i>hemN</i>	EC1.3.99.22

5.6 Gene pairs with enhanced deletion effects disturb key reactions or metabolites

In the following section, we will discuss those N-N pairs with enhanced deletion effects, i.e. gene pairs whose double deletion effects have a relative larger impact on the whole network in comparison with the sum of their individual's. Pathway analysis shows the diverse pathways they are arising from. In *E. coli*, the most frequently pathway combinations are Amino Acid Metabolism & Carbohydrate metabolism, and Amino Acid Metabolism & Metabolism of cofactors and vitamins (**Table 24**), whereas in yeast, the combination is shifted to: Amino Acid Metabolism & Lipid metabolism (**Table 25**).

Further examination on the damage lists for these N-N pairs from *E. coli* suggests that essential genes can be impaired in response to these double gene deletions, although it seldom occurs for individual gene deletion. The summarized results are listed for N-N pairs with epistasis larger than 5 (**Table 26**). It is noticed that b0159/b2436 (*mtn/hemF*), b0908/b2265 (*aroA/menF*), and b2265/b2329 (*menF/aroC*) can affect essential gene b3835 (*ubiB*) and reaction OPHHX3. Interestingly, these two are found to be synthetic lethal pairs in other studies [125]. Besides, gene pairs b1107/b3809 (*nagZ/dapF*), b1119/b3809 (*nagK/dapF*), and b1640/b3809 (*anmK/dapF*) can impair the function of essential genes, such as b0085 (*murE*), b0086 (*murF*), b0087 (*mraY*), b0088 (*murD*), b0090 (*murG*), b0091 (*murC*), b2472 (*dapE*), b3189 (*murA*), and

b3972 (*murB*). It is discussed in **Chapter 3** that they can form associated gene sets, whose impairment can disrupt the production of certain key metabolites. We will address this point in details later.

Table 24 N-N gene pairs with enhanced deletion effects (*E. coli*)

Gene1	Gene2	Symbol1	Symbol2	Pathway Category for Gene1	Pathway Category for Gene2
b1054	b2378	<i>lpxL</i>	<i>ddg</i>	Glycan biosynthesis and metabolism	-
b1107	b3809	<i>nagZ</i>	<i>dapF</i>	Carbohydrate metabolism	Amino acid metabolism
b1119	b3809	<i>nagK</i>	<i>dapF</i>	Carbohydrate metabolism	Amino acid metabolism
b1640	b3809	<i>anmK</i>	<i>dapF</i>	-	Amino acid metabolism
b1849	b4079	<i>purT</i>	<i>fdhF</i>	Metabolism of cofactors and vitamins	Carbohydrate metabolism
b0159	b2436	<i>mtn</i>	<i>hemF</i>	Amino acid metabolism	Metabolism of cofactors and vitamins
b0908	b2265	<i>aroA</i>	<i>menF</i>	Amino acid metabolism	Metabolism of cofactors and vitamins
b2265	b2329	<i>menF</i>	<i>aroC</i>	Metabolism of cofactors and vitamins	Amino acid metabolism

Table 25 N-N gene pairs with enhanced deletion effects (*S. cerevisiae*)

Gene1	Gene2	Symbol1	Symbol2	Pathway Category for Gene1	Pathway Category for Gene2
YDR127W	YGL012W	<i>aro1</i>	<i>erg4</i>	Amino acid metabolism	Lipid metabolism
YDR354W	YGL012W	<i>trp4</i>	<i>erg4</i>	Amino acid metabolism	Lipid metabolism
YDR538W	YGL012W	<i>pad1</i>	<i>erg4</i>	Metabolism of cofactors and vitamins	Lipid metabolism
YER055C	YGL012W	<i>his1</i>	<i>erg4</i>	Amino acid metabolism	Lipid metabolism
YGL012W	YIL020C	<i>erg4</i>	<i>his6</i>	Lipid metabolism	Amino acid metabolism
YHR208W	YJR148W	<i>bat1</i>	<i>bat2</i>	Amino acid metabolism	Amino acid metabolism
YAL012W	YGR012W	<i>cys3</i>	-	Amino acid metabolism	Amino acid metabolism
YBR041W	YGL012W	<i>fat1</i>	<i>erg4</i>	-	Lipid metabolism
YBR176W	YGL012W	<i>ecm31</i>	<i>erg4</i>	Metabolism of cofactors and vitamins	Lipid metabolism
YBR184W	YHR046C	-	<i>inm1</i>	-	Metabolism of cofactors and vitamins
YDR007W	YGL012W	<i>trp1</i>	<i>erg4</i>	Amino acid metabolism	Lipid metabolism
YGL012W	YGL026C	<i>erg4</i>	<i>trp5</i>	Lipid metabolism	Amino acid metabolism
YGL012W	YJR078W	<i>erg4</i>	<i>bnal2</i>	Lipid metabolism	Amino acid metabolism
YGL012W	YKL184W	<i>erg4</i>	<i>spe1</i>	Lipid metabolism	Amino acid metabolism
YGL012W	YKL211C	<i>erg4</i>	<i>trp3</i>	Lipid metabolism	Amino acid metabolism
YGL012W	YOL052C	<i>erg4</i>	<i>spe2</i>	Lipid metabolism	-
YGL012W	YPR069C	<i>erg4</i>	<i>spe3</i>	Lipid metabolism	Amino acid metabolism

Table 26 Detailed analysis of N-N gene pairs with enhanced deletion effects. Some additional essential genes can be affected in response to double gene deletions.

Gene1	Gene2	Additionally affected essential genes
b1054	b2378	b0096, b0179, b0182, b0524, b0915, b0918, b1215, b3623, b3633, b3793
b1107	b3809	b0085, b0086, b0087, b0088, b0090, b0091, b2472, b3189, b3972
b1119	b3809	b0085, b0086, b0087, b0088, b0090, b0091, b2472, b3189, b3972
b1640	b3809	b0085, b0086, b0087, b0088, b0090, b0091, b2472, b3189, b3972
b1297	b3870	b0142, b0680, b2315, b2514, b2780, b3729
b1849	b4079	b0414, b0415, b1277, b1662, b2153, b3041
b0159	b2436	b0369, b0475, b3804, b3805, b3835, b3850, b4040
b0908	b2265	b0142, b2315, b3187, b3835, b4040
b2265	b2329	b0142, b2315, b3187, b3835, b4040

5.7 Discussions

In this chapter, a genome-scale double gene deletion is carried out, aiming to reveal some higher level functional or structural organizations of the complex metabolic networks. A comparison with single gene deletion revealed some gene pairs with either enhanced or reduced deletion effects. It is shown that for most gene pairs, the absolute damage size difference between double gene deletions and single gene deletion ranges from 1 to 3, whereas for a small fraction of gene pairs, the difference can be up to 6 or even higher. As observed, the number of gene pairs with enhanced deletion effects is much more than those with reduced effects, indicating that multiple gene mutants frequently introduce deleterious effects and a decreased fitness of organism. Furthermore, the existence of gene pairs with reduced deletion demonstrates that the deleterious phenotypes can be masked due to certain genetic interactions.

Pathway analysis indicates that gene pairs with reduced deletion effects tend to be from the same pathway, whereas gene pairs with enhanced deletion effects are more likely to be from different pathways. Such observations are strengthened for E-E pairs with reduced deletion and N-N pairs with enhanced deletion. The reason we concerned about E-E pairs with reduced double deletion effects and N-N with enhanced deletion effects is that they are the ideal candidates for synthetic rescue and synthetic lethal pairs. Synthetic rescue are defined as essential gene pairs

whose double deletion cannot lead to lethality anymore, whereas synthetic lethal are non-essential gene pairs whose double deletion can lead to lethality.

Back to our analysis, gene pairs with a shorter distance are more likely to impact on each other and thus lead to a truncated damage list. On the other hand, for gene pairs from diverse pathways, it is more likely that the impact from the other gene occurs at the end of the cascade failure procedures, instead of the intermediate, considering the damage list size of single gene deletion. Thus, an enhanced deletion effect is expected for this case. The strengthened effects of E-E with reduced deletion effects can be explained by the structural organization of essential gene pairs, as they are more likely to be linked via low-degree metabolites [201].

Further analysis on the shortest path distance between E-E gene pairs with reduced deletion effects shows a significantly shortened distance compared with the whole pool of E-E gene pairs. On the other hand, for N-N gene pairs with enhanced deletion effects, the distances are relatively higher than the control N-N gene pairs.

In **Chapter 3**, we have discussed how essentiality is arisen. For essential genes, their damage lists showed the ability to affect other essential genes and non-essential genes. The cooperative effects of these essential genes finally can disrupt

some associated gene sets and block the production or consumption of certain key metabolites, the loss of which finally impact on the cell survival. An in-depth analysis on gene pairs with enhanced deletion effects offers some functional organization clues. For example, for gene pair b2265 (*menF*) and b2329 (*aroC*), it will affect the following essential genes: b0142 (*folK*), b2315 (*folC*), b3187 (*ispB*), b3835 (*ubiB*), and b4040 (*ubiA*), whereas their individual gene deletion cannot. Other than these essential genes, some reactions such as OPHHX3 can also be impaired. Interestingly, b3835 (*ubiB*) and OPHHX3 are identified as synthetic lethal pairs in other studies [125], implying the synthetic lethal relations between b2265 (*menF*) and b2329 (*aroC*). Another example is: b1107/b3809 (*nagZ/dapF*), b1119/b3809 (*nagK/dapF*), and b1640/b3809 (*anmK/dapF*). All of these three gene pairs can disrupt the following essential genes: b0085 (*murE*), b0086 (*murF*), b0087 (*mraY*), b0088 (*murD*), b0090 (*murG*), b0091 (*murC*), b2472 (*dapE*), b3189 (*murA*), and b3972 (*murB*). Among these genes, b0085/b0088 (*murE/murD*), b0087/b0090 (*mraY/murG*), and b3189/b3972 (*murA/murB*) form three associated gene sets as identified in our previous studies. Perturbations on these modules can impact on some key metabolites, such as UDP-N-acetylmuramoyl-L-alanyl-D-gamma-glutamyl-meso-2,6-diaminopimelate, UDP-N-acetylmuramoyl-L-alanyl-D-glutamate, Undecaprenyl-diphospho-N-acetylmuramoyl-L-alanyl-D-glutamyl-meso-2,6-diaminopimeloyl-D-alanyl-D-alanine, Undecaprenyl-diphospho-N-acetylmuramoyl-(N-acetylglucosamine)-L-ala-D-glu-meso-2,6-diaminopimeloyl-D-ala-D-ala, UDP-N-acetyl-D-glucosamine and UDP-N-acetyl-3-O-(1-carboxyvinyl)-D-glucosamine. These metabolites are indispensable for the cell envelope biosynthesis.

Our analysis is further extended to gene pairs with reduced deletion effects, especially those essential gene pairs. Two essential genes with reduced deletion effects are revealed to be from the same pathway. For *E. coli*, these genes are mainly located in the following pathways: Glycan Biosynthesis and Metabolism, Metabolism of Terpenoids and Polyketides, Metabolism of Cofactors and Vitamins and Metabolism of other Amino Acids. Studies on *S. cerevisiae* imply the conservation of pathways as gene pairs from Metabolism of Terpenoids and Polyketides, and Metabolism of Cofactors and Vitamins are observed. The former are the substances that are required for the activity of an enzyme or protein. It is actively bound to the chemical groups and then carries them between different reactions. These cofactors act as the important intermediates for metabolism reactions. From our analysis, it is suggested that the functional organization of essential genes in some key processes are quite conserved across species.

Besides these shared pathways, each species have its own specific pathways which include E-E pairs with reduced deletion effects. For example, Glycan Biosynthesis and Metabolism involves the biosynthesis of peptidoglycan. Peptidoglycan is the fundamental component of bacteria cell wall, which surrounds and protects individual bacteria cell. This process is unique to bacteria. So while some essential gene pairs with reduced deletion effects are from some functional conserved pathway, there are a few species-specific pathways.

Our scope is not limited to *E. coli* and a parallel study is also carried out in *S. cerevisiae*. Investigations on gene pairs with enhanced and reduced deletion effects support our findings, implying some evolutionary clues. As discussed, there is a tendency that essential gene pairs with reduced deletion effects are arising from the same pathways. Detailed investigation demonstrates that genes from Metabolism of Cofactors and Vitamins are indeed orthologous genes between *E. coli* and *S. cerevisiae*, further suggesting the conservations between species. Apart from those conserved pathways, there are some species-specific pathways involved, such as Glycan Biosynthesis and Metabolism. Analysis of N-N pairs with enhanced deletion effects implies that the mapped pathways are quite diverse. For example, in *E. coli*, the most frequent pathway combination is Amino Acid Metabolism & Metabolism of Cofactors and Vitamins, whereas it is shifted to Amino Acid Metabolism & Lipid Metabolism in yeast.

In summary, genome-scale analysis of double gene deletion offers some new insights into the functional and structural organizations of complex biological networks. From the structural point of view, gene pairs with enhanced deletion effects tend to be from different pathways, meanwhile those with reduced deletion effects are from the same ones. From the functional perspective, our analysis indicates a possible mechanism regarding how N-N gene pairs can lead to lethality, i.e. they can affect some essential genes/reactions/metabolites, and the deletion of which finally lead to lethality. The cross comparison between *E. coli* and yeast also implies some evolutionary clues.

Chapter 6 Summary and future work

In this chapter, the major findings and contributions are summarized. We also discuss the limitations of this work and some possible future directions.

6.1 Major findings and contributions

Considering the indispensable role of essential genes in cellular growth, gene essentiality has been intensively studied, including gene deletion analysis [88, 109], essential gene prediction [159, 160], and genetic epistasis studies [50, 129-131]. Three types of mechanisms are deemed to be the causes of gene dispensability: (1) Conditional essentiality, i.e. those genes identified as non-essential genes are actually essential in other conditions that have not been examined in current laboratory conditions [103]; (2) Genetic buffering by duplicated genes or genes with overlapping functions [103, 106]; (3) Functional compensation by alternative pathways or redistribution of fluxes [103, 107]. However, knowledge about the intermediate biological processes following essential and non-essential gene deletion is extremely limited. Therefore, in this thesis, we tried to fill this gap by studying the topological features and functional organizations of genes at the metabolic network context, with the aim to enrich our understanding on the causality of gene essentiality.

We firstly introduced a novel way to measure single gene deletion effects by ‘damage list’, which is defined as a set of affected genes in response to target gene deletion. For *E. coli* iAF1260 metabolic network, we found that the deletion of essential genes generally cause a wider range of network failure. Further analyzing the composition of the damage list revealed that essential genes tend to spread the deletion effect to other essential and non-essential genes, whereas non-essential genes seldom affect other essential genes. Based on the distinct deletion patterns, we made a conjecture that genes sharing similar damage lists are more likely to exhibit similar type of essentiality. Furthermore, we identified some associated gene sets, which were able to perturb the production of certain key metabolites, and whose failure might induce lethality [109, 168]. Network structural analysis also suggested the distinguishable topological organizations of essential and non-essential genes in the metabolic system. Neighboring essential gene pairs tend to be interconnected through low-degree metabolites (namely there are no redundant paths between them), whereas neighboring non-essential genes pairs are mainly linked by high-degree metabolites. Therefore, the failure of one essential gene can easily spread its deletion effects to neighboring essential genes through the ‘fragile’ irreplaceable path. This served as an important evidence for the observed phenomenon that the damage list of one essential gene is mainly made up of other essential genes.

Inspired by the above findings that some structural and functional features are tightly associated with gene essentiality, we developed a novel approach that aims

to predict unknown genes from a limited number of genes with known essentiality type. By using the gene essentiality associated features and optimized parameters, we tested the prediction accuracy of our algorithm for the number of unknown genes ranging from 100 to 1200 for *E. coli*. It is found out that the prediction accuracy is quite high and robust. As some genes cannot be predicted using this method, we combined our approach with some existing ones (such as FBA). The hybrid method can predict all the target genes while giving high-level prediction accuracy. Similar results were also obtained for yeast metabolic network, indicating the robustness of our algorithm.

The epistatic effects between mutants were also studied in order to gain in-depth insights on the higher order structural and functional organization principles of biological networks. Gene pairs with enhanced and reduced deletion effects were identified and studied. It is revealed that gene pairs with reduced deletion effects tend to be arisen from the same pathway whereas those pairs with enhanced deletion effects tend to be arisen from different pathways. This is further confirmed by the shortest path studies which implied that the shortest path distance between gene pairs (particularly E-E pairs) with reduced deletion effects are significantly decreased, whereas the shortest path distance between gene pairs (especially N-N pairs) with enhanced deletion effects are significantly increased. The observed properties are quite robust, as also had been found in yeast metabolic network, indicating evolutionarily conserved features. Further mechanism studies showed that lethality caused by non-essential gene pairs with

enhanced deletion effects are due to the fact that their deletion effects can disrupt some key metabolites/reactions/genes, leading to cellular lethality. In addition, essential gene pairs with reduced deletion effects are quite conserved across species.

In summary, our studies proposed a reasonable explanation for the differential deletion effects between essential and non-essential genes from the network perspective, which is seldom investigated by previous studies. Other than the heavy focus on essential gene identification, understanding the functional and structural organization principles of essential and non-essential genes is equally important since they can help us to interpret cellular processes clearly and to design more efficient drug targets. Furthermore, we proposed an effective computational strategy to predict essential genes. By applying features tightly associated with essential genes, our algorithm for gene essentiality prediction outperformed the traditional constraint-based and machine learning approaches. This further emphasized the fact that understanding the nature of essential genes is a prerequisite for designing high quality gene essentiality prediction approaches. Last but not least, our studies enriched the understanding on the epistatic effects between mutants. The essential/essential (E-E) gene pairs with reduced deletion effects and non-essential/non-essential (N-N) gene pairs with enhanced deletion effects were intensively studied from the topological, mechanistic, and evolutionary perspectives, which warranted a new way to identify synthetic lethality and synthetic rescue gene pairs.

6.2 Future work

A comprehensive understanding of gene essentiality can provide novel insights for drug design and development [70, 72, 189]. It would be much easier and more efficient to select certain genes as drug targets if we have a clear picture of gene essentiality and epistatic interactions. Therefore, our future work will dig deep into gene deletion effects not only from the context of metabolic network but also from other types of biological networks such as protein-protein interaction networks and transcriptional networks. We will also attempt to unveil other network topological features that tightly associated with gene essentiality, for example, finding key functioning motifs that are associated to essential genes [202]. Those network measurements combining with biological classifications such as GO terms [203] could also increase the power to discriminate between essential and non-essential genes.

We may also design some efficient therapeutic strategies to avoid negative effects arising from traditional ‘one gene, one target, one disease’ paradigm. By analyzing biological networks under pathological states, we may propose sophisticated and personalized strategies, such as simultaneously enhance, reduce, or even totally inhibit the activities of a group of genes, proteins, and metabolites to optimize therapeutic effects.

The concept of synthetic lethality has also been widely accepted in the context of cancer therapy [204]. For example, experimental studies indicated that BRCA1 and BRCA2 are key genes involved in double-strand break repair and inhibition of Poly (ADP-ribose) polymerase (PARP) (forming synthetic lethal pairs with BRCA1/2), an important enzyme involved in base excision repair, can lead tumor cells carrying BRCA1 or BRCA2 deficient to apoptosis while maintaining the activities of normal cells [205]. This emphasizes the significance of gaining a comprehensive understanding on synthetic lethality, especially in eukaryotes such as human beings. However, evolutionary analysis of essential genes suggested that they are not conserved in eukaryotes [119, 120]. For example, some non-essential genes in yeast had become essential in mouse. Biological network rewiring is considered to be an important mechanism for gene essentiality change during evolution [206]. One may expect that essentiality studies from the perspective of network biology can offer some novel findings on how synthetic lethality changes across species.

In this thesis, our analyses are limited to understand the cause and evolution of essentiality in model organisms, such as *E. coli* and yeast. A more comprehensive description of topological features and functional characteristics that tightly associated with genes that are essential for cancer cell survival and proliferation might shed light on cancer therapy [207]. Genome-wide screenings on short-hairpin RNA (shRNA) have identified genes indispensable for cancer cell growth [208]. Considering the promising results shown in our works, integrating network

analysis into genome-wide shRNA screens might offer new insights on the cause of gene essentiality in cancer cells which may serve as basis for drug development.

Bibliography

1. Kircher, M. and J. Kelso, *High-throughput DNA sequencing--concepts and limitations*. *Bioessays*, 2010. **32**(6): p. 524-36.
2. Hertzberg, R.P. and A.J. Pope, *High-throughput screening: new technology for the 21st century*. *Curr Opin Chem Biol*, 2000. **4**(4): p. 445-51.
3. Sobek, J., et al., *Microarray technology as a universal tool for high-throughput analysis of biological systems*. *Comb Chem High Throughput Screen*, 2006. **9**(5): p. 365-80.
4. Wang, Z., M. Gerstein, and M. Snyder, *RNA-Seq: a revolutionary tool for transcriptomics*. *Nat Rev Genet*, 2009. **10**(1): p. 57-63.
5. Shendure, J. and H. Ji, *Next-generation DNA sequencing*. *Nat Biotechnol*, 2008. **26**(10): p. 1135-45.
6. Lashkari, D.A., et al., *Yeast microarrays for genome wide parallel genetic and gene expression analysis*. *Proc Natl Acad Sci U S A*, 1997. **94**(24): p. 13057-62.
7. Cookson, W., et al., *Mapping complex disease traits with global gene expression*. *Nat Rev Genet*, 2009. **10**(3): p. 184-94.
8. Lewis, N.E., et al., *Gene expression profiling and the use of genome-scale in silico models of Escherichia coli for analysis: providing context for content*. *J Bacteriol*, 2009. **191**(11): p. 3437-44.
9. Raghavan, R., A. Sage, and H. Ochman, *Genome-wide identification of transcription start sites yields a novel thermosensing RNA and new cyclic*

- AMP receptor protein-regulated genes in Escherichia coli.* J Bacteriol, 2011. **193**(11): p. 2871-4.
10. David, L., et al., *A high-resolution map of transcription in the yeast genome.* Proc Natl Acad Sci U S A, 2006. **103**(14): p. 5320-5.
 11. Hesselberth, J.R., et al., *Global mapping of protein-DNA interactions in vivo by digital genomic footprinting.* Nat Methods, 2009. **6**(4): p. 283-9.
 12. Kim, T.H. and B. Ren, *Genome-wide analysis of protein-DNA interactions.* Annu Rev Genomics Hum Genet, 2006. **7**: p. 81-102.
 13. Plaimas, K., et al., *Machine learning based analyses on metabolic networks supports high-throughput knockout screens.* BMC Syst Biol, 2008. **2**: p. 67.
 14. Plaimas, K., R. Eils, and R. König, *Identifying essential genes in bacterial metabolic networks with machine learning methods.* BMC Syst Biol, 2010. **4**: p. 56.
 15. Feist, A.M., et al., *Reconstruction of biochemical networks in microorganisms.* Nat Rev Microbiol, 2009. **7**(2): p. 129-43.
 16. Henry, C.S., et al., *iBsu1103: a new genome-scale metabolic model of Bacillus subtilis based on SEED annotations.* Genome Biol, 2009. **10**(6): p. R69.
 17. Feist, A.M., et al., *A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information.* Mol Syst Biol, 2007. **3**: p. 121.

18. Mahadevan, R., et al., *Characterization of metabolism in the Fe(III)-reducing organism Geobacter sulfurreducens by constraint-based modeling*. Appl Environ Microbiol, 2006. **72**(2): p. 1558-68.
19. Schilling, C.H. and B.O. Palsson, *Assessment of the metabolic capabilities of Haemophilus influenzae Rd through a genome-scale pathway analysis*. J Theor Biol, 2000. **203**(3): p. 249-83.
20. Thiele, I., et al., *Expanded metabolic reconstruction of Helicobacter pylori (iIT341 GSM/GPR): an in silico genome-scale characterization of single- and double-deletion mutants*. J Bacteriol, 2005. **187**(16): p. 5818-30.
21. Oliveira, A.P., J. Nielsen, and J. Forster, *Modeling Lactococcus lactis using a genome-scale flux model*. BMC Microbiol, 2005. **5**: p. 39.
22. Teusink, B., et al., *Analysis of growth of Lactobacillus plantarum WCFS1 on a complex medium using a genome-scale metabolic model*. J Biol Chem, 2006. **281**(52): p. 40041-8.
23. Kim, T.Y., et al., *Genome-scale analysis of Mannheimia succiniciproducens metabolism*. Biotechnol Bioeng, 2007. **97**(4): p. 657-71.
24. Beste, D.J., et al., *GSMN-TB: a web-based genome-scale network model of Mycobacterium tuberculosis metabolism*. Genome Biol, 2007. **8**(5): p. R89.
25. Suthers, P.F., et al., *A genome-scale metabolic reconstruction of Mycoplasma genitalium, iPS189*. PLoS Comput Biol, 2009. **5**(2): p. e1000285.
26. Baart, G.J., et al., *Modeling Neisseria meningitidis metabolism: from genome to metabolic fluxes*. Genome Biol, 2007. **8**(7): p. R136.

27. Oberhardt, M.A., et al., *Genome-scale metabolic network analysis of the opportunistic pathogen Pseudomonas aeruginosa PAO1*. J Bacteriol, 2008. **190**(8): p. 2790-803.
28. Resendis-Antonio, O., et al., *Metabolic reconstruction and modeling of nitrogen fixation in Rhizobium etli*. PLoS Comput Biol, 2007. **3**(10): p. 1887-95.
29. Heinemann, M., et al., *In silico genome-scale reconstruction and validation of the Staphylococcus aureus metabolic network*. Biotechnol Bioeng, 2005. **92**(7): p. 850-64.
30. Alam, M.T., et al., *Metabolic modeling and analysis of the metabolic switch in Streptomyces coelicolor*. BMC Genomics, 2010. **11**: p. 202.
31. Higashino, K., et al., *Preservation of C7 spinous process does not influence the long-term outcome after laminoplasty for cervical spondylotic myelopathy*. Int Orthop, 2006. **30**(5): p. 362-5.
32. Gonzalez, O., et al., *Reconstruction, modeling & analysis of Halobacterium salinarum R-1 metabolism*. Mol Biosyst, 2008. **4**(2): p. 148-59.
33. Duarte, N.C., et al., *Global reconstruction of the human metabolic network based on genomic and bibliomic data*. Proc Natl Acad Sci U S A, 2007. **104**(6): p. 1777-82.
34. Sigurdsson, M.I., et al., *A detailed genome-wide reconstruction of mouse metabolism based on human Recon 1*. BMC Syst Biol, 2010. **4**: p. 140.

35. Mo, M.L., B.O. Palsson, and M.J. Herrgard, *Connecting extracellular metabolomic measurements to intracellular flux states in yeast*. BMC Syst Biol, 2009. **3**: p. 37.
36. Barabasi, A.L. and Z.N. Oltvai, *Network biology: understanding the cell's functional organization*. Nat Rev Genet, 2004. **5**(2): p. 101-13.
37. Barabasi, A.L., *Scale-free networks: a decade and beyond*. Science, 2009. **325**(5939): p. 412-3.
38. Li, S., et al., *A map of the interactome network of the metazoan C. elegans*. Science, 2004. **303**(5657): p. 540-3.
39. Albert, R., H. Jeong, and A.-L. Barabási, *Diameter of the World-Wide Web*. Nature, 1999. **401**: p. 130-131.
40. Giot, L., et al., *A protein interaction map of Drosophila melanogaster*. Science, 2003. **302**(5651): p. 1727-36.
41. Barabasi, A.L. and R. Albert, *Emergence of scaling in random networks*. Science, 1999. **286**(5439): p. 509-12.
42. Albert, R., H. Jeong, and A.L. Barabasi, *Error and attack tolerance of complex networks*. Nature, 2000. **406**(6794): p. 378-82.
43. Pastor-Satorras, R. and A. Vespignani, *Epidemic spreading in scale-free networks*. Phys Rev Lett, 2001. **86**(14): p. 3200-3.
44. Pastor-Satorras, R. and A. Vespignani, *Immunization of complex networks*. Phys Rev E Stat Nonlin Soft Matter Phys, 2002. **65**(3 Pt 2A): p. 036104.
45. Albert, R. and A.-L. Barabási, *Statistical mechanics of complex networks*. Reviews of Modern Physics, 2002. **74**(1): p. 47-97.

46. Folger, O., et al., *Predicting selective drug targets in cancer through metabolic networks*. Mol Syst Biol, 2011. **7**: p. 501.
47. Guimera, R., M. Sales-Pardo, and L.A. Amaral, *A network-based method for target selection in metabolic networks*. Bioinformatics, 2007. **23**(13): p. 1616-22.
48. Sridhar, P., et al., *Mining metabolic networks for optimal drug targets*. Pac Symp Biocomput, 2008: p. 291-302.
49. Ravasz, E., et al., *Hierarchical organization of modularity in metabolic networks*. Science, 2002. **297**(5586): p. 1551-5.
50. Segre, D., et al., *Modular epistasis in yeast metabolism*. Nat Genet, 2005. **37**(1): p. 77-83.
51. Mazurie, A., et al., *Evolution of metabolic network organization*. BMC Syst Biol, 2010. **4**: p. 59.
52. Gat-Viks, I., A. Tanay, and R. Shamir, *Modeling and analysis of heterogeneous regulation in biological networks*. J Comput Biol, 2004. **11**(6): p. 1034-49.
53. Arrell, D.K. and A. Terzic, *Network systems biology for drug discovery*. Clin Pharmacol Ther, 2010. **88**(1): p. 120-5.
54. Pujol, A., et al., *Unveiling the role of network and systems biology in drug discovery*. Trends Pharmacol Sci, 2010. **31**(3): p. 115-23.
55. Kunkel, E.J., et al., *An integrative biology approach for analysis of drug action in models of human vascular inflammation*. FASEB J, 2004. **18**(11): p. 1279-81.

56. Lorimer, I.A., *Mutant epidermal growth factor receptors as targets for cancer therapy*. *Curr Cancer Drug Targets*, 2002. **2**(2): p. 91-102.
57. Primiano, T., et al., *Identification of potential anticancer drug targets through the selection of growth-inhibitory genetic suppressor elements*. *Cancer Cell*, 2003. **4**(1): p. 41-53.
58. Plump, A.S. and P.Y. Lum, *Genomics and cardiovascular drug development*. *J Am Coll Cardiol*, 2009. **53**(13): p. 1089-100.
59. Mayr, L.M. and D. Bojanic, *Novel trends in high-throughput screening*. *Curr Opin Pharmacol*, 2009. **9**(5): p. 580-8.
60. Koehn, F.E., *High impact technologies for natural products screening*. *Prog Drug Res*, 2008. **65**: p. 175, 177-210.
61. Langtry, H.D. and A. Markham, *Sildenafil: a review of its use in erectile dysfunction*. *Drugs*, 1999. **57**(6): p. 967-89.
62. Wilde, M.I. and D. McTavish, *Tamsulosin. A review of its pharmacological properties and therapeutic potential in the management of symptomatic benign prostatic hyperplasia*. *Drugs*, 1996. **52**(6): p. 883-98.
63. Hornberg, J.J., et al., *Cancer: a Systems Biology disease*. *Biosystems*, 2006. **83**(2-3): p. 81-90.
64. Sams-Dodd, F., *Target-based drug discovery: is something wrong?* *Drug Discov Today*, 2005. **10**(2): p. 139-47.
65. Morphy, R., C. Kay, and Z. Rankovic, *From magic bullets to designed multiple ligands*. *Drug Discov Today*, 2004. **9**(15): p. 641-51.

66. Baba, T., et al., *Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection*. Mol Syst Biol, 2006. **2**: p. 2006 0008.
67. Yildirim, M.A., et al., *Drug-target network*. Nat Biotechnol, 2007. **25**(10): p. 1119-26.
68. Iglehart, J.D. and D.P. Silver, *Synthetic lethality--a new direction in cancer-drug development*. N Engl J Med, 2009. **361**(2): p. 189-91.
69. Motter, A.E., *Improved network performance via antagonism: From synthetic rescues to multi-drug combinations*. Bioessays, 2010. **32**(3): p. 236-45.
70. Hopkins, A.L., *Network pharmacology: the next paradigm in drug discovery*. Nat Chem Biol, 2008. **4**(11): p. 682-90.
71. Nelander, S., et al., *Models from experiments: combinatorial drug perturbations of cancer cells*. Mol Syst Biol, 2008. **4**: p. 216.
72. Csermely, P., V. Agoston, and S. Pongor, *The efficiency of multi-target drugs: the network approach might help drug design*. Trends Pharmacol Sci, 2005. **26**(4): p. 178-82.
73. Orth, J.D., I. Thiele, and B.O. Palsson, *What is flux balance analysis?* Nat Biotechnol, 2010. **28**(3): p. 245-8.
74. Smart, A.G., L.A. Amaral, and J.M. Ottino, *Cascading failure and robustness in metabolic networks*. Proc Natl Acad Sci U S A, 2008. **105**(36): p. 13223-8.
75. Bloom, B.R. and C.J. Murray, *Tuberculosis: commentary on a reemergent killer*. Science, 1992. **257**(5073): p. 1055-64.

76. Appelbaum, P.C., *Antimicrobial resistance in Streptococcus pneumoniae: an overview*. Clin Infect Dis, 1992. **15**(1): p. 77-83.
77. Lange, R.P., et al., *The targets of currently used antibacterial agents: lessons for drug discovery*. Curr Pharm Des, 2007. **13**(30): p. 3140-54.
78. Gerdes, S., et al., *Essential genes on metabolic maps*. Curr Opin Biotechnol, 2006. **17**(5): p. 448-56.
79. Simon, N.M., et al., *Pneumatoxis cystoides intestinalis. Treatment with oxygen via close-fitting mask*. JAMA, 1975. **231**(13): p. 1354-6.
80. Burrus, V. and M.K. Waldor, *Shaping bacterial genomes with integrative and conjugative elements*. Res Microbiol, 2004. **155**(5): p. 376-86.
81. Aminetzach, Y.T., J.M. Macpherson, and D.A. Petrov, *Pesticide resistance via transposition-mediated adaptive gene truncation in Drosophila*. Science, 2005. **309**(5735): p. 764-7.
82. Hastings, P.J., et al., *Mechanisms of change in gene copy number*. Nat Rev Genet, 2009. **10**(8): p. 551-64.
83. Eyre-Walker, A. and P.D. Keightley, *The distribution of fitness effects of new mutations*. Nat Rev Genet, 2007. **8**(8): p. 610-8.
84. S. Carroll, J.G., S. Weatherbee, *From DNA to diversity: molecular genetics and the evolution of animal design*. 2004.
85. Clancy, S., *Genetic Mutation*. Nature Education, 2008. **1**(1).
86. Antoniou, A., et al., *Average risks of breast and ovarian cancer associated with BRCA1 or BRCA2 mutations detected in case Series unselected for family history: a combined analysis of 22 studies*. Am J Hum Genet, 2003. **72**(5): p. 1117-30.

87. Kettleborough, R.N., et al., *A systematic genome-wide analysis of zebrafish protein-coding gene function*. Nature, 2013. **496**(7446): p. 494-7.
88. Giaever, G., et al., *Functional profiling of the *Saccharomyces cerevisiae* genome*. Nature, 2002. **418**(6896): p. 387-91.
89. Winzeler, E.A., et al., *Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis*. Science, 1999. **285**(5429): p. 901-6.
90. Motter, A.E., et al., *Predicting synthetic rescues in metabolic networks*. Mol Syst Biol, 2008. **4**: p. 168.
91. Gerdes, S.Y., et al., *Experimental determination and system level analysis of essential genes in *Escherichia coli* MG1655*. J Bacteriol, 2003. **185**(19): p. 5673-84.
92. Joyce, A.R., et al., *Experimental and computational assessment of conditionally essential genes in *Escherichia coli**. J Bacteriol, 2006. **188**(23): p. 8259-71.
93. Zhang, R. and Y. Lin, *DEG 5.0, a database of essential genes in both prokaryotes and eukaryotes*. Nucleic Acids Res, 2009. **37**(Database issue): p. D455-8.
94. Kato, J. and M. Hashimoto, *Construction of consecutive deletions of the *Escherichia coli* chromosome*. Mol Syst Biol, 2007. **3**: p. 132.
95. Raman, K. and N. Chandra, *Flux balance analysis of biological systems: applications and challenges*. Brief Bioinform, 2009. **10**(4): p. 435-49.

96. Segre, D., D. Vitkup, and G.M. Church, *Analysis of optimality in natural and perturbed metabolic networks*. Proc Natl Acad Sci U S A, 2002. **99**(23): p. 15112-7.
97. Wunderlich, Z. and L.A. Mirny, *Using the topology of metabolic networks to predict viability of mutant strains*. Biophys J, 2006. **91**(6): p. 2304-11.
98. Seringhaus, M., et al., *Predicting essential genes in fungal genomes*. Genome Res, 2006. **16**(9): p. 1126-35.
99. de Berardinis, V., et al., *A complete collection of single-gene deletion mutants of Acinetobacter baylyi ADPI*. Mol Syst Biol, 2008. **4**: p. 174.
100. Lemke, N., et al., *Essentiality and damage in metabolic networks*. Bioinformatics, 2004. **20**(1): p. 115-9.
101. Almaas, E., et al., *Global organization of metabolic fluxes in the bacterium Escherichia coli*. Nature, 2004. **427**(6977): p. 839-43.
102. Edwards, J.S. and B.O. Palsson, *Robustness analysis of the Escherichia coli metabolic network*. Biotechnol Prog, 2000. **16**(6): p. 927-39.
103. Papp, B., C. Pal, and L.D. Hurst, *Metabolic network analysis of the causes and evolution of enzyme dispensability in yeast*. Nature, 2004. **429**(6992): p. 661-4.
104. Tong, X., et al., *Genome-scale identification of conditionally essential genes in E. coli by DNA microarrays*. Biochem Biophys Res Commun, 2004. **322**(1): p. 347-54.
105. Sasseti, C.M., D.H. Boyd, and E.J. Rubin, *Comprehensive identification of conditionally essential genes in mycobacteria*. Proc Natl Acad Sci U S A, 2001. **98**(22): p. 12712-7.

106. Gu, Z., et al., *Role of duplicate genes in genetic robustness against null mutations*. Nature, 2003. **421**(6918): p. 63-6.
107. Wagner, A., *Robustness against mutations in genetic networks of yeast*. Nat Genet, 2000. **24**(4): p. 355-61.
108. Deutscher, D., et al., *Multiple knockout analysis of genetic robustness in the yeast metabolic network*. Nat Genet, 2006. **38**(9): p. 993-8.
109. Kim, P.J., et al., *Metabolite essentiality elucidates robustness of Escherichia coli metabolism*. Proc Natl Acad Sci U S A, 2007. **104**(34): p. 13638-42.
110. Samal, A., et al., *Low degree metabolites explain essential reactions and enhance modularity in biological networks*. BMC Bioinformatics, 2006. **7**: p. 118.
111. Stelling, J., et al., *Metabolic network structure determines key aspects of functionality and regulation*. Nature, 2002. **420**(6912): p. 190-3.
112. Jordan, I.K., et al., *Essential genes are more evolutionarily conserved than are nonessential genes in bacteria*. Genome Res, 2002. **12**(6): p. 962-8.
113. Dotsch, A., et al., *Evolutionary conservation of essential and highly expressed genes in Pseudomonas aeruginosa*. BMC Genomics, 2010. **11**: p. 234.
114. Forsyth, R.A., et al., *A genome-wide strategy for the identification of essential genes in Staphylococcus aureus*. Mol Microbiol, 2002. **43**(6): p. 1387-400.

115. Freiberg, C., et al., *Identification of novel essential Escherichia coli genes conserved among pathogenic bacteria*. J Mol Microbiol Biotechnol, 2001. **3**(3): p. 483-9.
116. Duffield, M., et al., *Predicting conserved essential genes in bacteria: in silico identification of putative drug targets*. Mol Biosyst, 2010. **6**(12): p. 2482-9.
117. Bergmiller, T., M. Ackermann, and O.K. Silander, *Patterns of evolutionary conservation of essential genes correlate with their compensability*. PLoS Genet, 2012. **8**(6): p. e1002803.
118. Silander, O.K. and M. Ackermann, *The constancy of gene conservation across divergent bacterial orders*. BMC Res Notes, 2009. **2**: p. 2.
119. Hirsh, A.E. and H.B. Fraser, *Protein dispensability and rate of evolution*. Nature, 2001. **411**(6841): p. 1046-9.
120. Hurst, L.D. and N.G. Smith, *Do essential genes evolve slowly?* Curr Biol, 1999. **9**(14): p. 747-50.
121. Moore, J.H., *A global view of epistasis*. Nat Genet, 2005. **37**(1): p. 13-4.
122. Cordell, H.J., *Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans*. Hum Mol Genet, 2002. **11**(20): p. 2463-8.
123. Boone, C., H. Bussey, and B.J. Andrews, *Exploring genetic interactions and networks with yeast*. Nat Rev Genet, 2007. **8**(6): p. 437-49.
124. Kaelin, W.G., Jr., *Synthetic lethality: a framework for the development of wiser cancer therapeutics*. Genome Med, 2009. **1**(10): p. 99.

125. Suthers, P.F., A. Zomorodi, and C.D. Maranas, *Genome-scale gene/reaction essentiality and synthetic lethality analysis*. Mol Syst Biol, 2009. **5**: p. 301.
126. Tong, A.H., et al., *Systematic genetic analysis with ordered arrays of yeast deletion mutants*. Science, 2001. **294**(5550): p. 2364-8.
127. Ooi, S.L., D.D. Shoemaker, and J.D. Boeke, *DNA helicase gene interaction network defined using synthetic lethality analyzed by microarray*. Nat Genet, 2003. **35**(3): p. 277-86.
128. Tong, A.H., et al., *Global mapping of the yeast genetic interaction network*. Science, 2004. **303**(5659): p. 808-13.
129. Jasnos, L. and R. Korona, *Epistatic buffering of fitness loss in yeast double deletion strains*. Nat Genet, 2007. **39**(4): p. 550-4.
130. He, X., et al., *Prevalent positive epistasis in Escherichia coli and Saccharomyces cerevisiae metabolic networks*. Nat Genet, 2010. **42**(3): p. 272-6.
131. Ma, X., A.M. Tarone, and W. Li, *Mapping genetically compensatory pathways from synthetic lethal interactions in yeast*. PLoS One, 2008. **3**(4): p. e1922.
132. Tischler, J., B. Lehner, and A.G. Fraser, *Evolutionary plasticity of genetic interaction networks*. Nat Genet, 2008. **40**(4): p. 390-1.
133. Dixon, S.J., et al., *Significant conservation of synthetic lethal genetic interaction networks between distantly related eukaryotes*. Proc Natl Acad Sci U S A, 2008. **105**(43): p. 16653-8.

134. Kanehisa, M., et al., *KEGG for integration and interpretation of large-scale molecular data sets*. Nucleic Acids Res, 2012. **40**(Database issue): p. D109-14.
135. Ma, H. and A.P. Zeng, *Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms*. Bioinformatics, 2003. **19**(2): p. 270-7.
136. Schellenberger, J., et al., *BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions*. BMC Bioinformatics, 2010. **11**: p. 213.
137. Thiele, I. and B.O. Palsson, *A protocol for generating a high-quality genome-scale metabolic reconstruction*. Nat Protoc, 2010. **5**(1): p. 93-121.
138. Duarte, N.C., M.J. Herrgard, and B.O. Palsson, *Reconstruction and validation of *Saccharomyces cerevisiae* iND750, a fully compartmentalized genome-scale metabolic model*. Genome Res, 2004. **14**(7): p. 1298-309.
139. Huss, M. and P. Holme, *Currency and commodity metabolites: their identification and relation to the modularity of metabolic networks*. IET Syst Biol, 2007. **1**(5): p. 280-5.
140. Li, Y., et al., *Metabolic pathway alignment between species using a comprehensive and flexible similarity measure*. BMC Syst Biol, 2008. **2**: p. 111.
141. Shannon, P., et al., *Cytoscape: a software environment for integrated models of biomolecular interaction networks*. Genome Res, 2003. **13**(11): p. 2498-504.

142. Lee, J.M., E.P. Gianchandani, and J.A. Papin, *Flux balance analysis in the era of metabolomics*. *Brief Bioinform*, 2006. **7**(2): p. 140-50.
143. Edwards, J.S., R.U. Ibarra, and B.O. Palsson, *In silico predictions of Escherichia coli metabolic capabilities are consistent with experimental data*. *Nat Biotechnol*, 2001. **19**(2): p. 125-30.
144. Baumlér, D.J., et al., *The evolution of metabolic networks of E. coli*. *BMC Syst Biol*, 2011. **5**: p. 182.
145. Matias Rodrigues, J.F. and A. Wagner, *Evolutionary plasticity and innovations in complex metabolic reaction networks*. *PLoS Comput Biol*, 2009. **5**(12): p. e1000613.
146. Burgard, A.P. and C.D. Maranas, *Probing the performance limits of the Escherichia coli metabolic network subject to gene additions or deletions*. *Biotechnol Bioeng*, 2001. **74**(5): p. 364-75.
147. Oh, Y.K., et al., *Genome-scale reconstruction of metabolic network in Bacillus subtilis based on high-throughput phenotyping and gene essentiality data*. *J Biol Chem*, 2007. **282**(39): p. 28791-9.
148. Holzhütter, H.G., *The generalized flux-minimization method and its application to metabolic networks affected by enzyme deficiencies*. *Biosystems*, 2006. **83**(2-3): p. 98-107.
149. Park, J.M., T.Y. Kim, and S.Y. Lee, *Prediction of metabolic fluxes by incorporating genomic context and flux-converging pattern analyses*. *Proc Natl Acad Sci U S A*, 2010. **107**(33): p. 14931-6.
150. Klipp, E., et al., *Integrative model of the response of yeast to osmotic shock*. *Nat Biotechnol*, 2005. **23**(8): p. 975-82.

151. Covert, M.W., et al., *Integrating metabolic, transcriptional regulatory and signal transduction models in Escherichia coli*. Bioinformatics, 2008. **24**(18): p. 2044-50.
152. Lee, J.M., et al., *Dynamic analysis of integrated signaling, metabolic, and regulatory networks*. PLoS Comput Biol, 2008. **4**(5): p. e1000086.
153. Reed, J.L., et al., *An expanded genome-scale model of Escherichia coli K-12 (iJR904 GSM/GPR)*. Genome Biol, 2003. **4**(9): p. R54.
154. Kitano, H., *Biological robustness*. Nat Rev Genet, 2004. **5**(11): p. 826-37.
155. Hu, W., et al., *Essential gene identification and drug target prioritization in Aspergillus fumigatus*. PLoS Pathog, 2007. **3**(3): p. e24.
156. Buysse, J.M., *The role of genomics in antibacterial target discovery*. Curr Med Chem, 2001. **8**(14): p. 1713-26.
157. Chalker, A.F. and R.D. Lunsford, *Rational identification of new antibacterial drug targets that are essential for viability using a genomics-based approach*. Pharmacol Ther, 2002. **95**(1): p. 1-20.
158. Hopkins, A.L. and C.R. Groom, *The druggable genome*. Nat Rev Drug Discov, 2002. **1**(9): p. 727-30.
159. Becker, S.A., et al., *Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox*. Nat Protoc, 2007. **2**(3): p. 727-38.
160. Martelli, C., et al., *Identifying essential genes in Escherichia coli from a metabolic optimization principle*. Proc Natl Acad Sci U S A, 2009. **106**(8): p. 2607-11.

161. Joyce, A.R. and B.O. Palsson, *Predicting gene essentiality using genome-scale in silico models*. Methods Mol Biol, 2008. **416**: p. 433-57.
162. Whelan, K.E. and R.D. King, *Using a logical model to predict the growth of yeast*. BMC Bioinformatics, 2008. **9**: p. 97.
163. Metris, A., et al., *In vivo and in silico determination of essential genes of Campylobacter jejuni*. BMC Genomics, 2011. **12**: p. 535.
164. Zhao, J., et al., *Global metabolic response of Escherichia coli to gnd or zwf gene-knockout, based on 13C-labeling experiments and the measurement of enzyme activities*. Appl Microbiol Biotechnol, 2004. **64**(1): p. 91-8.
165. Kumar, R. and K. Shimizu, *Transcriptional regulation of main metabolic pathways of cyoA, cydB, fnr, and fur gene knockout Escherichia coli in C-limited and N-limited aerobic continuous cultures*. Microb Cell Fact, 2011. **10**: p. 3.
166. Jaccard, P., *Nouvelle recherche sur la distribution florale*. Bulletin de la Societe Vaudoise des Sciences Naturelles, 1908. **44**: p. 223-270.
167. Onishi, H.R., et al., *Antibacterial agents that inhibit lipid A biosynthesis*. Science, 1996. **274**(5289): p. 980-2.
168. Ling, J., N. Reynolds, and M. Ibba, *Aminoacyl-tRNA synthesis and translational quality control*. Annu Rev Microbiol, 2009. **63**: p. 61-78.
169. Bian, J., et al., *The riboswitch regulates a thiamine pyrophosphate ABC transporter of the oral spirochete Treponema denticola*. J Bacteriol, 2011. **193**(15): p. 3912-22.

170. Gong, L., K. Takayama, and S. Kjelleberg, *Role of spoT-dependent ppGpp accumulation in the survival of light-exposed starved bacteria*. Microbiology, 2002. **148**(Pt 2): p. 559-70.
171. Poon, W.W., et al., *Identification of Escherichia coli ubiB, a gene required for the first monooxygenase step in ubiquinone biosynthesis*. J Bacteriol, 2000. **182**(18): p. 5139-46.
172. Martin, V.J., et al., *Engineering a mevalonate pathway in Escherichia coli for production of terpenoids*. Nat Biotechnol, 2003. **21**(7): p. 796-802.
173. Kuzuyama, T., S. Takahashi, and H. Seto, *Construction and characterization of Escherichia coli disruptants defective in the yaeM gene*. Biosci Biotechnol Biochem, 1999. **63**(4): p. 776-8.
174. Battersby, A.R., et al., *Biosynthesis of the pigments of life: formation of the macrocycle*. Nature, 1980. **285**(5759): p. 17-21.
175. Nihei, C., et al., *Abortive assembly of succinate-ubiquinone reductase (complex II) in a ferrenchelatase-deficient mutant of Escherichia coli*. Mol Genet Genomics, 2001. **265**(3): p. 394-404.
176. Inada, T. and Y. Nakamura, *Lethal double-stranded RNA processing activity of ribonuclease III in the absence of suhB protein of Escherichia coli*. Biochimie, 1995. **77**(4): p. 294-302.
177. Keseler, I.M., et al., *EcoCyc: a comprehensive database of Escherichia coli biology*. Nucleic Acids Res, 2011. **39**(Database issue): p. D583-90.
178. Kim, C., S. Song, and C. Park, *The D-allose operon of Escherichia coli K-12*. J Bacteriol, 1997. **179**(24): p. 7631-7.

179. Torti, S.V. and J.T. Park, *Lipoprotein of gram-negative bacteria is essential for growth and division*. Nature, 1976. **263**(5575): p. 323-6.
180. Phadtare, S., et al., *Analysis of Escherichia coli global gene expression profiles in response to overexpression and deletion of CspC and CspE*. J Bacteriol, 2006. **188**(7): p. 2521-7.
181. Robbins-Manke, J.L., et al., *Analysis of global gene expression and double-strand-break formation in DNA adenine methyltransferase- and mismatch repair-deficient Escherichia coli*. J Bacteriol, 2005. **187**(20): p. 7027-37.
182. Kabir, M.M. and K. Shimizu, *Gene expression patterns for metabolic pathway in pgi knockout Escherichia coli with and without phb genes based on RT-PCR*. J Biotechnol, 2003. **105**(1-2): p. 11-31.
183. D'Elia, M.A., M.P. Pereira, and E.D. Brown, *Are essential genes really essential?* Trends Microbiol, 2009. **17**(10): p. 433-8.
184. Davierwala, A.P., et al., *The synthetic genetic interaction spectrum of essential genes*. Nat Genet, 2005. **37**(10): p. 1147-52.
185. Phillips, P.C., *Epistasis--the essential role of gene interactions in the structure and evolution of genetic systems*. Nat Rev Genet, 2008. **9**(11): p. 855-67.
186. Yu, L., et al., *A survey of essential gene function in the yeast cell division cycle*. Mol Biol Cell, 2006. **17**(11): p. 4736-47.
187. Li, Z., et al., *Systematic exploration of essential yeast gene function with temperature-sensitive mutants*. Nat Biotechnol, 2011. **29**(4): p. 361-7.

188. Juhas, M., L. Eberl, and J.I. Glass, *Essence of life: essential genes of minimal genomes*. Trends Cell Biol, 2011. **21**(10): p. 562-8.
189. Haselbeck, R., et al., *Comprehensive essential gene identification as a platform for novel anti-infective drug discovery*. Curr Pharm Des, 2002. **8**(13): p. 1155-72.
190. Salama, N.R., B. Shepherd, and S. Falkow, *Global transposon mutagenesis and essential gene analysis of Helicobacter pylori*. J Bacteriol, 2004. **186**(23): p. 7926-35.
191. Chaudhuri, R.R., et al., *Comprehensive identification of essential Staphylococcus aureus genes using Transposon-Mediated Differential Hybridisation (TMDH)*. BMC Genomics, 2009. **10**: p. 291.
192. Sakamoto, H., et al., *Towards systematic identification of Plasmodium essential genes by transposon shuttle mutagenesis*. Nucleic Acids Res, 2005. **33**(20): p. e174.
193. Hwang, Y.C., et al., *Predicting essential genes based on network and sequence analysis*. Mol Biosyst, 2009. **5**(12): p. 1672-8.
194. Gustafson, A.M., et al., *Towards the identification of essential genes using targeted genome sequencing and comparative analysis*. BMC Genomics, 2006. **7**: p. 265.
195. D'Cunha, J., *Exploring the optimal treatment strategy for surgically resected T4 esophageal tumors*. J Surg Oncol, 2012. **105**(8): p. 741-2.
196. Xu, P., et al., *Genome-wide essential gene identification in Streptococcus sanguinis*. Sci Rep, 2011. **1**: p. 125.

197. Sauer, U., *Metabolic networks in motion: 13C-based flux analysis*. Mol Syst Biol, 2006. **2**: p. 62.
198. Acencio, M.L. and N. Lemke, *Towards the prediction of essential genes by integration of network topology, cellular localization and biological process information*. BMC Bioinformatics, 2009. **10**: p. 290.
199. Sprossmann, F., et al., *Inducible knockout mutagenesis reveals compensatory mechanisms elicited by constitutive BK channel deficiency in overactive murine bladder*. FEBS J, 2009. **276**(6): p. 1680-97.
200. Dixon, S.J., et al., *Systematic mapping of genetic interaction networks*. Annu Rev Genet, 2009. **43**: p. 601-25.
201. Ma, J., et al., *Metabolic network analysis revealed distinct routes of deletion effects between essential and non-essential genes*. Mol Biosyst, 2012. **8**(4): p. 1179-86.
202. Mossesso, E. and C.D. Lima, *Ulp1-SUMO crystal structure and genetic analysis reveal conserved interactions and a regulatory element essential for cell growth in yeast*. Mol Cell, 2000. **5**(5): p. 865-76.
203. Ashburner, M., et al., *Gene ontology: tool for the unification of biology*. The Gene Ontology Consortium. Nat Genet, 2000. **25**(1): p. 25-9.
204. Kaelin, W.G., Jr., *The concept of synthetic lethality in the context of anticancer therapy*. Nat Rev Cancer, 2005. **5**(9): p. 689-98.
205. Farmer, H., et al., *Targeting the DNA repair defect in BRCA mutant cells as a therapeutic strategy*. Nature, 2005. **434**(7035): p. 917-21.
206. Kim, J., et al., *Network rewiring is an important mechanism of gene essentiality change*. Sci Rep, 2012. **2**: p. 900.

-
207. Marcotte, R., et al., *Essential gene profiles in breast, pancreatic, and ovarian cancer cells*. *Cancer Discov*, 2012. **2**(2): p. 172-89.
208. Bernards, R., T.R. Brummelkamp, and R.L. Beijersbergen, *shRNA libraries and their use in cancer genetics*. *Nat Methods*, 2006. **3**(9): p. 701-6.

Appendix 1: C++ Code for identifying corresponding reactions

In the following C++ code, `r2g_map` is extracted from SBML file exported from BiGG database. It stores the gene-reaction association information. Another data structure used is: `reaction_sets`, which stores all the reactions available in this network in a vector. The *i*-th element in `r2g_map` is the gene structure associated with the *i*-th reaction in `reaction_sets`.

```
vector<vector<set<string>>> r2g_map;
vector<string> reaction_sets;
void corresponding_reactions(string gene, set<string>& rt_sets)
{
    for(int i=0; i<r2g_map.size(); i++)
    {
        int count=0;
        if(r2g_map[i].size()!=0)
        {
            for(int j=0; j<r2g_map[i].size(); j++)
            {
                if(r2g_map[i][j].find(gene)!=r2g_map[i][j].end())
                    count++;
            }
            if(count==r2g_map[i].size())
                rt_sets.insert(reaction_sets[i]);
        }
    }
}
```

Appendix 2: Pseudo Code for identifying damage lists for single gene deletion

As the codes for identifying damage lists for single gene deletion is too long, with more than 200 lines, we illustrated the pseudo code for this algorithm.

Input: graph G, corresponding reaction lists CRs

Output: Damage Reaction List DRs

Step1: Identifying the affected reaction lists in response to single gene deletion;

BEGIN

For each reaction in CRs

 IF the reaction is REVERSE

 Retrieve all its neighbors NN;

 Assign each node in NN to NP and NS;

 Insert reaction in DRs;

 Erase reaction node and its links from G;

 Else

 Retrieve all its parent nodes and store them NP;

 Retrieve all its son nodes and store them in NS;

 Insert reaction in DRs;

 Erase reaction node and its links from G;

 END IF

Do

 FOR each metabolite M in NP

 IF the out-degree of M is zero

 Retrieve the parent nodes and store them in R;

 Erase metabolite node and its links from G;

 END IF

 END FOR

 FOR each metabolite M in NS

 IF the in-degree of M is zero

 Retrieve the son nodes and store them in R;

 Erase metabolite node and its links from G;

 END IF

 END FOR

 FOR each reaction node r in R

 Retrieve the parent metabolite nodes and store them in NP;

```
        Retrieve the son metabolite nodes and store them in NS;
        Insert reaction in DRs;
        Erase reaction node and its links from G;
    END FOR
WHILE (NP and ND are non-empty)
END
```

Step2: Identifying the affected gene lists in response to single gene deletion;

Input: Damage Reaction List DRs identified in Step 1;

Output: Damage Gene List DGs;

```
BEGIN
    FOR each reaction in DRs
        IF it is the corresponding reaction of gene GENE
            Insert GENE into DGs;
        END IF
    END FOR
END
```

The obtained DGs is the affected gene lists in response to single gene deletion.