ECAI 2010
H. Coelho et al. (Eds.)
IOS Press, 2010

doi:10.3233/978-1-60750-606-5-277

Identifying Necessary Reactions in Metabolic Pathways by Minimal Model Generation

Takehide Soh [1,2] and **Katsumi Inoue** [3,1]

Abstract. In systems biology, identifying vital functions like glycolysis from a given metabolic pathway is important to understand living organisms. In this paper, we focus on the problem of finding minimal sub-pathways producing target metabolites from source metabolites. We translate laws of biochemical reactions into propositional formulas and compute its minimal models to solve the problem. An advantage of our method is that it can treat reversible reactions. Moreover the translation enables us to obtain solutions for large pathways. We apply our method to a whole *Escherichia coli* metabolic pathway. As a result, we have found the conventional glycolysis sub-pathway described in a biological database EcoCyc.

1 Introduction

Living organisms are kept alive by a huge number of chemical reactions. In *systems biology*, interactions of such chemical reactions are represented in a network called *pathway*. Analyses of pathways have been active research field in the last decade and several methods have been proposed [7, 18].

A longstanding approach is to represent pathways as systems of differential equations. This method allows detailed analyses e.g. concentrations of each metabolite with time variation. However, it is not applicable to a large network due to its difficult parameter tuning. This is a problem because scalability is an important feature for a macroscopical analysis of complex networks like cells, organisms and life, which is a fundamental goal in systems biology. Therefore other methods aiming for scalable and abstracted analyses have been proposed [2, 13, 12, 16]. Although these methods are different from each others in their problem formalization and solving methods, their purpose is the same, that is, to identify biologically necessary reactions from a given pathway.

One of these methods proposed by Schuster *et al.* is called elementary mode analyses. It focuses on a flux distribution, which is computed by matrix calculus, corresponding to a set of reactions in metabolic pathways [16]. This method can treat multi-molecular reactions while taking into account stoichiometry, and its computational scalability is enough to analyze large pathways. However it tends to generate a large number of solutions without ordering e.g. over 20000 solutions are generated for a pathway including 100 reactions [8]. Even though found solutions are potentially interesting, analyzing all of them through biological experiments would be infeasible task. We thus need a method which generates lower number of solutions keeping its quality. Another approach relying on graphs is proposed by Croes *et al.* [2]. They represent a pathway in a weighted bipartite directed graph and apply a depth-first search algorithm to find the lightest paths from a source compound to a target compound. Planes and Beasley proposed to solve the same problem using a constraint-based method [12]. An advantage of these two methods is that an evaluation of the quality of the solution is provided. We can then choose an objective value to reduce the number of solutions that should be provided to biologists. However, this approach can only generate paths while sub-graphs would be a more natural representation. Moreover, this approach sometimes generates invalid paths from a biological viewpoint because it can easily take non-meaningful shortcuts via common metabolites, such as water, hydrogen and adenosine triphosphate (ATP).

In this paper, we propose a new analysis method for metabolic pathways which identifies *sub-pathways*, whose forms are given as sub-graphs, producing a set of target metabolites from a set of source metabolites. In particular, we formalize the problem of finding minimal sub-pathways, which has the property of not containing any other sub-pathways. That is, all elements of each minimal sub-pathway are qualitatively essential to produce target metabolites. We represent laws of biochemical multi-molecular reactions in propositional formulas and translate the problem into conjunctive normal form (CNF) formulas. We then use a minimal model generator based on state-of-the-art SAT solver to solve the problem efficiently. Our translation and recent progresses in SAT domain now make it possible to apply our method to huge pathways. Realistic metabolic pathways include a lot of reversible reactions. Previous approaches thus needed pre-processing or post-processing, which is possibly costly, to deal with reversible reactions in a pathway [12, 17]. We also show how our method treats such reversible reactions by minimal model generation.

We compare our method with previously proposed approaches [1, 12] for a simplified pathways of *E. coli* consisting of 880 reactions. We also test our method with a whole *Escherichia coli* (*E. coli*) pathway [4] consisting of 1777 reactions. In order to evaluate computed sub-pathways, we use conventional sub-pathways described in the literature [1] and EcoCyc [4], which are provided by biological experiments and existing knowledge. As a result, we have identified every conventional sub-pathway of 11 pathways we used in the experiments.

In the reminder of this paper, we explain propositional formulas and its minimal models in Section 2. In Section 3, we formalize the sub-pathway finding problem. We show the translation from the sub-pathway finding problem into propositional formulas in Section 4. In Section 5, we show the experimental result. In Section 6 and 7 respectively discuss related work and future work.

[1] Department of Informatics, The Graduate University for Advanced Studies, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan, E-mail: soh@nii.ac.jp
[2] Research Fellow of the Japan Society for the Promotion of Science
[3] Principles of Informatics Division, National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan, E-mail: ki@nii.ac.jp

2 Propositional Formulas and Minimal Model Generation

This section reviews propositional formulas and its *minimal models*. Let $V = \{v_1, v_2, \ldots, v_i\}$ be a set of propositional *variables*. A *literal* is a propositional variable v_i or its negation $\neg v_i$. A *clause* is a disjunction of literals. A *conjunctive normal form (CNF) formula* is a conjunction of clauses and is also identified with a set of clauses.

The truth value of a propositional variable is either $true$ (T) or $false$ (F). A *(partial) truth assignment* for V is a function $f : V \rightarrow \{T, F\}$. A literal v_i is said to be *satisfied* by a truth assignment f if its variable is mapped to T; a literal $\neg v_i$ is satisfied by a truth assignment f if its variable is mapped to F. A clause is satisfied if at least one of its literals is satisfied. A *model* for a CNF formula Ψ is a truth assignment f where all clauses are satisfied. Models can also be represented in the set of propositional variables to which it maps T. For instance, the model mapping v_1 to T, v_2 to F, v_3 to T is represented by the set $\{v_1, v_3\}$.

Niemelä report a theorem which is the basis of the computational treatment of *minimal models* [11]. Koshimura *et al.* also report a theorem which is an extension of that theorem [9]. This gives a method to compute a minimal model with respect to a set of propositional variables. We here give a definition and a theorem by [9]:

Definition 1 Let V_p be a set of propositional variables and Ψ a CNF formula. A model I is a *minimal model* of Ψ with respect to V_p iff I is a model of Ψ and there is no model I' of Ψ such that $I' \cap V_p \subset I \cap V_p$.

Theorem 1 Let Ψ be a CNF formula, I a model of Ψ, and V_p a set of propositional variables. I is a minimal model of ψ with respect to V_p iff a formula $\Psi_c = \Psi \wedge \neg(x_1 \wedge x_2 \wedge \ldots \wedge x_i) \wedge \neg y_1 \wedge \neg y_2 \wedge \ldots \wedge \neg y_j$ is unsatisfiable, where $I \cap V_p = \{x_1, x_2, \ldots, x_i\}$, $\overline{I} \cap V_p = \{y_1, y_2, \ldots, y_j\}$.

For instance, suppose that Ψ is a propositional formula $(v_1 \vee v_2) \wedge (\neg v_1 \vee \neg v_2) \wedge (\neg v_2 \vee v_3)$. Then all models of Ψ are $\{v_1\}$, $\{v_2, v_3\}$, $\{v_1, v_3\}$ and the minimal models of Ψ are $\{v_1\}$ and $\{v_2, v_3\}$.

Koshimura *et al.* report a minimal model generator based on a SAT solver by utilizing above theorem (see Figure 1). In the figure, The function `Solve` corresponds to a SAT solver which returns `SAT` and its model when a given formula is satisfiable. The function returns `UNSAT` otherwise.

3 Sub-pathway Finding Problem

This section provides the definition of the *sub-pathway finding problem* on which we are focusing. Let $M = \{m_1, m_2, \ldots, m_i\}$ be a set of metabolites, $R = \{r_1, r_2, \ldots, r_j\}$ a set of reactions, and $A \subseteq (R \times M) \cup (M \times R)$ a set of arcs. A pathway is represented in a directed bipartite graph $G = (M, R, A)$ where M and R are two sets of nodes, A is a set of arcs. A metabolite $m \in M$ is called a *reactant* of a reaction $r \in R$ if there is an arc $(m, r) \in A$. On the other hand, a metabolite $m \in M$ is called a *product* of a reaction $r \in R$ if there is an arc $(r, m) \in A$. A reaction is called a *reversible reaction* if it can occur in either of two directions. We distinguish a reversible reaction as two reactions.

Let $s : R \rightarrow 2^M$ be a mapping from a set of reactions to a set of metabolites such that $s(r) = \{m \in M | (m, r) \in A\}$ represents the set of metabolites which are needed for activating a reaction r. Let $p : R \rightarrow 2^M$ be a mapping from a set of reactions to a set of metabolites such that $p(r) = \{m \in M | (r, m) \in A\}$ represents the set of metabolites which are produced by a reaction r. Let s^{-1} and p^{-1} be inverse mappings of s and p, respectively. Let t be an integer variable representing a time and e be an integer value for a variable t. Let $M' \subset M$ be a subset of metabolites.

Minimal Model Generation Procedure (Ψ, V_p)
begin
  $\Sigma := \emptyset$;
  loop
    (res, I) = **Solve**(Ψ) ;
    if res = `UNSAT` **then** return Σ ;
    else
      $V_x := I \cap V_p$;
      $V_y := \overline{I} \cap V_p$;
      $\Psi_c := \Psi \wedge \left(\bigvee_{x_i \in V_x} \neg x_i\right) \wedge \left(\bigwedge_{y_j \in V_y} \neg y_j\right)$;
      (res, I_c) = **Solve**(Ψ_c) ;
      if res = `UNSAT` **then** $\Sigma := \Sigma \cup \{I\}$;
      $\Psi := \Psi \wedge \left(\bigvee_{x_i \in V_x} \neg x_i\right)$;
end

Figure 1. Minimal Model Generation

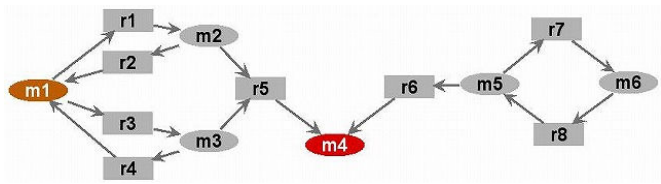

Figure 2. A Pathway including Reversible Reactions

A metabolite $m \in M$ is *producible* at time $t = 0$ from M' if $m \in M'$ holds. A reaction $r \in R$ is *activatable* at time $t = e\,(0 < e)$ from M' if for every $m \in s(r)$, m is producible at time $t = e - 1$ from M'. A metabolite $m \in M$ is *producible* at time $t = e\,(0 < e)$ from M' if $m \in p(r)$ holds for at least one reaction r which is activatable at time $t = e$ from M'. If r is activatable at time $t = e$ then r is activatable at a time $t = e + 1$. If m is producible at time $t = e$ then m is producible at time $t = e + 1$.

Let $M_i \subset M$ be a subset of metabolites representing initial metabolites, $M_s \subset M$ a subset of metabolites representing source metabolites and $M_t \subset M$ a subset of metabolites representing target metabolites. Note that we distinguish M_s from M_i. Every metabolite $m \in M_i$ represents universal metabolites which are always producible in pathways, such as `WATER`, `ATP` and `PROTON`. On the other hand, M_s and M_t represent particular source metabolites and target metabolites in which we are interested, respectively.

Definition 2 Let π be a 6-tuple (M, R, A, M_i, M_s, M_t) and $G = (M, R, A)$ a bipartite directed graph. A sub-graph G' of G is a *sub-pathway* of π if $G' = (M', R', A')$ and it holds the following conditions: (i) $M_s \subset M'$ and $M_t \subset M'$, (ii) for every $m \in M'$, m is producible from $M_i \cup M_s$ at time $t \geq e$ for some $e \in \mathbb{N}$, (iii) for every $r \in R'$, r is activatable from $M_i \cup M_s$ at time $t \geq e$ for some $e \in \mathbb{N}$ and $p(r) \in M'$. In addition, a sub-pathway G' is called *minimal* if it holds that (vi) there is no sub-pathway G'' of π such that $G'' \subset G'$.

Definition 3 *Sub-pathway Finding Problem*

Input A 6-tuple $\pi = (M, R, A, M_i, M_s, M_t)$, where $M = \{m_1, m_2, ..., m_i\}$ is a set of metabolites, $R = \{r_1, r_2, ..., r_j\}$, $A \subseteq (R \times M) \cup (M \times R)$ is a set of arcs, $M_i \subset M$ is a set of initial compounds, $M_s \subset M$ is a set of source compounds, $M_t \subset M$ is a set of target compounds.

Output All minimal sub-pathways of π.

In practice, we compute more restricted solutions of the problem since the number of all minimal sub-pathways tends to be large. We describe how to restrict solutions in the next session.

We here describe the difference between our problem and the *path finding problem* which has been studied [13, 2, 12]. While our problem can treat multiple source metabolites and its outputs are given by sub-graphs satisfying the specific properties, the path finding problem is basically given by the problem of finding paths between a source metabolite and a target metabolite. For instance, we consider a pathway shown in Figure 2. Three sets of metabolites $M_s = \{m_1\}$, $M_i = \{\}$ and $M_t = \{m_4\}$ are given. We find an output $G' = (M', R', A')$ for the input, where $M' = \{m_1, m_2, m_3, m_4\}$, $R' = \{r_1, r_3, r_5\}$ and $A' = \{(m_1, r_1), (m_1, r_3), (r_1, m_2), (r_3, m_3), (m_2, r_5), (m_3, r_5), (r_5, m_4)\}$. On the other hand, the outputs of the path finding problem are two paths $\{m_1, r_1, m_2, r_5, m_4\}$ and $\{m_1, r_3, m_3, r_5, m_4\}$. The point is that the reactions r_1 and r_3 must be needed to be activatable since metabolites m_2 and m_3 are the reactants of the reaction r_5. The output of the sub-pathway finding problem correctly reflects the law of the reaction r_5. However the both outputs of the path finding problem represent the activation of r_5 without producing both necessary reactants. Figueiredo *et al.* summarised problems for path finding approach [2, 13] by a specific example [3]. Obviously, the output of the sub-pathway finding problem correctly reflects the necessary reactions in the pathway.

```
Assign Earliest Activatable Time (M')
begin
  d := 0;
  while (M' ≠ ∅)
    ∀m_i ∈ M', mark m_i as visited;
    M'' := ∅;
    d := d + 1;
    loop for m_i ∈ M'
      loop for unvisited r_j ∈ s^{-1}(m_i)
        if ∀m_k ∈ s(r_j), m_k is visited then
          mark r_j as visited;
          f_e := f_e ∪ {(r_j, d)};
          loop for unvisited m_k ∈ p(r_j)
            M'' := M'' ∪ {m_k};
    M' := M'';
return (f_e, d);
end
```

```
Assign Unique Time (f_e)
begin
  u := 0;
  loop for d ∈ {1, ..., d_max}
    R_sorted := sort {r_i | (r_i, d) ∈ f_e};
    loop for r_j ∈ R_sorted
      u := u + 1;
      f_u := f_u ∪ {(r_j, u)};
return f_u;
end
```

Figure 3. Procedures for f_e and f_u

4 Translation into Propositional Formulas

4.1 Translation of Reaction Laws

This section provides a translation of the sub-pathway finding problem. Let e be an integer for time t and V the set of propositional variables which are used in this translation. Let $rt_{n,e} \in V$ be a propositional variable which is *true* if a reaction $r_n \in R$ is activatable at time $t = e$ and later. Let $mt_{i,e} \in V$ be a propositional variable which is *true* if a metabolite $m_i \in M$ is producible at time $t = e$ and later. For every reaction and time, we have the supplemental formula $rt_{n,e} \rightarrow rt_{n,e+1}$. For every metabolites and time, we have the supplemental formula $mt_{i,e} \rightarrow mt_{i,e+1}$. Let ψ_s be a supplemental formula representing the conjunction of those formulas.

For each reaction r_n, we have the following formula representing that if a reaction r_n is activatable at time $t = e$ then its reactants must be producible at time $t = e - 1$.

$$rt_{n,e} \rightarrow \bigwedge_{m_i \in s(r_n)} mt_{i,e-1} \tag{1}$$

For each reaction r_n, we have the following formula representing that if a reaction r_n is activatable at time $t = e$ then its products must be producible at time $t = e$.

$$rt_{n,e} \rightarrow \bigwedge_{m_j \in p(r_n)} mt_{j,e} \tag{2}$$

In a naive way, above formulas are generated for every time t and every reaction. However it results in the expansion of translated clauses. We thus need to reduce the size of the translated formulas. A time $t = e$ is called the *earliest activatable time* of a reaction $r \in R$ if r cannot be activatable at time $0 < t < e$ and can be activatable $e \leq t$. Let $M' = M_s \cup M_i$ be a set of metabolites, c and d integers, R' the set of reactions which are activatable from M', T a set of integers $\{1, \ldots, |R'|\}$. Let $f_e : R' \rightarrow T$ be a mapping from a set of reduced reactions to a set of integers representing each reaction $r_i \in R'$ and its earliest activatable time $e \in T$. The mapping f_e can be represented in a set of pairs (r_i, e) of a reaction $r_i \in R$ and its earliest activatable time $e \in T$.

We show a procedure *Assign Earliest Activatable Time* to form the mapping f_e in Figure 3. This procedure takes at most $O(|A|)$. Let d_{max} be a constant represents the output integer value d of the procedure. It can also be seen a filtering method for a given π, that is, it deletes the reactions which are not activatable from M'. Moreover, the earliest activatable time is useful to reduce the size of translated formulas. If e is the earliest activatable time for a reaction r then we obviously do not need to consider a time $t < e$ for the reaction. However the size of translated formulas still tends to be large.

Let $f_u : R' \rightarrow T$ be a bijection from a set of reactions to a set of integers representing each reaction and its *unique time*. The mapping f_u can be represented in a set of pairs (r_i, e) of a reaction $r_i \in R'$ and its unique time $e \in T$. In Figure 3, we show a procedure *Assign Unique Time* to form the bijection f_u. To complete the procedure, we need to consider how to sort elements of a set of reactions $\{r_i \mid (r_i, d) \in f_e\}$ for each d (see the line five in the procedure in Figure 3). We use a mapping $deg(r_i)$ which denotes the outdegree of a node r_i. We sort a set of reactions $\{r_i \mid (r_i, d) \in f_e\}$ according to increasing order of the value of $deg(r_i)$.

For each reaction r_n and its unique time $f_u(r_n)$, we have the third formula representing that if a reaction r_n is not activatable then metabolites $m_j \in p(r_n)$ keep its state from time $f_u(r_n) - 1$.

$$\neg rt_{n,f_u(r_n)} \rightarrow \bigwedge_{m_j \in p(r_n)} \left(\neg mt_{j,f_u(r_n)-1} \rightarrow \neg mt_{j,f_u(r_n)}\right) \tag{3}$$

Note that this formula does not mean that if r_n is not activatable then metabolites $m_j \in p(r_n)$ is not producible for any time. Some of those metabolites can be made to producible at a different time by some reactions since each reaction has its unique time. According

to our translation, the cardinality of f_u corresponds to $|R'|$. Thus, the formulas (1), (2) and (3) are generated for only $r_n \in R'$ with its unique time. Although the size of translated formulas is enough tractable, we sometimes cannot find objective solutions since the translation is incomplete.

To extend this limitation, we need to have *step*. Let z be an integer representing step and k an integer variable such that $1 \leq k \leq z$. Let $o_{k,n}$ be an integer such that $o_{k,n} = |R'| * (k-1) + f_u(r_n)$. We have the conjunction of the formulas (1), (2) and (3) as the following formula $D^k_{r_n}$:

$$D^k_{r_n} = \left(rt_{n,o_{k,n}} \rightarrow \bigwedge_{m_i \in s(r_n)} mt_{i,o_{k,n}-1} \wedge \bigwedge_{m_j \in p(r_n)} mt_{j,o_{k,n}} \right) \wedge \left(\neg rt_{n,o_{k,n}} \rightarrow \bigwedge_{m_j \in p(r_n)} \left(\neg mt_{j,o_{k,n}-1} \rightarrow \neg mt_{j,o_{k,n}} \right) \right) \quad (4)$$

Then we have the formula $\bigwedge_{k=1}^{z} \bigwedge_{n=1}^{|R'|} \left(D^k_{r_n} \right)$ which represents the effect of the activation and the inactivation of reactions with step z. In practice, step $z = 3$ is enough to obtain the objective sub-pathways of the pathways we used this time.

4.2 Translation of the Problem

To translate the problem, we need to have an initial condition and a target condition as follows:

$$C(0) = \bigwedge_{m_i \in M_s \cup M_i} mt_{i,0} \wedge \bigwedge_{m_j \in M \setminus (M_s \cup M_i)} \neg mt_{j,0} \quad (5)$$

$$C(|R'| * z) = \bigwedge_{m_i \in M_t} mt_{i,|R'|*z} \quad (6)$$

Finally, we have the translated formula Ψ as follows:

$$\Psi = C(0) \wedge C(|R'| * z) \wedge \psi_s \wedge \bigwedge_{k=1}^{z} \bigwedge_{n=1}^{|R'|} \left(D^k_{r_n} \right) \quad (7)$$

The size of the translated clause is $O(|A|)$. Let I be a model of a given propositional formula Ψ and V_z a set of propositional variables such that $V_z = \{mt_{i,t} \mid mt_{i,t} \in V, t = |R'| * z\} \cup \{rt_{j,t} \mid rt_{j,t} \in V, t = |R'| * z\}$. Let $f_v : V_z \rightarrow M \cup R$ be a mapping such that $f_v(mt_{i,t}) = m_i$ and $f_v(rt_{j,t}) = r_j$. An output of the sub-pathway finding problem is given by the following.

Proposition 1 Given $\pi = (M, R, A, M_i, M_s, M_t)$ and step z, let Ψ be the translated formula as above. If I is a minimal model of Ψ with respect to V_z then $G'' = (M'', R'', A'')$ is a minimal sub-pathway of π, where $M'' = \{f_v(mt_{i,t}) \mid mt_{i,t} \in I \cap V_z\}$, $R'' = \{f_v(rt_{j,t}) \mid rt_{j,t} \in I \cap V_z\}$, and $A'' = \{(m_j, r_i) \mid m_j \in s(r_i), r_i \in R''\} \cup \{(r_i, m_j) \mid m_j \in p(r_i), r_i \in R''\}$.

Note that, by the translation, once a metabolite (resp. a reaction) is made to be producible (resp. activatable), its producibility (resp. activatability) must be maintained until the end due to the supplemental formula ψ_s. We thus need to decode the state of metabolites and reactions only at time $t = |R'| * z$.

4.3 Treating Reversible Reactions

Treatment of reversible reactions frequently becomes a problem in pathway analyses. Some previous approaches took pre-processing or post-processing which breaks reversible reactions in a pathway [1, 12, 17]. Unlike those approaches, our method resolves the problem by considering the notion of activatablity and producibility and finding minimal models of translated formulas.

For instance, we consider the example including reversible reactions shown in Figure 2. Three sets of metabolites $M_s = \{m_1\}$, $M_i = \{\}$ and $M_t = \{m_4\}$ and $z = 1$ are given. A set of variables containing any elements of $\{rt_{6,8}, rt_{7,8}, rt_{8,8}\} \cup \{mt_{5,8}, mt_{6,8}\}$ cannot be a model of the translated formula due to the formula (3). The formula (3) traces the origin of the producibility of the metabolite as well as its state maintenance, that is, if a metabolite is producible at time $t = e$ then the formula (3) guarantees either the metabolite is producible at a time $t < e$ or the reaction is activatable at time $t = e$. Therefore reversible reactions without feeding from $M_s \cup M_i$ are not activatable. Practically, such reactions are deleted by the procedure shown in Figure 3 and we obtain a reduced set of reactions such that $|R'| = 5$. A model I_1 such that $I_1 \cap V_z = \{rt_{1,5}, rt_{2,5}, rt_{3,5}, rt_{4,5}, rt_{5,5}\} \cup \{mt_{1,5}, mt_{2,5}, mt_{3,5}, mt_{4,5}\}$ includes reversible reactions. However it cannot be a minimal model because there is a model I_2 such that $I_2 \cap V_z = \{rt_{1,5}, rt_{3,5}, rt_{5,5}\} \cup \{mt_{1,5}, mt_{2,5}, mt_{3,5}, mt_{4,5}\}$. Finally, we obtain the minimal model I_2 since there is no model such that $I' \cap V_z \subset I_2 \cap V_z$. The minimal model I_2 is decoded to a minimal sub-pathway G_2 consisting of $M_2 = \{m_1, m_2, m_3, m_4\}$, $R_2 = \{r_1, r_3, r_5\}$ and $A_2 = \{(m_1, r_1), (m_1, r_3), (r_1, m_2), (r_3, m_3), (m_2, r_5), (m_3, r_5), (r_5, m_4)\}$.

4.4 Other Biological Applications

Simulating Effects of Deletion of Enzymes. The method allows us to simulate the difference between pathways of wild-type organisms and pathways of mutants or gene knockout organisms. For instance, we can obtain the effect of a gene knock out by removing the reaction r_i related to the gene we want to delete. This is achieved by adding the following formula.

$$\neg rt_{i,|R'|*z} \quad (8)$$

Simulating Effects of Inhibition. In metabolic pathways, each reaction is catalyzed by enzymes. Inhibition relations in some enzymes have been studied through biological experiments. Our method is capable to treat this relation by adding the following formula:

$$\neg rt_{i,|R'|*z} \vee \neg rt_{j,|R'|*z} \quad (9)$$

where reactions r_i and r_j are catalyzed by inhibited enzymes, respectively. This inhibition relation refines output sub-pathways of the method.

Forbidden Metabolites. A further potential application is in drug design, which restricts bi-products by the effect of compounds included in the drug. In this case, we can test by adding drug compounds as sources and unexpected bi-products as forbidden metabolites. This is achieved by adding the following formulas.

$$\bigwedge_{m_i \in M_f} \neg mt_{i,|R'|*z} \quad (10)$$

where M_f is a set of metabolites which are forbidden to be producible. Those constraints are useful to refine outputs when we know such forbidden metabolites in advance.

Table 1. Results for Pathways from [1]

Pathway#	Proposal			[1]		[12]
	#Steps	#Sols.	res.	res. (a)	res. (b)	res.
1	3	1	yes	yes	no	no
2	1	1	yes	yes	no	yes
3	2	37	yes	yes	yes	no
4	1	1	yes	yes	no	no
5	3	4	yes	no	no	yes
6	2	7	yes	yes	no	yes
7	1	1	yes	yes	no	yes
8	3	28	yes	no	yes	no
9	1	4	yes	yes	no	yes
10	1	1	yes	yes	no	yes
Total # of yes in res.			10	8	2	6

5 Experiments and Results

To evaluate the proposed method, we use two reaction databases of *E. coli* K-12. One is the reaction database from supplemental data of the literature [1]. Another one is from a well-known biological database *EcoCyc* [4] which gathers results of biological experiments and existence knowledge of *E. coli*. We downloaded the latest version 13.6 of the reaction database of EcoCyc.

In the following experiments, we use conventional sub-pathways as right solutions, which are respectively obtained from the literature [1] and the database EcoCyc [4]. We modified the Main class of the SAT solver *Minisat* [5] and used it as a minimal model generator shown in Section 2.

Each experiment has been done using a PC (2.53GHz CPU and 2GB RAM) running Ubuntu Linux 9.04. We have developed a graphical user interface integrating the proposed method, which aims for smooth evaluation. To place the nodes, we use the fast organic layout in the Java library Jgraph [6]. In this layout method, vertexes connected by edges should be drawn close to one another and other vertexes should not be drawn to close to one another. Figures 4 and 5 are screen shots of our experimental results shown in Section 5.2 on the interface.

5.1 Comparison with Previous Methods

We compared our method with two previous methods. One is a method using optimization modeling for pathway analyses [1]. An input of this method is a reaction database with stoichiometry. Another one is a constraint based method for path finding [12]. An input of this method is a reaction database without stoichiometry as same as the proposed method. The comparison between these two methods [1, 12] has also shown in the literature [12]. We use same source, initial, and target metabolites according to the literature [1]. As right solutions, the method by [12] used liner paths which are chosen from the conventional sub-pathways of [1]. Similarly, we used those conventional sub-pathways deleted bypass reactions as right solutions.

The results are shown in Table 1. First column shows the following pathways: #1 gluconeogenesis, #2 glycogen, #3 glycolysis, #4 proline bio-synthesis, #5 ketogluconate metabolism, #6 pentose phosphate, #7 salvage pathway deoxythymidine phosphate, #8 Kreb's cycle, #9 NAD biosynthesis, #10 arginine biosynthesis. Each experiment has been done in a second. Second column shows the number of steps where the conventional sub-pathway was found. Third column shows the number of solutions found by the step shown in the second column. Columns 4-7 show each result of whether each method could find the sub-pathway or the path exactly corresponding to the conventional one. In columns 5 and 6, (a) represents the objective of minimizing the total number of reactions and (b) represents the objective of maximizing the production of ATP in the literature [1].

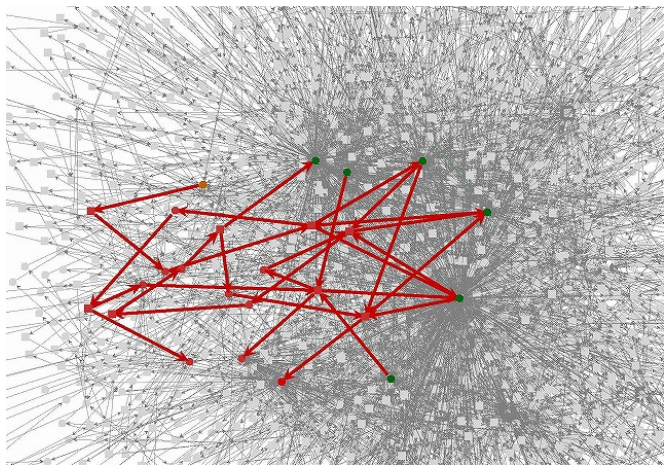

Figure 4. A Glycolysis Sub-pathway on a Whole *E. coli* Pathway

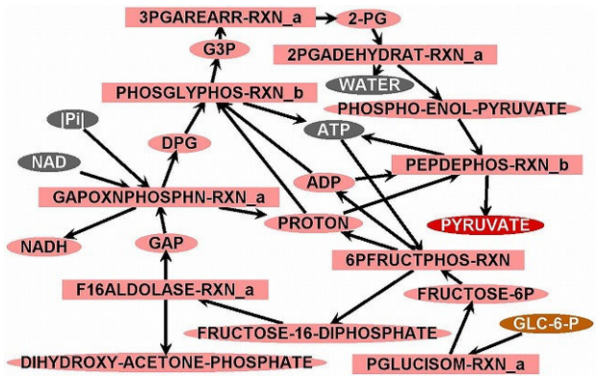

Figure 5. A Glycolysis Sub-pathway of the *E. coli* Pathway

As a result, we found every sub-pathway corresponding to the conventional sub-pathway with step $z \leq 3$. Moreover, the number of solutions are less than 10 except the pathway #3 and #8. Even for these two pathways, we found each conventional sub-pathway in the first 10 solutions by ordering the sub-pathways according to the numbers of reactions. Due to the differences of each input, problem formalization and the number of solutions, it is difficult to make a direct comparison. While the optimization modeling using stoichiometry information by [1] generates one solution for each pathway, it cannot identify two sub-pathways. Constraint based path finding approach [12] outputs the best 10 paths for each pathway but it cannot identify four sub-pathways. Among them, only proposed method identifies all conventional sub-pathways.

5.2 Evaluation on the whole *E. coli* Metabolic Pathway from EcoCyc

We also apply our method to a whole metabolic pathway of *E. coli*. A bipartite directed graph representation of the pathway is shown in Figure 4.

We choose initial metabolites, which are recognized as common metabolites, by calculating percentage of the presence of each metabolites as same as the literature [1]. In order to decide initial metabolites, we define the percentage of the presence of a metabolite m as $pr_m = (n_m \div |R|) \times 100$, where n_m represents the number of reactions in which the metabolite m appears.

According to the value of pr_m, we particularly choose metabo-

lites which are the first 6 of 1073 metabolites: WATER, PROTON, ATP, ADP, |pi| and NAD. In addition, GLC-6-P and PYRUVATE are given as the source metabolite and the target metabolite, respectively. We then apply the method to find a glycolysis sub-pathway in a whole *E. coli* pathway. As a result, we found 4880 minimal sub-pathways and ordered those sub-pathways according to the number of reactions. This experiment has been done in a minute. Figure 5 shows a sub-pathway found in the best 10 solutions corresponding to the conventional glycolysis sub-pathway described in EcoCyc [4]. We here consider the computed sub-pathway shown in Figure 5. All reactions included in the computed sub-pathway are included in the conventional sub-pathway. However, some reactions included in the conventional glycolysis sub-pathway are not included in the computed sub-pathway. This is because conventional sub-pathways from EcoCyc frequently contains bypass reactions which may be needed from a stoichiometry viewpoint. In the case of the glycolysis sub-pathway, TRIOSEPISOMERIZATION-RXN is such a bypass reaction, which consumes DIHYDROXY-ACETONE-PHOSPHATE as a reactant and produces GAP. To support such a bypass reaction is considered to be a future work.

6 Related Work

As far as the authors are aware, the exactly same problem of the sub-pathway finding problem has not yet been formalized. Küffer *et al.* report an approach using a petri net [10]. Although their approach considers producibility and activatability, they do not consider subset minimality of the solution. Schuster *et al.* propose a concept of elementary flux modes and find minimal flux distribution [16]. Although their problem closes to our problem, they use stoichiometry information to solve their problem while our problem only needs the topology of a pathway. Croes *et al.* report the path finding problem with weighted graphs. They add a weight for each metabolite node according to its degree. The results are improved compared with the original graph but there is still a remaining problem shown in the Section 3.

Tiwari *et al.* propose a method using a weighted Max-SAT solver [19] to analyze pathways. They translate reaction laws into soft constraint represented in weighted clauses to compute ordered solutions. However, its ordering is sometimes not acceptable from a biological viewpoint since reaction laws must be held are sometimes violated.

Ray *et al.* report a method using answer set programming (ASP) to compute the steady states of a given pathway and complete lacking reactions [14]. Schaub and Thiele also apply ASP to complete pathways and to identify necessary source metabolites from target metabolites [15]. Unlike their approach, we use minimal model generation to compute essential reactions to produce target metabolites.

7 Conclusion

In this paper, we formalized the sub-pathway finding problem which identifies necessary reactions to produce target metabolites and presented a translation into a propositional formula. Our method uses a SAT solver as a model generator and it has the following features. First, our method can treat reversible reactions without pre-processing and post-processing. Second, it is capable to treat a whole *E. coli* metabolic pathway. Third, it can restrict the number of solutions to be tractable. As far as the authors know, there are few methods have been reported for analyses of a whole organism pathway. We believe that our method provides new analyses for a whole cell and more extended pathways, such as signaling, and gene regulatory networks. Future topics are as follows. For more general evaluation, statistical analyses with more number of pathways are needed. We also need to consider the quality of solutions as well as ranking. Translating more biological knowledge is important to find sub-pathways of more extended pathways.

ACKNOWLEDGEMENTS

This research is supported in part by the 2008-2011 JSPS Grant-in-Aid for Scientific Research (A) (No.20240016) and by the JSPS Research Fellowships for Young Scientists. We would like to thank Gauvain Bourgne and colleagues for their helpful comments. We also thank Oliver Ray for useful discussions.

REFERENCES

[1] John E. Beasley and Francisco J. Planes, 'Recovering metabolic pathways via optimization', *Bioinformatics*, **23**(1), 92–98, (2007).
[2] Didier Croes, Fabian Couche, Shoshana J. Wodak, and Jacques van Helden, 'Inferring meaningful pathways in weighted metabolic networks', *Journal of Molecular Biology*, **356**(1), 222–236, (2006).
[3] Luis F. de Figueiredo, Stefan Schuster, Christoph Kaleta, and David A. Fell, 'Can sugars be produced from fatty acids? a test case for pathway analysis tools', *Bioinformatics*, **24**(22), 2615–2621, (2008).
[4] EcoCyc. http://biocyc.org/download.shtml.
[5] Niklas Eén and Niklas Sörensson, 'An extensible SAT-solver', in *Proceedings of SAT*, pp. 502–518, (2003).
[6] Jgraph. http://www.jgraph.com/pub/jgraphmanual.pdf.
[7] Hidde De Jong, 'Modeling and simulation of genetic regulatory systems: A literature review', *Journal of Computational Biology*, **9**, 67–103, (2002).
[8] Steffen Klamt and Jörg Stelling, 'Combinatorial complexity of pathway analysis in metabolic networks', *Molecular Biology Reports*, **29**(1-2), 233–236, (2002).
[9] Miyuki Koshimura, Hidetomo Nabeshima, Hiroshi Fujita, and Ryuzo Hasegawa, 'Minimal model generation with respect to an atom set', in *Proceedings of FTP'09*, pp. 49–59, (2009).
[10] Robert Küffner, Ralf Zimmer, and Thomas Lengauer, 'Pathway analysis in metabolic databases via differetial metabolic display (DMD)', in *German Conference on Bioinformatics*, pp. 141–147, (1999).
[11] Ilkka Niemelä, 'A tableau calculus for minimal model reasoning', in *Proceedings of the TABLEAU '96*, pp. 278–294, (1996).
[12] Francisco J. Planes and John E. Beasley, 'Path finding approaches and metabolic pathways', *Discrete Applied Mathematics*, **157**(10), 2244–2256, (2009).
[13] Syed Asad Rahman, P. Advani, R. Schunk, Rainer Schrader, and Dietmar Schomburg, 'Metabolic pathway analysis web service (pathway hunter tool at cubic)', *Bioinformatics*, **21**(7), 1189–1193, (2005).
[14] Oliver Ray, Ken E. Whelan, and Ross D. King, 'Logic-based steady-state analysis and revision of metabolic networks with inhibition', in *CISIS*, pp. 661–666, (2010).
[15] Torsten Schaub and Sven Thiele, 'Metabolic network expansion with answer set programming', in *Proceedings of the ICLP '09*, pp. 312–326, Berlin, Heidelberg, (2009). Springer-Verlag.
[16] Stefan Schuster, David A. Fell, and Thomas Dandekar, 'A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks', *Nature Biotechnology*, **18**, 326–332, (2000).
[17] Takeyuki Tamura, Kazuhiro Takemoto, and Tatsuya Akutsu, 'Measuring structural robustness of metabolic networks under a boolean model using integer programming and feedback vertex sets', in *CISIS*, pp. 819–824, (2009).
[18] Marco Terzer, Nathaniel D. Maynard, Markus W. Covert, and Jörg Stelling, 'Genome-scale metabolit networks', *Systems Biology and Medicine*, **1**(3), 285 – 297, (2009).
[19] Ashish Tiwari, Carolyn L. Talcott, Merrill Knapp, Patrick Lincoln, and Keith Laderoute, 'Analyzing pathways using sat-based approaches', in *AB*, pp. 155–169, (2007).