1E1-3

モデル生成を用いた代謝ネットワークにおける極小部分パスウェイの同定

Finding Minimal Sub-pathways in Metabolic Pathways by Model Generation

宋 剛秀[*1] 井上 克巳[*2*1]
Takehide Soh Katsumi Inoue

[*1]総合研究大学院大学 The graduate university for advanced studies
[*2]国立情報学研究所 National Institute of Informatics

In systems biology, identifying vital functions like glycolysis from a given metabolic pathway is important to understand living organisms. In this paper, we particularly focus on the problem of finding minimal sub-pathways producing target metabolites from source metabolites. We represent laws of biochemical reactions in propositional formulas and use a minimal model generator based on a state-of-the-art SAT solver to solve the problem efficiently. An advantage of our method is that it can treat reversible reactions represented by cycles. Moreover recent advances of SAT technologies enables us to obtain solutions for large pathways. We have applied our method to a whole *Escherichia coli* pathway. As a result, we found 5 sets of reactions including the conventional glycolysis sub-pathway described in a biological database EcoCyc.

1. はじめに

生物は非常に多くの化学反応によりその生命活動を維持している．**システム生物学**において，そのような化学反応の相互作用は**パスウェイ**と呼ばれるネットワークによって表現される．近年パスウェイの解析は活発に研究されており，これまで様々なアプローチが提案されてきた．代表的なものの 1 つは微分方程式を用いたアプローチである．この手法はパスウェイの詳細な解析が可能であり，パスウェイに含まれるそれぞれの化合物濃度の時間変化等をシミュレーションすることができる．しかしながら微分方程式中のパラメータチューニングは困難なことが多く，またチューニングが完了しても一般的に微分方程式の解を求めるのも難しい問題である．このため比較的小さなネットワークの解析に対して適用されており，大きなパスウェイの解析には向かない．しかしシステム生物学では細胞，組織，生命全体の解析を目標としているため，大きなネットワークに対する解析手法は必須である．この目標を達成するために，より簡略化された解析であるが，大きなネットワークに適用可能な解析手法が提案されている [Schuster 2000, Croes 2006, Planes and Beasley, 2009]．これらの手法はそのパスウェイの解析に対する問題定式化と適用手法について異なってはいるが，目的は同じある，すなわち与えられたパスウェイから生物学的に意味のある部分パスウェイを同定することである．

そのような従来手法の 1 つに Shuster らによって提案された Elementary mode (EM) 解析がある [Schuster 2000]．この手法では行列計算によって計算される反応流束に着目し，与えられたパスウェイの中から非零の流束を持つ反応を活性化される反応として捉えることができる．この EM 解析では複数入力，出力の化学反応を扱うことができ，大きなパスウェイに対しても適用可能である．しかしながら，EMA は巨大な数の解を出力する，例えば 100 の反応を含むパスウェイに対して 2 万以上の解を出力することが報告されている．これら出力された解の中には，生物学的に興味深いものも存在すると考えられるが，出力された解全てを解析することは不可能である．このため出力される解を少なくする，もしくは生物学的に意味がある解を優先して出力するような手法が必要である．これを解決する手法として Croes らはパスウェイを重み付きの二部有向グラフで表現し，その重みを最小にするような 2 つの頂点間の経路を求める問題を定式化した [Croes 2006]．2 つの頂点として入力代謝物と目標代謝物が与えられ，グラフ探索アルゴリズムにより計算された経路はどのような経路を使用して目標代謝物が生成されるかを表している．また解は重みによって尤もらしい順序で出力される．また同じ問題に対するアプローチとして Planes と Beasley は制約プログラミングを用いた手法で解いている [Planes and Beasley, 2009]．この問題を解くこれら 2 つの手法のメリットは出力された解の質の評価が行えることである．しかしながら，この手法は複数の入出力を持つ化学反応式を考慮しておらず，出力されう解は常に 2 頂点間の経路であり，より多くの情報を持つ部分パスウェイを出力することができない．

本研究では，与えられた入力代謝物より目標代謝物を生成するような部分パスウェイを同定する問題およびその解法を提案する．特に部分パスウェイの中のどの要素を削除しても，削除されたパスウェイが部分パスウェイにならないような**極小部分パスウェイ**に着目する．このようなパスウェイは目標代謝物を出力するために必要不可欠な要素しか含んでおらず，極小部分パスウェイを計算することで解の数を限定しつつ生物学的に興味深い解を出力することができる．本手法では化学反応式を命題論理式で表現し，問題を CNF 式へと符号化する．次に SAT ソルバーを極小モデル生成器として利用し，与えられた命題論理式の極小モデルを求める．近年の SAT ソルバーの進歩は劇的であり，大きなパスウェイに対しても手法を適用可能にできると考えられる．本手法の利点として，複数入出力の化学反応式を扱うことができるということがある．またパスウェイ中に含まれる可逆反応はしばしば手法に前処理，後処理を必要とさせるが [Beasley 2007], 本手法はこれらの処理なしに解を出力することができる．計算機実験において本手法を 1777 の反応と 1073 の代謝物で構成される大腸菌の代謝経路全体に対して適用した．結果として 5 つの極小部分パスウェイが得られ，その中の 1 つは，これまで生物学の分野で知られている解糖系のパスウェイと一致することを確認することができた．

連絡先: 宋剛秀 総合研究大学院大学情報学専攻
〒 101-8430 東京都千代田区一ツ橋 2-1-2
Email: soh@nii.ac.jp

2. 命題論理式と極小モデル

本研究では，パスウェイにおける化学反応式の定性的な性質を**命題論理式**を用いて表現し，その**極小モデル**を求めることによって最終的な解を出力する．手法の説明に必要な定義を以下に述べる．ある命題論理式 Ψ のモデルは，その命題論理式中の命題変数の集合 V から真偽値 $\{True, False\}$ への写像で表現される．この他にモデルは真 $(True)$ に割り当てられた命題変数の集合としても表現可能である．以降，本論文ではモデルを命題変数の集合で表現する．ある 2 つのモデルの大小関係と極小モデルは以下のように定義される [Koshimura 2009]:

Definition 1 V_1, V_2, V_p を命題変数の集合とする．モデル V_1 は V_p に関してモデル V_2 よりも小さいとは，$V_1 \cap V_p \subset V_2 \cap V_p$ が成り立つことをいう．

Definition 2 Ψ_A を命題論理式，V_p を命題変数の集合 V_m を Ψ_A のモデルとする． このとき V_m が V_p に関して Ψ_A の極小モデルであるとは，V_p に関して V_m より小さい Ψ_A が存在しないことをいう．

ここで極小モデルは明らかにモデルの要素数には依存しないことに注意されたい．命題論理式の極小モデルを求める方法はいくつか提案されているが，本研究では越村らによって提案された SAT ソルバーを用いた極小モデル生成法を使用する [Koshimura 2009].

3. 極小部分パスウェイ同定問題

この章では**極小部分パスウェイ同定問題**を定義する．$M = \{m_1, m_2, \ldots, m_e\}$ をパスウェイ上に存在する代謝物の集合，$R = \{r_1, r_2, \ldots, r_f\}$ を反応の集合，$A \subseteq (R \times M) \cup (M \times R)$ を弧の集合とする．パスウェイを二部有向グラフ $G = (M, R, A)$ として表す．もしある代謝物 $m \in M$ に対して (m, r) となる弧が存在するならば，そのとき m を r の**反応物**と呼ぶ．もしある代謝物 $m \in M$ に対して (r, m) となる弧が存在するならば，そのとき m を r の**生成物**と呼ぶ．$s : R \to 2^M, s(r) = \{m \in M | (m, r) \in A\}$ を反応の集合からからその反応物の集合への写像とする．また $p : R \to 2^M, p(r) = \{m \in M | (r, m) \in A\}$ を反応の集合からその生成物への写像とする．s^{-1} と p^{-1} をそれぞれ s と p の逆写像とする．t を時間を表す整数変数，e を整数定数とする．M' を M の部分集合とする．ある代謝物 $m \in M$ が時間 $t = 0$ において M' より**供給可能**であるとは $m \in M'$ となるとき，またそのときに限る．ある反応 $r \in R$ が時間 $t = e\,(0 < e)$ において M' より**活性可能**であるとは，時間 $t = e - 1$ において任意の代謝物に対して $m \in s(r)$ となるとき，またそのときに限る．ある代謝物 $m \in M$ が時間 $t = e\,(0 < e)$ において M' より供給可能であるとは，$m \in p(r)$ に対して少なくとも 1 つの反応 r が時間 $t = e$ において M' より活性可能であるとき，またその時に限る．ある反応 $r \in R$ が時間 $t = e$ において活性可能ならば時間 $t = e + 1$ においても活性可能である．またある代謝物 m が時間 $t = e$ において供給可能ならば時間 $t = e + 1$ においても供給可能である．極小部分パスウェイ同定問題は以下で与えられる．

Definition 3 極小部分パスウェイ同定問題

入力 *6* 項組 (M, R, A, M_i, M_s, M_t) で与えられる．ここで $M = \{m_1, m_2, ..., m_x\}$ は代謝物の集合，$R = \{r_1, r_2, ..., r_y\}$ は反応の集合，$A \subseteq (R \times M) \cup (M \times R)$ は弧の集合，$M_i \subset M$ は初期代謝物の集合，$M_s \subset M$ 入力代謝物の集合，$M_t \subset M$ は目標代謝物の集合である．

出力 *6* 項組 (M, R, A, M_i, M_s, M_t) に関する全ての極小部分パスウェイ $G' = (M', R', A')$．ここで部分パスウェイとは以下の条件 *(i), (ii), (iii)* を満たすパスウェイであり，極小部分パスウェイとは条件 *(vi)* を満たす部分パスウェイである：*(i)* $M_s \subset M'$, $M_t \subset M'$, *(ii)* $\forall m \in M'$, m は $M_i \cup M_s$ から時間 $t \geq e\,(\exists e \in Z^+)$ において供給可能，*(iii)* $\forall r \in R'$, r は $M_i \cup M_s$ から時間 $t \geq e\,(\exists e \in Z^+)$ において活性可能，*(vi)* $G'' \subset G'$ となるような G'' が存在しない．

```
Time Assignment (M')
begin
  enqueue ∀m_i ∈ M'
  mark ∀m_i ∈ M' as visited
  while queue is not empty
    dequeue m_i
    loop for r_j ∈ s^{-1}(m_i)
      if r_j ∉ O_r and s(r_j) ⊆ M' ∪ ⋃_{r_i ∈ O_r} p(r_i) then
        add r_j to O_r
        loop for m_k ∈ p(r_j)
          if m_k is not visited then
            enqueue m_k
            mark m_k as visited
  return O_r
end
```

図 1: 反応への時間割当手続き

実際のパスウェイにおいて全ての極小部分パスウェイは数多く存在するので本研究ではより制限された極小部分パスウェイを求める．解の制限については次の節で詳細を述べる．

4. 命題論理式への符号化

この節では極小部分パスウェイ同定問題の命題論理式への符号化方法について説明を行う．e, f を整数定数とする．命題変数 $rt_{n,e}$ は反応 $r_n \in R$ が時間 $t = e$ において活性されるときに真である．命題変数 $mt_{i,e}$ は代謝物 $m_i \in M$ が時間 $t = e$ において生成される時に真である．それぞれの変数毎に以下の節が得られる：

$$rt_{n,e-1} \to rt_{n,e}, \quad mt_{i,f-1} \to mt_{i,f}$$

任意の反応 r_n に対して，以下の論理式が与えられる．これは反応 r_n が時間 $t = e$ に活性されるのであればその反応物は時間 $t = e - 1$ において供給されていなければならないことを表している．

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

任意の反応 r_n に対して，以下の論理式が与えられる．これは反応 r_n が時間 $t = e$ に活性されるのであればその生成物は $t = e$ において供給されることを表している．

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

素朴な符号化方法ではこれら 2 つの論理式は任意の時間と反応に対して出力されるが，符号化された論理式の増大を招く

ことになる．それを避けるために，本手法ではそれぞれの反応が最も早く活性できる時間を基にユニークな時間を割り当てる．その時間の割り当て方法は幅優先探索を基にした図 1 に示す手順で与えられる．ここで O_r は反応を要素とする順序集合である．この手続きにより各反応には初期代謝物集合と入力代謝物集合から活性可能な最も早い時間がユニークに割り当てられる．ここでパスウェイにおける各代謝物の濃度について，入力または初期代謝物集合から反応を経由するばするほどその生成される代謝物の濃度は低くくなっていくことを前提にしており，これによると最も早い活性可能時間を反応に与えることは濃度が高い代謝物による反応を優先することを意味する．

それぞれの反応 r_n について, 以下の論理式を得る．この式はもし反応 r_n が不活性ならば代謝物 $m_j \in p(r_n)$ は時間 $e-1$ から e においてその状態を保つことを意味している．

$$\neg rt_{n,e} \rightarrow \bigwedge_{m_j \in p(r_n)} (\neg mt_{j,e-1} \rightarrow \neg mt_{j,e}) \qquad (3)$$

この論理式において代謝物 m_j はは他の時間において活性可能であり，反応 r_n が不活性ならば代謝物 $m_j \in p(r_n)$ が供給されないことを表していないことに注意されたい．

ψ_1, ψ_2, ψ_3 をそれぞれ論理式 (1), (2), (3) を表すものとする．反応 r_n の時間 $e-1$ から e の活性化状態は $D_{r_n}(e-1, e) = \psi_1 \wedge \psi_2 \wedge \psi_3$ によって与えられる．本手法では時間 t は基本的に高々反応集合 R の要素数である $n(R)$ で与えられる．もしより多くの解が必要であれば時間 t を $z * t (z > 1)$ に拡張する．ここで z は整数変数である．全ての極小部分パスウェイを計算するには十分に大きな z を考慮する必要があるが，実際のパスウェイでは $z = 5$ で十分である．初期状態と目標状態を以下の論理式によって与える：

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

$$I(n(R) * z) = \bigwedge_{m_i \in M_t} mt_{i,n(R)*z}$$

最終的に符号化された命題論理式 Ψ は $\bigwedge_1^z D_{r_n}(1 \leq n \leq n(R))$, $I(0)$, $I(n(R) * z)$ の連言によって得られる．命題変数の集合 $V_p = \{mt_{i,n(R)*z} | m_i \in M\} \cup \{rt_{j,n(R)*z} | r_j \in R\}$ とし，命題論理式 Ψ の V_p に関する極小モデルを求めることにより，極小部分パスウェイ同定問題の解を求めることができる． なお $V_p = \{mt_{i,n(R)*z} | m_i \in M\}$ とすることで最終的に生成される代謝物に関する極小部分パスウェイを求めることも可能である．

5. 実験結果と考察

5.1 計算環境

計算機実験にあたり，生物学分野においてよく知られた大腸菌のデータベースである *EcoCyc* [Ecocyc] を用いた．このデータベースは過去の生物学における実験データや文献の情報を集積したものであり，今回の計算機実験においては最新バージョンである 13.6 を使用した．データベース中の可逆反応は弧の方向が異なる 2 つの反応として表し，それぞれの反応名の後ろに _a, _b を付加することにより区別している．この実験において，生物学分野における実験によって検証され一般的な知識として定着している代謝経路を実験パスウェイと呼ぶことにする．極小モデル生成器として SAT ソルバー *Minisat 2.0* のコードを変更したものを用いた．計算環境として，1.84GHz の CPU，1GB のメモリを搭載した PC を使用し，解の計算時間は 1 秒以下である．また結果の評価をスムーズに行うために GUI を作成した．そのスナップショットを図 2 と図 3 に示す．図 3 において楕円形の頂点は代謝物，長方形の頂点は化学反応を表す．また代謝物のうち黄色で表されたものは入力代謝物，紫色のものは初期代謝物，ピンク色のものは目標代謝物を表す．大腸菌における解糖系のパスウェイを同定するためにグルコース－ 6 －リン酸 (G6P: Glucose-6-phosphate) を入力代謝物とし，ピルビン酸を (PYRUVATE) を目標代謝物とした．また $z = 1$, 命題変数集合 $V_p = \{mt_{i,n(R)*z} | m_i \in M\}$ と設定した．初期代謝物集合は各代謝物のデータベースにおいて最も出現率の高いものから 6 つを選択した．なおこの出現率の計算は論文 [Beasley 2007] を参考に行った．具体的には 1073 ある代謝物の中から出現率に従って WATER, PROTON, ATP, ADP, |pi|, NAD を初期代謝物とした．

表 1: 計算された大腸菌における極小部分パスウェイ

Reaction Name (by [Ecocyc])	M1	M2	M3	M4	M5	EcoCyc
2PGADEHYDRAT-RXN				○	○	○
3PGAREARR-RXN				○	○	○
6PFRUCTPHOS-RXN		○			○	○
6PGLUCONOLACT-RXN			○			
DLACTDEHYDROGNAD-RXN	○	○				
F16ALDOLASE-RXN		○			○	○
F16BDEPHOS-RXN						○
GAPOXNPHOSPHN-RXN				○	○	○
GLU6PDEHYDROG-RXN			○			
GLYOXIII-RXN	○	○				
KDPGALDOL-RXN			○			
METHGLYSYN-RXN	○	○				
NAD-KIN-RXN			○			
PEPDEPHOS-RXN				○	○	○
PEPSYNTH-RXN						○
PGLUCISOM-RXN	○	○		○	○	○
PGLUCONDEHYDRAT-RXN			○			
PHOSGLYPHOS-RXN				○	○	○
RXN0-313	○			○		
TRIOSEPISOMERIZATION-RXN	○					○

5.2 大腸菌の代謝パスウェイにおける解糖系パスウェイの同定

以上の設定の基に，大腸菌の代謝パスウェイ全体に対して極小部分パスウェイを計算し，結果として 5 つの極小部分パスウェイを同定した (表 1 参照). 結果を有向 2 部グラフで表現したものを図 2，それを拡大したものを図 3 に記載する．表 1 から分かるように極小部分パスウェイ M5 の反応は全ては実験パスウェイの反応と一致していることが分かる．従来多く用いられてきた評価方法に基づくと [Planes and Beasley, 2009], 極小部分パスウェイ M5 は感度 (sensitivity) $Sn = 0.727$, 陽性的中率 (positive predictive value) $PPV = 1$, 精度 (accuracy) $Ac = 0.864$ として評価される．陽性的中率 $PPV = 1$ であるとは計算された極小部分パスウェイにおける反応の集合の要素が全て，実験パスウェイにも含まれていることを表している．感度 $Sn = 0.727$ であるとは，実験パスウェイにおけるいくつかの反応が計算されたパスウェイに含まれていないことを表している．この理由であるが，実験パスウェイはしばしばある代謝物からある代謝物へのバイパス反応を含んでいることがある．極小部分パスウェイではこのようなバイパス反応は考慮されないことが理由の 1 つとしてあげられる．今回の実験では計算された極小部分パスウェイ M5 は実験パスウェイと比較して，反応 PEPSYNTH-RXN が含まれていない．この中で PEPSYNTH-RXN は DIHYDROXY-ACETONE-PHOSPHATE から GAP を生成する反応であり，反応 F16ALDOLASE-RXN のバイパス経路となっている．このようなバイパス経路が必要な理由であるが，生物学的には生成

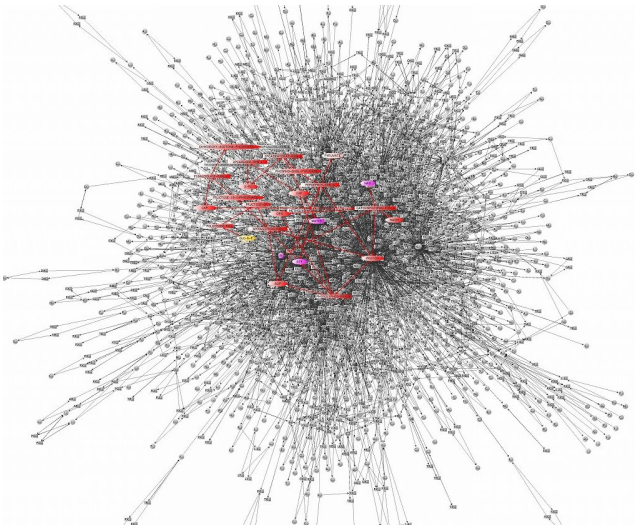

図 2: 大腸菌の代謝パスウェイと解糖パスウェイ (M5)

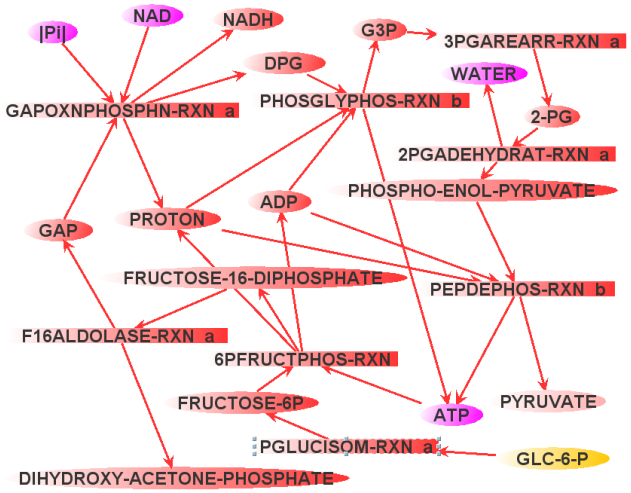

図 3: 計算された解糖パスウェイ (M5)

物 GAP の生成量をコントロールするために必要であることが知られている．現在の問題定式化ではこのような経路を解として出力することはできないため，これを考慮することは今後の課題の 1 つである．他に 2 つの反応 TRIOSEPISOMERIZATION-RXN と F16BDEPHOS-RXN が含まれていないが，これらの反応は一見して必須でないように見える．この 2 つの反応が実験パスウェイに含まれている理由を考察すること，他の 4 つの極小部分パスウェイについて生物学的な評価および考察を行うことは重要な課題である．

5.3 関連研究

本研究は [宋 2009] の手法をより大きなパスウェイに適用できるように改良したものであり，次の点で従来のものと異なる．従来の手法における命題論理式への符号化手法では出力される部分パスウェイの中に閉路が含まれる可能性があった．そのため比較的小さなパスウェイに適用する際には出力された解に閉路が含まれるかどうかをチェックすることで必要な解を抽出することができるが，今回の大腸菌の代謝パスウェイ全体のように大きなパスウェイに対しては解の数が非常に多くなり適用することが困難であった．本論文における極小部分パスウェイの定義および符号化では，そのような閉路を含む部分パスウェイは解にならず，よって大きなパスウェイに対しても解の数が制限され，手法を適用することが可能となっている．また計算の際に与える整数変数 z を変更することで活性可能時間に基づいた順序で複数の解を出力することが可能である．

6. おわりに

本研究では目標代謝物を生成するのに必要である化学反応夫の集合を同定する極小部分パスウェイ同定問題を定式化し，その命題論理式への符号化方法を提案した．本手法では符号化された命題論理式より極小モデルを求める方法として SAT ソルバーを極小モデル生成器として用いた．評価のために大腸菌の代謝パスウェイ全体に対して本手法を適用し，結果として生物学において認められている解糖系の実験パスウェイを計算によって求めることに成功した．本手法の良い点の 1 つとして大きなパスウェイに対しても解の数を抑えながら計算可能な点が挙げられる．これは今後より大きな代謝パスウェイを解析する際，またシグナル伝達パスウェイや遺伝調節パスウェイのように異なったパスウェイを横断した解析を行う際に重要な特徴である．今後の課題として，出力された解を生物学的視点から解析し，専門家の意見を取り入れながらより精度の高いパスウェイを出力できるように手法を改良することが挙げられる．

参考文献

[Beasley 2007] Beasley, J. E. and Planes, F. J. : Recovering metabolic pathways via optimization, *Bioinformatics*, vol. 23, 1, pp.92-98, 2007.

[Croes 2006] Croes, D., Couche, F., Wodak, S. J., and Helden, J. V.: Inferring Meaningful Pathways in Weighted Metabolic Networks, *Journal of Molecular Biology*, vol. 356, 1, pp.222-236, 2006.

[Ecocyc] `http://biocyc.org/download.shtml`

[Koshimura 2009] Koshimura, M., Nabeshima, H., Fujita, H., and Hasegawa, R.: Minimal Model Generation with respect to an Atom Set, Proceedings of FTP'09, 2009.

[Planes and Beasley, 2009] Planes, F. J., and Beasley, J. E.: Path finding approaches and metabolic pathways, *Discrete Applied Mathematics*, vol. 157, 10, pp.2244-2256, 2009.

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

[Tamura 2009] Tamura, T., Takemoto, K., and Akutsu, T.: Measuring structural robustness of metabolic networks under a Boolean model using integer programming and feedback vertex sets, IIBM2009, pp.819–824, March, 2009.

[宋 2009] 宋　剛秀，井上克巳: SAT 問題への変換を用いたフィードバックを含むパスウェイの解析, 人工知能学会全国大会, 高松, 2009.