Evaluation of the Prediction of Gene Knockout Effects by Minimal Pathway Enumeration

Takehide Soh[†]
Kobe University
1-1, Rokko-dai, Nada, Kobe
Hyogo, Japan
soh@lion.kobe-u.ac.jp

Katsumi Inoue
National Institute of Informatics
2-1-2, Hitotsubashi, Chiyoda-ku,
Tokyo, Japan
inoue@nii.ac.jp

Tomoya Baba
Transdisciplinary Research
Integration Center
1111 Yata, Mishima,
411-8540, Japan
tobaba@lab.nig.ac.jp

Toyoyuki Takada, Toshihiko Shiroishi
National Institute of Genetics
1111 Yata, Mishima,
411-8540, Japan
{ttakada,tshirois}@lab.nig.ac.jp

***Abstract*—In this paper, we propose a method to predict gene knockout effects for the cell growth by utilizing biological databases such as KEGG and EcoCyc, in which biological knowledge and experimental results have been collected. At first, biological networks are constructed from such databases and configure experimental conditions by giving source metabolites, target metabolites, and knockout genes. All minimal active pathways are then enumerated, which are minimal subsets of a given network using source metabolites to produce target metabolites. Finally, the effects of gene knockouts are predicted by measuring the difference of minimal active pathways between original networks and knockout ones. In the experiments, we applied it to predict the single gene knockout effects on the glycolysis pathway and amino acids biosynthesis in *Escherichia coli*. We also analyze which pathways are important to *Escherichia coli* and predict lethal pairs of knockout genes. In the results, our method predicted three out of four essential genes, which agree with the biological results of the Keio collection containing comprehensive cell growth data. In addition, predicted lethal pairs of genes also agree with the biological results of double gene knockouts.**

Keywords-metabolic pathways; gene knockout; prediction method; minimal pathway; Keio collection.

I. Introduction

This paper is an extended version of the previously published conference paper [1]. While the earlier paper only considered effects of single gene knockouts, this paper proceeds the analysis of those effects and shows the prediction result of double gene knockouts, which is compared with biological results of [2].

Living organisms, such as bacteria, fishes, animals, and humans, are kept alive by a huge number of intracellular chemical reactions. In *systems biology*, interactions of such chemical reactions are represented in a network called a *pathway*. Pathways have been actively researched in the last decade [3]–[5]. In addition, it is a biologically important subject to reveal the function of genes, which affect the phenotype of organisms. For model organisms such as *Escherichia coli* (*E. coli*), it has been approached by various methods. Constructing gene knockout organisms is an example of such methods [6]–[8]. However, it generally involves high costs and is limited by target genes and organisms.

In this paper, we propose a computation method to predict gene knockout effects by identifying *active pathways*, which are sub-pathways that produce target metabolites from source metabolites. We particularly focus on *minimal active pathways* [9], which do not contain any other active pathways. In other words, all elements of each minimal active pathway are qualitatively essential to produce target metabolites. To predict gene knockout effects by the enumeration of minimal active pathways, at first, *extended pathways*, which include relations between enzymatic reactions and genes, are introduced. Then, the problem of finding minimal active pathways on the extended pathway with gene knockouts is formalized. After computing the solution of the problem, our method predicts gene knockout effects by collecting minimal active pathways that are still active under given gene knockouts.

To evaluate our method, *E. coli* is chosen as our target organism, since it has been studied and much information is available on public resources such as KEGG [10] and EcoCyc [11]. The proposed method is applied to predict gene knockout effects on *E. coli* utilizing biological databases KEGG and EcoCyc, in which biological knowledge and experimental results have been collected. In the experiments, we compare our prediction with the cell growth of every single gene knockout *E. coli* strain, which was obtained from the Keio collection [6]. In addition to the above experiments, using the biological results of [2], we analyze which minimal active pathways are important and predict lethal pairs of gene knockouts for *E. coli*. We also apply our method to predict the gene knockout effects on amino acid biosynthesis and discuss which reactions are suspected to be lacked in databases.

This paper is organized as follows. At first, databases used in this paper and our research framework are explained in Section II. The extended pathway is defined in Section III. Then, Section IV formalizes the problem of finding minimal

[†]This work has been done while affiliated with Transdisciplinary Research Integration Center, 2-1-2, Hitotsubashi, Chiyoda-ku, Tokyo, Japan.

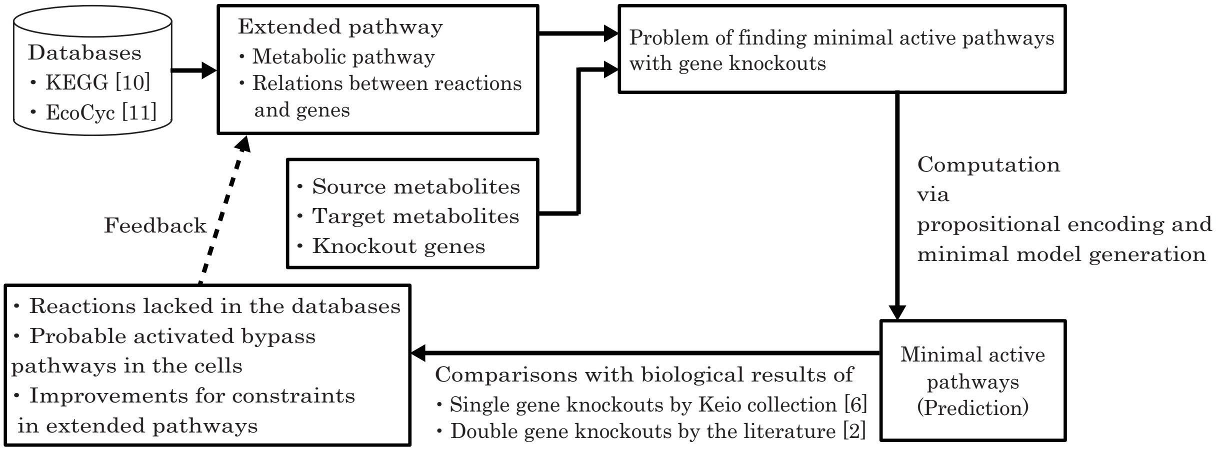

Databases
• KEGG [10]
• EcoCyc [11]
Extended pathway
• Metabolic pathway
• Relations between reactions and genes
Problem of finding minimal active pathways with gene knockouts
• Source metabolites
• Target metabolites
• Knockout genes
Feedback
Computation via propositional encoding and minimal model generation
• Reactions lacked in the databases
• Probable activated bypass pathways in the cells
• Improvements for constraints in extended pathways
Comparisons with biological results of
• Single gene knockouts by Keio collection [6]
• Double gene knockouts by the literature [2]
Minimal active pathways (Prediction)

Figure 1. Used Databases and our Research Framework

active pathways on the extended pathway. The effect of gene knockouts is formalized in Section V. Following that, our computational method is shown in Section VI. In Section VII, computational predictions are compared with results of biological experiments, and we have discussions. Following Section VIII of related work, this paper is concluded in Section IX.

II. Used Databases and Research Framework

This section explains used databases and our research framework shown in Figure 1. In this paper, we particularly focus on *E. coli*. The metabolic pathway has been revealed by biochemical, molecular, and genetic studies, and *E. coli* is the organism in most detail. A large number of *E. coli* studies has contributed to several kinds of biological databases. In particular, we used the following two databases to construct our input network, called an extended pathway.

One is *EcoCyc* [11]. It is a bioinformatics database that describes the genome and the biochemical machinery of *E. coli* K-12 MG1655. The EcoCyc project performs literature-based curation of the entire genome, metabolic pathways, etc. Specifically, it has been doing a literature-based curation from more than 19,000 publications. We constructed metabolic pathways with EcoCyc. The other one is *Kyoto encyclopedia of genes and genomes* (KEGG) [10], which is a database resource that integrates genomic, chemical, and systemic functional information. In particular, gene catalogs in the completely sequenced genomes, from bacteria to humans, are linked to higher-level systemic functions of the cell, the organism, and the ecosystem. A distinguished feature of KEGG is that it provides useful application program interfaces (API). We connected enzymatic reactions of metabolic pathways to genes with this API.

Figure 1 shows our research framework using the two databases. At first, the input network called *extended pathway* is constructed from those databases. Then the problem of finding minimal active pathways is constructed by giving source metabolites, target metabolites, knockout genes, and the extended pathway. Then, we compute minimal active pathways using source metabolites to produce target metabolites. In the case of wild cells, we usually obtain multiple minimal active pathways including bypass pathways. However, in the case of knockout cells, we lose some (or all) of them. In brief, we predict the effects of gene knockouts from how many pathways are lost from the case of wild cells.

To evaluate our prediction method, we usually need additional biological experiments. However, Baba *et al.* comprehensively experimented on the cell growth of every single gene knockout strain [6]. Thanks to this research, we can evaluate our method with comparative ease. We briefly explain this research as follows. The *E. coli* K-12 single gene knockout mutant set, named *Keio collection*, is constructed as a resource for systems biological analyses. Excluding repetitive genes, e.g., insertion sequences related genes, 4288 protein coding genes are targeted for the systematic single gene knockout experiments. Of those, 3985 genes are successfully disrupted, and those of single gene knockout mutants are constructed as the Keio collection. On the other hand, 303 genes are not disrupted and they are thought to be essential gene candidates. Those single gene knockout mutants have the same genome background, which results in an advantage for distinct functional analysis of the targeted gene. The genome-wide relationship between the genome structure, i.e., genotype, and the phenomena, i.e., phenotype, which are analyzed by using the Keio collection has become available. In addition to the Keio collection, results of double gene knockouts by the literature [2] is also used.

After the above evaluation, some differences between prediction results and biological results of [6] and [2] will be found. Those differences are used to found lacked reactions, bypass pathways actually used in cells, and improvements for pathway model, which refine extended pathways.

Although Figure 1 shows specific databases for *E. coli*,

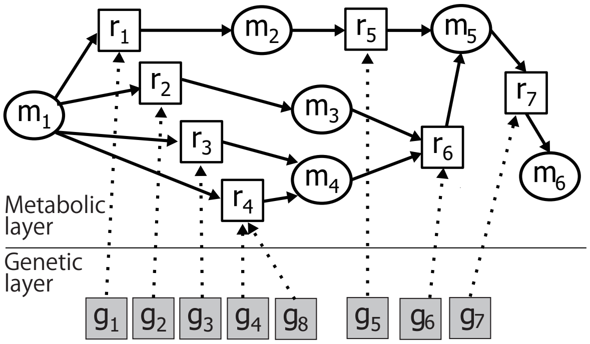


Figure 2. Example of an Extended Pathway

the research framework itself can be applied for other organisms whose pathway information is available. For instance, although there is a large difference between *E. coli* and mice, the central metabolism is similar and it could be a potential application.

III. Extended Pathways

In this section, we explain how to represent metabolic pathways and their relations to genes. We then define the extended pathway.

To represent metabolic pathways, we commonly use bipartite directed graph representation as follows. Let M be a set of metabolites and R be a set of reactions. For M and R, $M \cap R = \emptyset$ holds. Let $A_M \subseteq (R \times M) \cup (M \times R)$ be a set of arcs. A *metabolic pathway* is represented in a directed bipartite graph $\mathcal{G_M} = (M \cup R, A_M)$, where M and R are two sets of nodes, and A_M is a set of arcs. In addition to the metabolic pathway, we consider relations between enzymatic reactions and genes. Let G be a set of genes and A_G be a set of arcs such that $A_G \subseteq (G \times R)$. That is, A_G represents relations between enzymatic reactions and genes. Let N be a set of nodes such that $N = M \cup R \cup G$ and A be a set of arcs such that $A = A_M \cup A_G$. Then, the *extended pathway* is represented in a directed graph $\mathcal{G} = (N, A)$.

Figure 2 shows an example of the extended pathway. As the figure shows, it consists of two layers: the metabolic layer and the genetic layer. The genetic layer is the difference between the metabolic pathway and the extended pathway. In this example, the pathway consists of nodes of $M = \{m_1, m_2, \ldots, m_6\}$, $R = \{r_1, r_2, \ldots, r_7\}$, and $G = \{g_1, g_2, \ldots, g_8\}$. Each arc represents relations between elements. For instance, the activation of the reaction r_6 needs the production of metabolites m_3 and m_4 and the expression of g_6. We will explain the interpretation of the extended pathway in detail in the next section.

IV. Minimal Active Pathways with Gene Knockouts

In this section, we explain the notion of *producible*, *activatable* and *minimal active pathway* on the extended pathway, while the minimal active pathway is introduced only on the metabolic pathway in the literature [9].

We here define $M_S \subset M$ as a set of source metabolites and $M_T \subset M$ as a set of target metabolites such that $M_S \cap M_T = \emptyset$. An *extended pathway instance* is represented in a four tuple $\pi = (N, A, M_S, M_T)$, where $N = M \cup R \cup G$, $A = A_M \cup A_G$. Let K be a set of genes such that $K \subseteq G$. We use K as a set of knockout genes in a given pathway. A *knockout instance* is represented in a five tuple $\pi_K = (N, A, M_S, M_T, K)$. If $K = \emptyset$ then π_K corresponds to π.

Let m, r be a metabolite and a reaction such that $m \in M$ and $r \in R$, respectively. A metabolite $m \in M$ is called a *reactant* of a reaction $r \in R$ when there is an arc $(m, r) \in A$. On the other hand, a metabolite $m \in M$ is called a *product* of a reaction $r \in R$ when there is an arc $(r, m) \in A$. Furthermore, a gene $g \in G$ is called a *corresponding gene* of a reaction $r \in R$ when there is an arc $(g, r) \in A$.

A reaction is called a *reversible reaction* if it can occur in both directions between reactants and products. In this paper, we distinguish a reversible reaction as two reactions. Suppose that there is a reversible reaction r_1 that has m_1 and m_2 as reactants and m_3 and m_4 as products. In this case, we split the reaction r_1 into two reactions r_{1a} and r_{1b} such that one of them has m_1 and m_2 as products and m_3 and m_4 as reactants.

Let $s : R \to 2^M$ be a mapping from a set of reactions to a power set of metabolites such that $s(r) = \{m \in M \mid (m, r) \in A\}$ represents the set of metabolites that are needed to turn the reaction r to be active. Let $p : R \to 2^M$ be a mapping from a set of reactions to a power set of metabolites such that $p(r) = \{m \in M \mid (r, m) \in A\}$ represents the set of metabolites that are produced by the reaction r. Let $c : R \to 2^G$ be a mapping from a set of reactions to a power set of genes such that $c(r) = \{g \in G \mid (g, r) \in A\}$ represents the set of genes that are corresponding genes of the reaction r. Let $p' : M \to 2^R$ be a mapping from a set of metabolites to a power set of reactions such that $p'(m) = \{r \in R \mid (r, m) \in A\}$. Let $c' : G \to 2^R$ be a mapping from a set of genes to a power set of reactions such that $c'(g) = \{r \in R \mid (g, r) \in A\}$.

Let t be an integer variable representing time. In this paper, the time is used to represent order relation between reactions to produce target metabolites from source metabolites. In the following, we explain important notions related to production of metabolites, activation of reactions, and expression of genes. Since we focus on gene knockouts, we suppose that almost all genes exist in the cell of a given organism. We also suppose that if genes exist, then they are expressed and available to construct enzymes needed for enzymatic reactions. The reason for this condition is that we want to simulate how the lack of corresponding genes affects metabolic pathway rather than how the existence of genes affects other elements. Although our pathway modeling is simple, it allows us to analyze a whole cell scale pathway.

Let $\pi_K = (N, A, M_S, M_T, K)$ be a knockout instance, where $N = M \cup R \cup G$, $A = A_M \cup A_G$. Let $\mathcal{G} = (N, A)$ be an extended pathway. Let $M' \subset M$ be a subset of metabolites. A metabolite $m \in M$ is obviously *producible* at time $t = 0$ from M' on $\mathcal{G}$ if $m \in M'$ holds. A reaction $r \in R$ is *activatable* at time $t > 0$ from M' on $\mathcal{G}$ if the following two conditions are satisfied: (i) for every $m \in s(r)$, m is producible at time $t-1$ from M', (ii) at least one corresponding gene $g \in c(r)$ is not included in K. A metabolite $m \in M$ is *producible* at time $t > 0$ from M' on $\mathcal{G}$ if there is at least one activatable reaction r at time t such that $m \in p(r)$. If r is activatable at time t, then r is activatable at time $t+1$. If m is producible at time t, then m is producible at time $t+1$.

Let $\mathcal{G}' = (N', A')$ be a sub-graph of $\mathcal{G}$, where $N' = M' \cup R' \cup G'$ and $A' = A'_M \cup A'_G$. Then, an active pathway of $\pi_K = (N, A, M_S, M_T, K)$ is defined as follows.

Definition 1: Active Pathway of Knockout Instance
A bipartite directed graph $\mathcal{G}'$ is an *active pathway* of π_K if it satisfies the following conditions:

- $M_T \subset M'$
- $M' = M_S \cup \{m \in M \mid (m, r) \subseteq A, r \in R'\} \cup \{m \in M \mid (r, m) \subseteq A, r \in R'\}$
- $A' = \{(m, r) \in A \mid r \in R'\} \cup \{(r, m) \in A \mid r \in R'\} \cup \{(g, r) \in A \mid g \notin K, r \in R'\}$
- $G' = \{g \in G \mid (g, r) \in A', r \in R'\}$
- For every $m \in M'$, m is producible from M_S on $\mathcal{G}'$

From Definition 1, active pathways include a set of metabolites, reactions, and genes, which are producible and activatable from M_S on $\mathcal{G}'$ such that all target metabolites M_T become producible. The number of active pathways depends on the combination of M_S and M_T but an extended pathway generally has a large number of active pathways. We thus particularly focus on minimal ones rather than active pathways. We give the definition of minimal active pathways of π_K as follows. Let $\mathcal{G}$ and $\mathcal{G}'$ be extended pathways. We say that $\mathcal{G}$ is *smaller* than $\mathcal{G}'$ and represented in $\mathcal{G} \subset \mathcal{G}'$ if $R \subset R'$. An active pathway $\mathcal{G}$ is *minimal active pathway* of π_K *iff* there is no active pathway of π_K, which is smaller than $\mathcal{G}$. As this definition shows, we only need to see sets of reactions to compare two pathways. Thus, in the rest of this paper, we sometimes represent a minimal active pathway as a set of reactions.

Any reactions included in a minimal active pathway cannot be deleted to produce target metabolites. Intuitively, this indicates that each of the elements of a minimal active pathway is essential. In practice, minimal active pathways including a large number of reactions are considered to be biologically inefficient. We thus introduce a time limitation z and pathways that can make all target metabolites producible by $t = z$. In the following, we consider the problem of finding minimal active pathways with respect to π_K and z.

V. Knockout Effects

This section provides how to predict knockout effects. In the following, we give some definitions for the prediction. Let $\pi = (N, A, M_S, M_T)$ and $\pi_K = (N, A, M_S, M_T, K)$ be an extended pathway instance and a knockout instance, respectively. In addition, we denote the number of minimal active pathways of π as $|\pi|$ and the number of minimal active pathways of π_K as $|\pi_K|$. Obviously, $|\pi_K| \leq |\pi|$ holds. Then, the gene knockout effect, i.e., the prediction by the proposed method, is given by $E_K = |\pi| - |\pi_K|$. Let K_a and K_b be sets of knockout genes. If $E_{K_a} > E_{K_b}$ holds, then we say that the gene knockout effect of K_a is stronger than that of K_b. If $|\pi_K| = 0$, i.e., $E_K = |\pi|$, then we say that the knockout effect of K is *critical* to produce target metabolites. Various metabolites are known as vital metabolites, which means organisms cannot survive without them. That is, if some gene knockouts are critical to produce such metabolites, then a given organism cannot grow any more or dies. If $|K| = 1$ and its effect is critical to produce vital metabolites, then we say that the gene $g \in K$ is *essential*.

In the following, we explain the above definition with a specific example. Suppose that we are given a pathway instance $\pi = (N, A, M_S, M_T)$, where N and A are from the extended pathway shown in Figure 2, and the source metabolite is $M_S = \{m_1\}$ and the target metabolite is $M_T = \{m_6\}$. Obviously, $|\pi| = 3$ and the minimal active pathways of π are specifically as follows: $\{r_1, r_5, r_7\}, \{r_2, r_3, r_6, r_7\}, \{r_2, r_4, r_6, r_7\}$. Then, we consider the following knockout instances π_{K_1} and π_{K_2}, where $K_1 = \{g_1\}$ and $K_2 = \{g_6\}$. For π_{K_1}, minimal active pathways including r_1 can no longer be solutions, i.e., $|\pi_{K_1}| = 2$. For π_{K_2}, minimal active pathways including r_6 can no longer be solutions either. Thus, $\{r_2, r_3, r_6, r_7\}$ and $\{r_2, r_4, r_6, r_7\}$ are deleted from the solutions of π, i.e., $|\pi_{K_2}| = 1$. Consequently, we can say that the knockout effect of K_2 is stronger than that of K_1. Moreover, suppose that $K_3 = \{g_7\}$. Then, there is no minimal active pathway of π_{K_3} and we say that the knockout effect of K_3 is critical to produce m_6. If m_6 is a vital metabolite, we can simultaneously say that g_7 is an essential gene.

In addition to the number of remaining minimal active pathways after knockouts, an important factor in the prediction is the gain of ATPs. This is because pathways that are inefficient with respect to energy consumption will not be used in organisms. Let $|\pi^{a+}|$, $|\pi_K^{a+}|$ be the number of minimal active pathways of π and π_K, which gain ATPs, respectively. Then, the gene knockout effect with respect to ATP production is given by $E_K^{a+} = |\pi^{a+}| - |\pi_K^{a+}|$. In particular, it is important when we consider the glycolysis pathway since one of its main functions is to gain ATPs. However, we cannot find any pathways producing ATPs on some other pathways, i.e., minimal active pathways on them must consume ATPs. In this case, the number of minimal active pathways, which consume fewer ATPs, should be considered instead of $|\pi^{a+}|$ and $|\pi_K^{a+}|$.

VI. Computational Method

This section explains how to compute $|\pi_K|$. In this paper, we use the method of computing all minimal active pathways of π by Soh and Inoue [9]. This method computes pathways through propositional encoding and minimal model generation. An advantage is that this method is flexible for adding biological constraints, which is explored in [9]. Moreover, we can utilize SAT technologies, which have been developed actively in recent years.

In the following, we briefly explain the propositional encoding to compute minimal active pathways of π. Let i, j be integers denoting indices for metabolites and reactions. Let t be an integer variable representing time. Let $\pi = (N, A, M_S, M_T)$ be an extended pathway instance, where $N = M \cup R \cup G$, $A = A_M \cup A_G$. We introduce two kinds of propositional variables. Let $m^*_{i,t}$ be a propositional variable, which is $true$ if a metabolite $m_i \in M$ is producible at time t. Let $r^*_{j,t}$ be a propositional variable, which is $true$ if a reaction $r_j \in R$ is activatable at time t.

The encoding of the problem of finding minimal active pathways with respect to π_K and z is as follows.

$$\psi_1 = \bigwedge_{0\le t<z} \bigwedge_{m_i\in M} (m^*_{i,t} \rightarrow m^*_{i,t+1})$$

$$\psi_2 = \bigwedge_{0\le t<z} \bigwedge_{r_j\in R} (r^*_{j,t} \rightarrow r^*_{j,t+1})$$

$$\psi_3 = \bigwedge_{1\le t\le z} \bigwedge_{r_j\in R} \left(r^*_{j,t} \rightarrow \bigwedge_{m_i\in s(r_j)} m^*_{i,t-1} \right)$$

$$\psi_4 = \bigwedge_{1\le t\le z} \bigwedge_{r_j\in R} \left(r^*_{j,t} \rightarrow \bigwedge_{m_i\in p(r_j)} m^*_{i,t} \right)$$

$$\psi_5 = \bigwedge_{m_i\in(M\setminus M_S)} \bigwedge_{1\le t\le z} \left(m^*_{i,t} \rightarrow m^*_{i,t-1} \vee \bigvee_{r_j\in p'(m_i)} r^*_{j,t} \right)$$

$$\psi_6 = \bigwedge_{m_i\in M_S} m^*_{i,0} \wedge \bigwedge_{m_{i'}\in M\setminus M_S} \neg m^*_{i',0}$$

$$\psi_7 = \bigwedge_{m_i\in M_T} m^*_{i,z}$$

The formulas ψ_1 and ψ_2 represent that once a metabolite (or a reaction) is made to producible (or activatable), then it remains in the producible (or activatable) state. The formula ψ_3 represents that if a reaction r_j is activatable at time t then its reactants must be producible at time $t-1$. The formula ψ_4 represents that if a reaction r_j is activatable at time t then its products must be producible at time t. The formula ψ_5 represents that if a reaction m_i is producible then either two states hold: the metabolite m_i is producible at $t-1$ or at least one reaction r_j is activatable. The formulas ψ_6 and ψ_7 represent source metabolites and target metabolites. We denote the conjunction of $\psi_1, \ldots, \psi_7$ as Ψ_z. Then, we can enumerate minimal active pathways with respect to π_K and z by computing minimal models of Ψ_z with respect to $V^z = \{r^*_{i,z} | r_i \in R\}$.

The computation for π is always needed to compare a wild cell and its mutant. We thus explain a method to compute all minimal active pathways of π_K for a set of knockout genes K. Actually, when the minimal active pathways of π are obtained, we do not need much additional computation. All minimal active pathways of π_K are obtained by selecting pathways that do not contain some $r \in R_K$, where $R_K = \{r \in c'(g) \mid g \in K\}$. The procedure is given as follows: (i) enumerate all minimal active pathways with respect to π and z, (ii) delete minimal active pathways including some $r \in R_K$, where $R_K = \{r \in c'(g) \mid g \in K\}$. As well as the above procedure, there is another way to compute all minimal active pathways with respect to π_K and z. The same is achieved by adding constraints, which inhibit the activation of each reaction in R_K, to the formula Ψ_z.

VII. Experimental Results

This section provides experimental results and discussions. At first, we describe experimental conditions. Then, we show the results of our prediction of knockout effects for glycolysis and amino acids biosynthesis.

A. Experimental conditions

We constructed extended pathways from EcoCyc [11] and KEGG [10]. Specifically, we use EcoCyc to construct metabolic pathways, which consists of 1222 metabolites and 1920 reactions. Moreover, we use KEGG to construct relations between enzymatic reactions and genes. The entire extended pathway we used is constructed from these two databases. In the following experiments, we denote a reversible reaction in EcoCyc as two differently directed two reactions by adding suffixes *_a* and *_b*, respectively. In addition, some reactions such as *6PGLUCONDEHYDROG-RXN* can accept different metabolites as its input, e.g., *6PGLUCONDEHYDROG-RXN* is considered to be able to use NAD$^+$ and NADP$^+$. In this case, we distinguish it as two different reactions by adding suffixes _1 and _2, respectively.

Each experiment has been done using a PC (3.2GHz CPU) running on OS X 10.6. For computation, we use a SAT solver Minisat2 [12]. Koshimura *et al.* proposed a procedure computing minimal models with SAT solvers [13]. We follow their procedure to generate minimal models by using a SAT solver.

To evaluate our method, we use the Keio collection as is described in Section II. In particular, we use their results on the MOPS medium whose main nutrient is glucose.

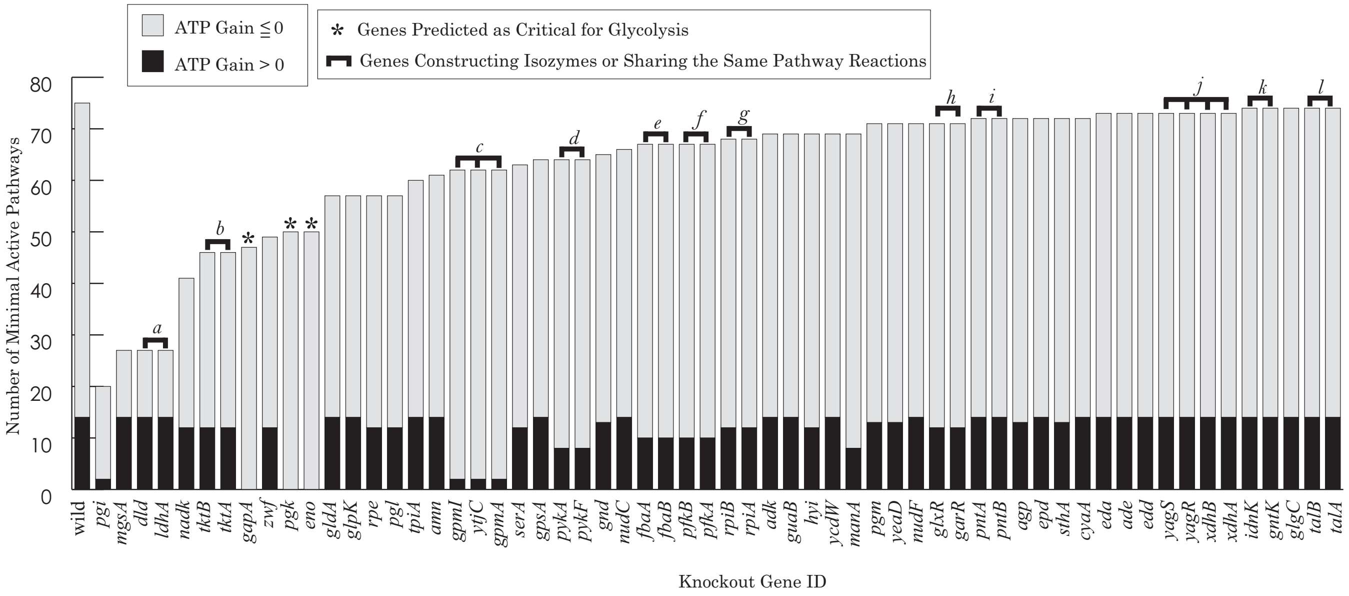
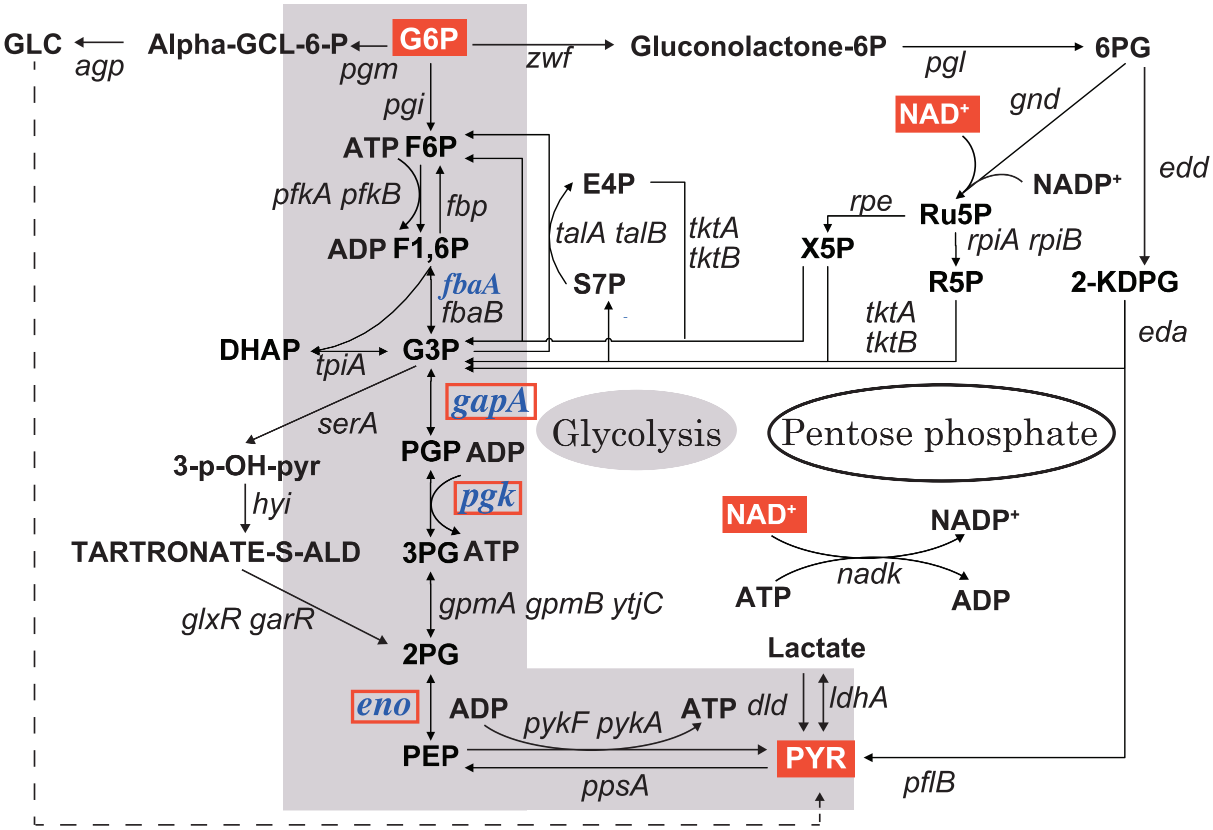


Figure 3. The Number of Minimal Active Pathways for each Gene Knockout on Glycolysis

Table I
SINGLE GENE KNOCKOUTS FOR GLYCOLYSIS
Symbols "*" denote genes predicted as critical and "*a*" to "*l*" denote genes constructing isozymes or sharing the same pathway reactions.

Gene	#Minimal Active Pathways			Keio Collection [6]	
	Total	ATP Gain		MOPS24hr	MOPS48hr
		plus	others		
wild	75	14	61	0.219-0.392	0.216-0.480
pgi	20	2	18	0.137	0.542
mgsA	27	14	13	0.293	0.371
dld^{a}	27	14	13	0.303	0.366
ldhA^{a}	27	14	13	0.357	0.393
nadk (yfjB)	41	12	29	N.A.	N.A.
tktB^{b}	46	12	34	0.311	0.315
tktA^{b}	46	12	34	0.317	0.327
*gapA**	47	0	47	N.A.	N.A.
zwf	49	12	37	0.231	0.223
*pgk**	50	0	50	N.A.	N.A.
*eno**	50	0	50	N.A.	N.A.
gldA	57	14	43	0.255	0.351
glpK	57	14	43	0.283	0.409
rpe	57	14	43	0.347	0.335
pgl(ybhE)	57	12	45	0.551	0.686
tpiA	60	14	46	0.345	0.321
amn	61	14	47	0.330	0.342
gpmI^{c}	62	2	60	0.303	0.292
ytjC^{c}	62	2	60	0.339	0.378
gpmA^{c}	62	2	60	0.383	0.240
serA	63	12	51	0.007	0.021
gpsA	64	14	50	N.A.	N.A.
pykA^{d}	64	8	56	0.266	0.299
pykF^{d}	64	8	56	0.310	0.320
gnd	64	12	52	0.251	0.282
nudC	66	14	52	0.445	0.518
fbaA^{e}	67	10	57	N.A.	N.A.
fbaB^{e}	67	10	57	0.371	0.447
pfkB^{f}	67	10	57	0.270	0.258
pfkA^{f}	67	10	57	0.087	0.554

Gene	#Minimal Active Pathways			Keio Collection [6]	
	Total	ATP Gain		MOPS24hr	MOPS48hr
		plus	minus		
rpiB^{g}	68	12	56	0.326	0.394
rpiA^{g}	68	12	56	0.340	0.372
adk	69	14	55	N.A.	N.A.
guaB	69	14	55	0.005	0.020
hyi	69	12	57	0.191	0.197
ycdW	69	14	55	0.255	0.301
manA	69	8	61	0.334	0.355
pgm	71	13	58	0.169	0.158
yeaD	71	13	58	0.233	0.289
nudF	71	14	57	0.376	0.373
glxR^{h}	71	12	59	0.226	0.231
garR^{h}	71	12	59	0.400	0.368
pntA^{i}	72	14	58	0.220	0.288
pntB^{i}	72	14	58	0.317	0.513
agp	72	13	59	0.319	0.528
epd	72	14	58	0.321	0.344
sthA	72	14	58	0.271	0.515
cyaA	72	14	58	0.455	0.295
eda	73	14	59	0.211	0.205
ade	73	14	59	0.246	0.304
edd	73	14	59	0.269	0.282
yagS^{j}	73	14	59	0.273	0.125
yagR^{j}	73	14	59	0.295	0.306
xdhB^{j}	73	14	59	0.324	0.447
xdhA^{j}	73	14	59	0.433	0.488
idnK^{k}	74	14	60	0.230	0.140
gntK^{k}	74	14	60	0.303	0.303
glgC	74	14	60	0.304	0.327
talB^{l}	74	14	60	0.254	0.169
talA^{l}	74	14	60	0.316	0.345



Source metabolites (M_S): β-D-glucose-6-phosphate, NAD+	G6P NAD+
Target metabolite (M_T): pyruvate	PYR
Four essential genes confirmed by KEIO collection	*fbaA* *gapA* *pgk* *eno*
Genes predicted by our method	□

Figure 4. Glycolysis and Pentose Phosphate Pathways of *E. coli* from the literature by Ishii *et al.* [8]

In addition, we use the results of the literature [2] for a comparison for double gene knockouts. For the former results, i.e., Keio collection, we consider that the set of knockout genes K consists of one gene. The cell growth of the wild cell is ranged from 0.216 to 0.480 in Keio collection and 0.378 in the literature [2]. For both data, if cell growth is less than 0.1, which is less than half of them, we then say that the cell is strongly affected by a gene knockout.

B. Results for Glycolysis Analysis

First, we analyze the glycolysis pathway of *E. coli*. In accordance with the MOPS medium of the Keio collection [6], a set of source metabolites M_S is chosen as follows: $\{\beta$-D-glucose-6-phosphate, H^+, H_2O, ATP, ADP, phosphate, and $NAD^+\}$. In addition, pyruvate is given as the target metabolite to analyze glycolysis, i.e., $M_T = \{\text{pyruvate}\}$. We then compute all minimal active pathways from the entire metabolic pathway of *E. coli*. As we can see in biological literature such as the work of Ferguson *et al.* [14], glycolysis is known to a pathway constructed by eight steps. However, if some reactions are disabled, then *E. coli* is expected to use other bypass pathways by using additional reactions. In this experiment, we consider four additional reactions, i.e., the number of reactions included in each pathway is limited to less than or equal to 12 as well as $z = 12$.

At first, we computed all minimal active pathways with the above conditions and obtained 75 minimal active pathways from the entire reactions database, which consists of 1920 reactions. We then connect 61 genes to reactions by API on KEGG. Since there is no data, some reactions are remaining unconnected. Next, we computed minimal active pathways with each gene knockout. This experiment was done within four seconds. Figure 3 shows the results of 61 gene knockouts. The x-axis denotes each gene knockout and the y-axis denotes the number of minimal active pathways. As is shown in the figure, we compute minimal active pathways of $\pi_{K_1}, \ldots, \pi_{K_{61}}$ such that $K_1 = \{pgi\}, K_2 = \{mgsA\}, \ldots, K_{61} = \{talA\}$. However, since some of the 61 genes construct isozymes, such single gene knockout K_i does not affect the number of minimal active pathways $|\pi_{K_i}|$. However, for reference, we compute the effect of the gene knockouts that disables all of isozymes. For instance, *tktA* and *tktB* construct isozymes. In this case, the number of minimal active pathways in the figure shows the case of the gene knockout of both *tktA* and *tktB*. For each gene knockout, we computed the gain of ATP in each minimal active pathway, which is calculated by counting the number of both reactions with the coefficient of ATP: ones consuming ATP and the other ones producing ATP. Minimal active pathways that produce the positive number of ATPs are more important than the others because producing ATP is a main function of glycolysis.

From the figure, we can see that *E. coli* keeps almost all minimal active pathways even by more than half of single gene knockouts. This is considered to indicate the robustness of *E. coli*. However, some gene knockouts dramatically reduce the number of minimal active pathways. In particular, the single gene knockouts of *gapA*, *pgk*, and *eno* destroy all minimal active pathways producing ATPs. Thus, they are predicted to strongly affect the glycolysis of *E. coli*.

To evaluate the above predictions, we compare them with the Keio collection. Table I compares all gene knockouts shown in Figure 3 regarding the number of lost minimal active pathways. Column 1, Gene, shows gene names except *wild*, which denotes an empty set of knockout genes, i.e., $K = \emptyset$. Other rows denote the result of single gene knockout. Column 2, Total, shows the total number of minimal active pathways, i.e., $|\pi_{K_i}|$. Columns 3 and 4 show the number of minimal active pathways, which gain ATPs, i.e, $|\pi_{K_i}| > 0$ and consume ATPs, i.e., $|\pi_{K_i}| \leq 0$. Column 7, MOPS24hr, and Column 8, MOPS48hr, show the cell growth of *E. coli* after 24 hours and 48 hours, respectively. Note that N.A. (not applicable) refers to essential genes [6]. As the first row of Table I shows, we found 14 minimal active pathways that produce the positive number of ATPs on the wild cell of *E. coli* while there are 75 in total[1].

Distinguished single gene knockouts are $K_8 = gapA$, $K_{10} = pgk$, and $K_{11} = eno$. Each gene knockout effect with respect to ATP production is $E^{a+}_{K_8} = E^{a+}_{K_{10}} = E^{a+}_{K_{11}} = 14$ and it is the strongest gene knockout effect with respect to ATP production, which is the important function of glycolysis. For this prediction, the Keio collection shows “N.A.” for each gene knockout. Thus, in glycolysis, our predictions successfully agree with the results of the Keio collection. However, there are other gene knockouts showing “N.A.” in the results of Keio collection, that is, *gpsA*, *fbaA*, *nadk (yfjB)* and *adk*. For those genes, the number of minimal active pathways is not so reduced. Then, in those gene knockouts, it can be considered that *E. coli* is damaged in other pathways rather than the glycolysis pathway, discussing at the following sections.

C. Discussion for Glycolysis Analysis

In this section, we first discuss about the difference of our prediction and the cell growth of the Keio collection. Figure 4 shows the glycolysis pathway modified from the one in the literature [8]. We pick up the figure of glycolysis and pentose phosphate pathways. Abbreviation is same as the literature [8]. Each node, e.g., G6P, denotes a metabolite and edge denote chemical reactions. Labels of edges denote corresponding genes to reactions. A dotted line denotes the abstraction of some reactions whose genes are not registered in the databases. The figure also shows generally known four essential genes in terms of the glycolysis pathway, which are also confirmed by the Keio collection. One of them, *fbaA* is not predicted to be critical for the cell growth since the gene knockout cell keeps almost all minimal active pathways producing ATPs even if we delete both *fbaB* and *fbaA*. Specifically, the knockout lost only four minimal active pathways producing ATPs (see Table I). Thus, two hypotheses come up. One is that the four lost minimal active pathways are the most important pathways in glycolysis. The other is that the essentiality is caused by the breakdown of other cell functions. However, the first hypothesis is considered not to be true by the following discussion.

As the results in the previous section show, our method predicted three out of four essential genes. Focusing on minimal active pathways lost by gene knockouts allows us to find an important part of glycolysis. Table II shows the minimal active pathways lost by each gene knockouts of *gapA*, *pgk* and *eno*. Column 1, shows reaction names from the database of Ecocyc [11]. Reactions in the glycolysis pathway are collected in the upper part of the table. Column 2, shows corresponding genes to the reaction in Column 1 that are shown in Figure 4. Columns 3 to 16 show the 14 minimal active pathways such as $p_1, \ldots, p_{14}$ disabled by the gene knockouts of *gapA*, *pgk* and *eno*. For each column, “x” denotes a reaction contained in each minimal active pathway. In the case that two reactions have the same corresponding gene, e.g., *gnd*, we show the lacking effect of the two reactions respectively for reference. All 14 pathways are producing ATPs. For instance, a minimal active pathway p_8 consists of 8 reactions corresponding to the following genes: $\{$*fbaB, fbaA*$\}$, *pgk*, *eno*, *pgi*, $\{$*pfkB, pfkA*$\}$, $\{$*pykF, pykA*$\}$, *gapA*, $\{$*gpmA, gpml, ytjC*$\}$. This pathway p_8 is known as a typical glycolysis pathway and p_5 is known as a bypass pathway using pentose phosphate pathway that is used when p_8 is not available [8]. The glycolysis pathway p_8

[1] Those 75 minimal active pathways are shown in a supporting online material in http://kix.istc.kobe-u.ac.jp/~soh/supplement/prediction.html.

Table II
14 PATHWAYS, WHICH PRODUCE ATP, DISABLED BY *gapA*, *pgk* AND *eno*
Symbols "*" denote genes predicted as critical for glycolysis.

Reaction Name	Gene	14 out of 75 pathways disabled by the single gene knockout of *gapA*, *pgk* and *eno*													
		p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8	p_9	p_{10}	p_{11}	p_{12}	p_{13}	p_{14}
PGLUCISOM-RXN_a	*pgi*	x	x	x	x			x	x	x	x	x	x	x	x
6PFRUCTPHOS-RXN	*pfkB,pfkA*								x	x				x	x
F16ALDOLASE-RXN_a	*fbaB,fbaA*								x	x				x	x
GAPOXNPHOSPHN-RXN_a	*gapA**	x	x	x	x	x	x	x	x	x	x	x	x	x	x
PHOSGLYPHOS-RXN_b	*pgk**	x	x	x	x	x	x	x	x	x	x	x	x	x	x
3PGAREARR-RXN_a	*gpmA,gpmI,ytjC*	x	x	x	x	x	x	x	x	x	x			x	x
2PGADEHYDRAT-RXN_a	*eno**	x	x	x	x	x	x	x	x	x	x	x	x	x	x
PEPDEPHOS-RXN_b	*pykF,pykA*	x				x	x		x			x	x		
PHOSPHOGLUCMUT-RXN_b	*pgm*										x				
GLU6PDEHYDROG-RXN_a	*zwf*					x	x								
6PGLUCONOLACT-RXN	*pgl(ybhE)*					x	x								
6PGLUCONDEHYDROG-RXN_1	*gnd*						x								
6PGLUCONDEHYDROG-RXN_2	*gnd*					x									
RIBULP3EPIM-RXN_a	*rpe*					x	x								
RIB5PISOM-RXN_b	*rpiA,rpiB*					x	x								
1TRANSKETO-RXN_b	*tktB,tktA*					x	x								
NAD-KIN-RXN	*nadk (yfjB)*					x	x								
MANNPISOM-RXN_b	*manA*		x	x	x					x				x	x
PGLYCDEHYDROG-RXN_a	*serA*											x	x		
RXN0-305_a	*hyi*											x	x		
RXN0-5289_b	*glxR,garR*												x		
TSA-REDUCT-RXN_1	*glxR,garR*											x			
GLUCOSE-6-PHOSPHATE-1-EPIMERASE-RXN_b	*yeaD*										x				
MANNKIN-RXN_b			x	x	x					x				x	x
GKI-RXN												x	x		
RXN0-6562												x	x		
RXN0-6418_b											x				
TRANS-RXN-158					x									x	
GLUCOSE-1-PHOSPHAT-RXN	*agp*										x				
RXN0-313_a		x	x	x	x			x			x	x	x		
TRANS-RXN-157											x				
TRANS-RXN-158A				x											x
2.7.1.121-RXN								x							
TRANS-RXN-165			x							x					
MANNOSE-ISOMERASE-RXN_a				x	x									x	x
Total Number of Reactions		7	9	10	10	12	12	7	8	10	11	11	11	11	11
Total Number of Corresponding Genes		6	6	6	6	12	12	5	8	8	8	8	8	8	8

is activated mainly and the activity of the pentose phosphate pathway p_5 is minimized at normal condition, however, p_5 is maximized when p_8 is inactivated [8]. It will be difficult to detect of all gene-expressions or enzymatic activities in the cell on every conditions, instead of this issue, our minimal pathway analysis will be a new approach for understanding of biological robustness and systems.

From Table II, we also can easily read which gene knockout disable which minimal active pathways. Note that minimal active pathways are disabled even if only one of its components is lacked. For instance, *gapA*, *pgk* and *eno* are contained in all 14 minimal active pathways and it means that the knockout of each of them immediately disables all 14 minimal active pathways. Then, the following is observed: considering all minimal active pathways of glycolysis from a whole reaction database of *E. coli*, those pathways producing ATPs always include known essential genes. In other words, the importance of 14 minimal active pathways are confirmed. In addition, it is also confirmed that even if there are computationally feasible pathways, they cannot be bypass pathways in *E. coli*. We thus can expect that *E. coli* cannot survive without those 14 pathways. This assumption allows us to predict the effect of *multiple gene knockouts*. The knockout of *pgi* disables 12 pathways but the Keio collection shows that *E. coli* is still alive with this gene knockout. In this case, it is supposed that *E. coli* manages to survive with only 2 out of 14 pathways, i.e., p_5 and p_6. In other words, those remaining pathways are supposed to be used as bypass pathways. For instance, *pgi* encoding glucosephosphate isomerase gene of glycolysis pathway that transfer D-glucose 6-phosphate to D-fluctose 6-phosphate. However, pentose phosphate pathway is available as a bypass pathway from D-glucose 6-phosphate, resulting in the gene knockout slow-growth at starting MOPS24hr and same level of the final growth of the wild cell at MOPS48hr. Then, what will happen if we additionally knockout *rpe*, *zwf*, *gnd* or *pgl(ybhE)*? Each of those gene knockouts disable both p_5 and p_6. Thus, the following double knockouts disable all 14 pathways: (a) *zwf* and *pgi*, (b) *rpe* and *pgi*, (c) *gnd* and *pgi*, and (d) *pgl(ybhE)* and *pgi*. Then, we can predict that they are critical for *E. coli*. As well as the Keio collection, Nakahigashi *et al.* measured the growth rate of *E. coli* with some combinations of double gene knockouts including the above (a), (b) and (c) [2]. Table III shows their results and both double gene knockouts of (a) and (b) affect so strong that the growth rate of *E. coli* becomes less than 0.1. Thus, as well as the prediction of single gene knockouts, our prediction agrees with the biological results for (a) and (b). However, for the knockouts of (c), our prediction disagrees with it. For this issue, comparing more combinations of genes to be critical is needed and biological evaluations for them are necessary future work.

Table III
GROWTH RATE OF *E. coli* WITH DOUBLE GENE KNOCKOUTS [2]

Knockout genes	Growth Rate [2]	
	24 hours	48 hours
(a) *zwf*, *pgi*	0.033	0.070
(b) *rpe*, *pgi*	0.001	0.140
(c) *gnd*, *pgi*	0.199	0.322

As Table II shows, four minimal active pathways disabled by the knockout of *fbaA* are also disabled by the knockout of *pgi* that is not an essential gene. Thus, the first hypothesis discussed in the former part of this section is not true. A gene *nadk (yfjB)* is similar to *fbaA*. Our method predicts that this knockout does not affect the cell growth in terms of glycolysis. However, the Keio collection shows that this is an essential gene for *E. coli*. In the case of *nadk (yfjB)*, we found that this gene knockout affects other function of *E. coli*. In relation to this, we have additional experiments for amino acid generation in the following section.

D. Results for Amino Acids Generation

We also applied our prediction method to predict gene knockout effects of the cell growth in terms of amino acid biosynthesis. Since we want to involve more genes for our prediction, we particularly focus on essential amino acids for humans, whose synthesis needs more reactions than others. In the experiments, we separately constructed pathway instances, each of which consists of the following eight amino acids as a target metabolites: L-valine (VAL), L-leucine (LEU), L-phenylalanine (PHE), L-isoleucine (ILE), L-threonine (THR), L-lysine (LYS), L-tryptophan (TRP) and L-methionine (MET). In addition, to produce the above amino acids, we added the following metabolites to the source metabolites used in the glycolysis analysis: coenzyme-A and sulfite. For each of the eight amino acids, the computation time is on average 255 seconds and the longest computation time is 877 seconds.

In contrast to the result of glycolysis, we found there are 11 single gene knockouts that destroy all minimal active pathways without the limitation of z. That is, no pathway can synthesize each target on the entire metabolic pathway of *E. coli* with those single gene knockouts. Obviously, they are predicted to be critical to produce each amino acid. Table IV shows the cell growth of Keio collection. Column 1, gene, shows knockout genes predicted as critical by our prediction. Column 2, unsynthesized target, shows target amino acids, which cannot be synthesized with the knockout of the gene in Column 1. Columns 3 and 4 show the cell growth of *E. coli* after 24 hours and 48 hours, respectively. At first, the gene knockout of *nadk (yfjB)* is predicted as critical for the cell growth in terms of six amino acids biosynthesis. This result is also supported by the Keio collection. We thus consider the essentiality of *nadk (yfjB)* to be caused by its knockout effect in amino acids biosynthesis rather than glycolysis. Table IV also shows that our method predicts that no way to produce target metabolites with each single gene knockout: *folE*, *thrC*, *thrB*, *glnA*, *aroC*, *lysA*, *aroB*, and *leuA*. However, except *folE* and *thrC*, the Keio collection shows that *E. coli* survives with very low cell growth. One explanation for the results is that they are suspected to keep living by consuming unsynthesized amino acids from other individual cells. In this case, since the amino acids cannot be sustainably produced, those genes are recognized to be almost essential for *E. coli*.

Table IV
CRITICAL GENE KNOCKOUTS FOR AMINO ACIDS BIOSYNTHESIS

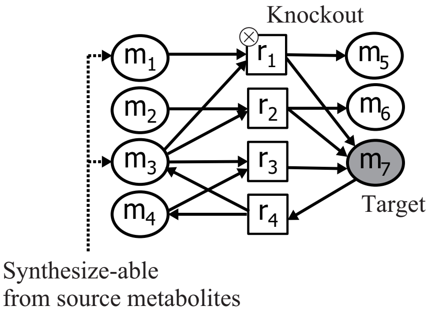
Gene	Unsynthesized Target	Keio Collection [6]	
		MOPS24hr	MOPS48hr
wild	-	0.219-0.392	0.216-0.480
folE	MET	N.A.	N.A.
nadk (yfjB)	VAL, LEU, THR, ILE, LYS, MET	N.A.	N.A.
thrC	THR	0.000	0.000
thrB	THR	0.004	0.010
glnA	TRP, MET	0.005	0.015
aroC	TRP, PHE, TRP	0.009	0.020
lysA	LYS	0.012	0.021
aroB	PHE, TRP, MET	0.010	0.032
leuA	LEU	0.026	0.034
folP	MET	0.283	0.293
metH	MET	0.357	0.509

Furthermore, the result of the Keio collection shows that the knockouts of *folP* and *metH* are not critical, although our method predicts them to be critical. We have detailed discussions on those gene knockouts in the following section.

E. Discussion for Amino Acids Generation

The difference between *folP* and *metH* in terms of amino acid biosynthesis also introduces interesting issues. At first, we consider *metH*, which constructs an enzymatic reaction methionine synthase. Its conversion is as follows: 5-methyltetrahydrofolate + L-homocysteine = tetrahydrofolate + L-methionine.

In both KEGG and EcoCyc databases, two alternative reactions exist to the above reaction. Figure 5 shows standard reaction and those two alternatives. An alternative reaction r_3, homocysteine S-methyltransferase, uses S-methyl-L-methionine (m_4) instead of 5-methyltetrahydrofolate (m_1). On the other hand, another alternative reaction r_2, 5-methyltetrahydropteroyltriglutamate–homocysteine methyltransferase, uses 5-methyltetrahydropteroyltri-L-glutamate (m_2). However, both metabolites cannot be synthesized from the source metabolites. Specifically, S-methyl-L-methionine (m_4) can be synthesized only from methionine (m_7), which is the target amino acid, and there is no reaction in the metabolic pathway of EcoCyc that can synthesize 5-methyltetrahydropteroyltri-L-glutamate (m_2), meaning that reactions are lacking in the database. The gene *folP* is on folate biosynthesis and there is no alternative in the databases. For the results of above genes, two hypotheses



r_1: 5-methyltetrahydrofolate--homocysteine methyltransferase (metH)
r_2: 5-methyltetrahydropteroyltriglutamate--homocysteine methyltransferase(metE)
r_3: homocysteine S-methyltransferase (mmuM)
r_4: homocysteine S-methyltransferase (mmuM, reverse)
m_1: 5-methyltetrahydrofolate
m_2: 5-methyltetrahydropteroyltri-L-glutamate
m_3: L-homocysteine
m_4: S-methyl-L-methionine
m_5: tetrahydrofolate
m_6: tetrahydropteroyltri-L-glutamate
m_7: L-methionine

Figure 5. Bypass Reactions for Synthesizing Methionine

are as follows: there are unknown complementary genes, or there are unknown bypasses. For the above issues, we need further analyses or researches on biological and computational level that would be creating new systems biology.

VIII. Related Work

There are several researches on metabolic pathway analyses. Schuster *et al.* proposed a method based on elementary mode analysis [15]. They focused on metabolic flux distributions corresponding to sets of reactions in metabolic pathways. A different point from our method is that their approach needs to define source metabolites strictly with a fixed amount that must be consumed in flux. In contrast, our method treats them as candidates that will be utilized; thus, we can flexibly give source metabolites. Handorf *et al.* proposed the *inverse scope problem* [16]. This is the problem of finding necessary source metabolites from target metabolites. The two differences between their problem and our proposal are as follows. One is that they only computed the cardinality minimal solution. Unlike their approach, we can generate subset minimal solution by minimal model generation. Another one is that each of their solutions includes all reactions, which are activatable from source metabolites needed to generate target metabolites. For instance, if there are two ways to produce a metabolite from source metabolites then both are mixed in one solution, that is, we cannot distinguish between them. On the other hand, our method can distinguish between the two ways, and we think that it is important to identify functionally minimal active pathways. Schaub and Thiele applied answer set programming (ASP) to solve the inverse scope problem [17], while we use propositional encoding and minimal model generation to compute minimal active pathways.

There are other researches using ASP [18], [19]. The literature [19] consider the most likely states of a reaction network with respect to given constraints and signaling pathways are analyzed with ASP. In [18], Ray *et al.* report a method using ASP to compute the steady states of a given pathway and complete lacking reactions. Unlike their approach, we use minimal model generation to compute essential reactions to produce target metabolites.

Küffer *et al.* report an approach using a Petri net [20]. Although their approach considers producibility and activatability of metabolites and reactions, they do not consider subset minimality of solutions.

IX. Conclusion and Future Work

In this paper, we propose a method to predict gene knockout effects by enumerating minimal active pathways. We formalize the extended pathway and show the definition of minimal active pathways on it. In addition, we present a computation method for the prediction. An advantage of our method is that it allows us to trace the reason for the prediction results, e.g., we can suggest the reason for the essentiality of three genes in the glycolysis pathway. This is an important feature that other methods do not have.

In the experiments, we applied our method to extended pathways of *E. coli* and made comparisons using the Keio collection. For the prediction of the knockout of 61 genes in the glycolysis pathway, our method predicted three essential genes, which correspond to the results of the Keio collection. Moreover, we analyze lethal 14 minimal active pathways and predict lethal pairs of gene knockouts, which also agree with the result of the literature [2]. In addition to the experiments in glycolysis, we found three essential genes and six almost essential genes in amino acids biosynthesis. We also discuss the reason for the difference between our prediction and results of the Keio collection with regard to the knockout of *metH* and *folP*. Although we treat relations between genes and enzymatic reactions that have one-to-one relations, we intend to extend them to relations that are more complex such as multiple relations and consider interactions among genes. Following that, we plan to apply our method to other organisms such as mice. In addition to *E. coli*, mice are well known model organisms for human study, and information available on them has been accumulated in the last decade. In particular, chromosome substitution strains are used to reveal the function of genes [21]. In addition to gene knockouts, we could adapt our method to such strains. Although there is a large difference between *E. coli* and mice, the basic metabolism is same. This fact tells us that our method can also be a potential prediction method for mice.

Acknowledgments

This research is supported in part by the 2008-2011 JSPS Grant-in-Aid for Scientific Research (A) No. 20240016, the research project of Systems Biology of Genetic Function, and the Transdisciplinary Seeds Research 2011 in the Transdisciplinary Research Integration Center. The authors would like to thank Kenji Nakahigashi and anonymous reviewers for their helpful comments.

References

[1] T. Soh, K. Inoue, T. Baba, T. Takada, and T. Shiroishi, "Predicting gene knockout effects by minimal pathway enumeration," in *The 4th International Conference on Bioinformatics, Biocomputational Systems and Biotechnologies (BIOTECHNO 2012)*, 2012, pp. 11–19.

[2] K. Nakahigashi, Y. Toya, N. Ishii, T. Soga, M. Hasegawa, H. Watanabe, Y. Takai, M. Honma, H. Mori, and M. Tomita, "Systematic phenome analysis of *Escherichia coli* multiple-knockout mutants reveals hidden reactions in central carbon metabolism," *Molecular Systems Biology*, vol. 5, no. 306, pp. 1–14, 2009.

[3] H. D. Jong, "Modeling and simulation of genetic regulatory systems: A literature review," *Journal of Computational Biology*, vol. 9, pp. 67–103, 2002.

[4] M. Terzer, N. D. Maynard, M. W. Covert, and J. Stelling, "Genome-scale metabolit networks," *Systems Biology and Medicine*, vol. 1, no. 3, pp. 285 – 297, 2009.

[5] C. J. Tomlin and J. D. Axelrod, "Biology by numbers: mathematical modelling in developmental biology," *Nature Reviews Genetics*, vol. 8, no. 5, pp. 331 – 340, 2007.

[6] T. Baba, T. Ara, M. Hasegawa, Y. Takai, Y. Okumura, M. Baba, K. A. Datsenko, M. Tomita, B. L. Wanner, and H. Mori, "Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection," *Molecular Systems Biology*, vol. 2, no. 2006.0008, 2006.

[7] H. Mizoguchi, H. Mori, and T. Fujio, "*Escherichia coli* minimum genome factory," *Biotechnology and Applied Biochemistry*, vol. 46, no. 3, pp. 157–167, 2007.

[8] N. Ishii, K. Nakahigashi, T. Baba, M. Robert, T. Soga, A. Kanai, T. Hirasawa, M. Naba, K. Hirai, A. Hoque, P. Y. Ho, Y. Kakazu, K. Sugawara, S. Igarashi, S. Harada, T. Masuda, N. Sugiyama, T. Togashi, M. Hasegawa, Y. Takai, K. Yugi, K. Arakawa, N. Iwata, Y. Toya, Y. Nakayama, T. Nishioka, K. Shimizu, H. Mori, and M. Tomita, "Multiple high-throughput analyses monitor the response of *E. coli* to perturbations," *Science*, vol. 316, no. 5824, pp. 593–597, 2007.

[9] T. Soh and K. Inoue, "Identifying necessary reactions in metabolic pathways by minimal model generation," in *PAIS 2010, Proc. ECAI 2010*, 2010, pp. 277–282.

[10] M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, and M. Tanabe, "KEGG for integration and interpretation of large-scale molecular data sets," *Nucleic Acids Research*, 2011.

[11] I. M. Keseler, J. Collado-Vides, A. Santos-Zavaleta, M. Peralta-Gil, S. Gama-Castro, L. Muiz-Rascado, C. Bonavides-Martinez, S. Paley, M. Krummenacker, T. Altman, P. Kaipa, A. Spaulding, J. Pacheco, M. Latendresse, C. Fulcher, M. Sarker, A. G. Shearer, A. Mackie, I. Paulsen, R. P. Gunsalus, and P. D. Karp, "EcoCyc: a comprehensive database of *Escherichia coli* biology," *Nucleic Acids Research*, vol. 39, no. suppl 1, pp. D583–D590, 2011.

[12] N. Eén and N. Sörensson, "An extensible SAT-solver," in *Proc. the 6th International Conference on Theory and Applications of Satisfiability Testing (SAT 2003)*, 2003, pp. 502–518.

[13] M. Koshimura, H. Nabeshima, H. Fujita, and R. Hasegawa, "Minimal model generation with respect to an atom set," in *Proc. the the 7th International Workshop on First-Order Theorem Proving (FTP '09)*, 2009, pp. 49–59.

[14] G. P. Ferguson, S. Totemeyer, M. J. MacLean, and I. R. Booth, "Methylglyoxal production in bacteria: suicide or survival?" *Archives of Microbiology*, vol. 170, no. 4, pp. 209–218, 1998.

[15] S. Schuster, D. A. Fell, and T. Dandekar, "A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks," *Nature Biotechnology*, vol. 18, pp. 326–332, 2000.

[16] T. Handorf, N. Christian, O. Ebenhöh, and D. Kahn, "An environmental perspective on metabolism," *Journal of Theoretical Biology*, vol. 252, no. 3, pp. 530 – 537, 2008.

[17] T. Schaub and S. Thiele, "Metabolic network expansion with answer set programming," in *Proc. the 25th International Conference on Logic Programming (ICLP 2009)*, 2009, pp. 312–326.

[18] O. Ray, K. E. Whelan, and R. D. King, "Logic-based steady-state analysis and revision of metabolic networks with inhibition," in *Proc. the 2nd International Conference on Complex, Intelligent and Software Intensive Systems (CISIS 2009)*, 2009, pp. 661–666.

[19] O. Ray, T. Soh, and K. Inoue, "Analyzing pathways using asp-based approaches," in *Algebraic and Numeric Biology: Proceedings of the 2010 International Conference (ANB'10)*, ser. LNCS, vol. 6479, 2011, pp. 167–183.

[20] R. Küffner, R. Zimmer, and T. Lengauer, "Pathway analysis in metabolic databases via differetial metabolic display (DMD)," in *German Conference on Bioinformatics*, 1999, pp. 141–147.

[21] T. Takada, A. Mita, A. Maeno, T. Sakai, H. Shitara, Y. Kikkawa, K. Moriwaki, H. Yonekawa, and T. Shiroishi, "Mouse inter-subspecific consomic strains for genetic dissection of quantitative complex traits," *Genome Research*, vol. 18, no. 3, pp. 500–508, 2008.