

多値論理を用いた生体ネットワークシステムのモデル検査 (2012年5月28日版)

坂本 悠^{1,a)} 山本 泰生² 岩沼 宏治²

概要: 近年、遺伝子制御系、シグナル伝達系や代謝系などの生体機構をひとつのシステムとして再構築する研究が進められている。実験機器のハイスループット化に伴い、生体システムに関与する観測データは急速に増加しており、システムと観測データ間の整合性を組織的に検証する技術が必要となってきた。本論文では、遺伝子制御系のようなネットワーク形式で表現される生体システムを多値論理の枠組みでモデル検査する手法を提案する。また酵母のグルコース抑制機構に関する生体ネットワークに対して本手法を適用した結果を報告する。

キーワード: モデル検査, 生体ネットワーク, 多値論理

Model Evaluation of Biological Networks with Multiple-valued Logic (version 2012/5/28)

HARUKA SAKAMOTO^{1,a)} YOSHITAKA YAMAMOTO² KOJI IWANUMA²

Abstract: Recently, a systematic approach has been evolved in biology to reconstruct biological mechanisms involved in genome, proteome and metabolome into one whole system. Along with high-throughput experimental tools like microarrays, it is required to analyze the consistency between a large amount of data and those reconstructed systems. In this paper, we propose a qualitative way with multiple-valued logic that enables to systematically evaluate those systems that are represented as networks, like gene regulatory networks. This paper also reports preliminary experimental results obtained by applying our technique to a biological network on the glucose repression system of *S. cerevisiae*.

Keywords: Model evaluation, biological networks, multiple-valued logic

1. はじめに

分子生物学の発展に伴い、遺伝子解析が急速に進み、各遺伝子からどのようなタンパク質がどのように生成されるか理解が深まってきている。近年、このような個々の生命機構をひとつのシステムとして再構築を図るシステム生物

学 [1][2][3] の分野が注目されて来ている。

DNA マイクロアレイや、クロマチン免疫沈降 (CHIP) などのハイスループットな実験手法が開発された結果、網羅的な実験データが近年得られるようになってきた [1]。これに伴い、システムと実験データ間の整合性を検証する技術が必要となってきた。本論文では、システムと実験データを定性的に解釈することで論理的にそれらの整合性を検証する手法を検討する。定性的解釈を用いる手法の利点は、システム検証を行うユーザーに対して論理的に考える全ての解 (無矛盾な解釈) を提示することができる点である。また、解がない、すなわちシステムと実験データが矛盾する場合には、矛盾を引き起こした原因を列挙する

¹ 山梨大学大学院医学工学総合教育部コンピュータ・メディア工学専攻

Interdisciplinary Graduate School of Medicine and Engineering, University of Yamanashi

² 山梨大学大学院医学工学総合研究部

Department of Research Interdisciplinary, Graduate School of Medicine and Engineering, University of Yamanashi

^{a)} g11mk022@yamanashi.ac.jp

ことも可能である。ここで重要となるのは、現実と出来るだけ乖離しないような定性的解釈を用いる必要がある点である。

論理ベースの生体ネットワークとして良く使用されているものに、ブーリアンネットワーク (*Boolean networks*) がある [1]。ブーリアンネットワークは、遺伝子制御などさまざまな分子間相互作用を表現する際に用いられる最も基礎的な数理モデルの一つである [1]。ネットワークを構成する各頂点は2つの状態を取るブール変数とみなされる。また各 (有向) 辺は始点となる頂点の論理積が終点となる頂点の状態に相当することを示す。このようにブーリアンネットワークを用いる場合、各頂点の状態を2値に解釈することが必要不可欠となる。例えば遺伝子ネットワークの場合、頂点は各遺伝子の発現状態を発現しているかそうでないかの2値に表す必要がある。しかしながら、検証に用いる観測データは必ずしもいつも2値の状態に解釈できるわけではなく、本来は多値を取る状態を2値に近似して利用する場合がある。

その一例として文献 [4] を取り上げる。文献 [4] では、出芽酵母のグルコース抑制機構がブーリアンネットワーク形式で表現され、このシステムの整合性が遺伝子発現データを用いて検証されている。遺伝子発現データは、野生株といくつかの変異株において機構に関与する各遺伝子の発現量である。この観測データから、統計的検定をもとに、野生株に対する変異株の発現量が有意に増加しているか減少しているか変化していないかがわかる。文献 [4] では、システムがブーリアンネットワーク形式を取るため、この3値の発現変化量のデータを、発現しているかどうかの2値に近似し、システムの検証を行っている。しかし、この解釈は実際の観測データの意味から乖離したものであり、検証結果に誤差が生じる可能性がある。

そこで本稿では、3値論理の枠組みを導入することで、観測データの持つ本来の意味論のもとで、グルコース抑制機構の生体ネットワークを再検証する。また再検証の中で、定量データから定性状態への解釈の違いにより、検証結果にどのような違いが出るかを示す。

本稿の構成は以下の通りである。2章では、先行研究として文献 [4] の生体ネットワークを紹介する。3章では、提案手法を示す。4章では、まとめと今後の課題を述べる。

2. 背景

この章では、まず、文献 [4] で構築された論理モデルについて説明し、その後モデル検査について説明する

2.1 グルコース抑制の論理モデル

文献 [4] では、グルコース抑制に関与する遺伝子とタンパク質間の分子間インタラクションの論理モデルを構築している。グルコース抑制 [5] は、周囲の環境にあるグル

コース等の好みの糖を感知し、そのほかの糖の利用を抑える糖代謝の調節のことである。グルコース抑制は、ほとんどの真核細胞を持つ基礎的な機構のひとつであり、糖尿病など糖代謝の疾患を理解する上でも重要となる。文献 [4] は、このグルコース抑制機構をハイパーグラフを用いて表現している。(有向) ハイパーグラフとは各辺に2つ以上の複数の頂点をゆるす (有向) グラフである。ここで、ある頂点が外部頂点であるとは、その頂点を終点とする有向辺が存在しないとき、またそのときにかぎる。また、内部頂点とは、外部頂点以外の頂点をいう。文献 [4] で作成されたハイパーグラフは、頂点数が83個ある。外部頂点は18個あり、内部頂点は65個ある。各頂点は、遺伝子頂点、たんぱく質頂点、代謝頂点の3種類あり、それぞれ遺伝子発現の有無、タンパク質の活性と不活性、代謝物の有無を表すブール変数と見なされる。辺の数は124本で構成されている。各辺は、その辺を構成する頂点間の因果関係を表し、終点の頂点の値を決定するために用いられる。例えば図1のグラフは、 x_1 を終点とする2つの辺から構成されている。このとき、 x_1 のブール値は以下のDNF式により決定される。

$$x_1 = \neg x_2 \vee (x_3 \wedge x_4). \quad (1)$$

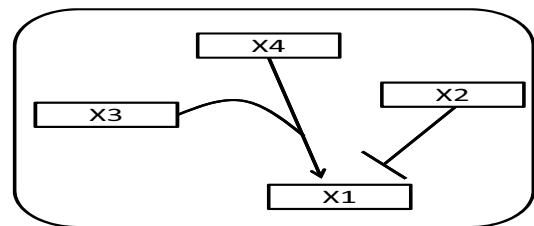


図1 ハイパーグラフの例

文献 [4] で構築された論理モデルは、(1) のような DNF 式が複数集まり構成されているため、ひとつのブーリアンネットワークとみなすことができる。文献 [4] で構築されたハイパーグラフの全体を図2に示す。

2.2 モデル検査

モデル検査に必要なものは、各頂点の因果関係を論理式で表した集合と、観測データである。本稿で使用する観測データは、野生株及び $\Delta mig1$, $\Delta mig1 \Delta mig2$, $\Delta grr1$ の3つの変異株の各遺伝子の発現量である。野生株とは、自然環境で最もよく見られる型の株である。また、変異株とは、野生株と異なる形質を獲得した株である [6]。例えば $\Delta mig1$ は、遺伝子 $mig1$ の機能を除外して得られる変異株を意味する。このとき、遺伝子 $mig1$ をノックアウト遺伝子という。ここで、観測データから得られた遺伝子 i の発現変化を表す変数の値を Y_i^{exp} とおく。 Y_i^{exp} の定義域は以

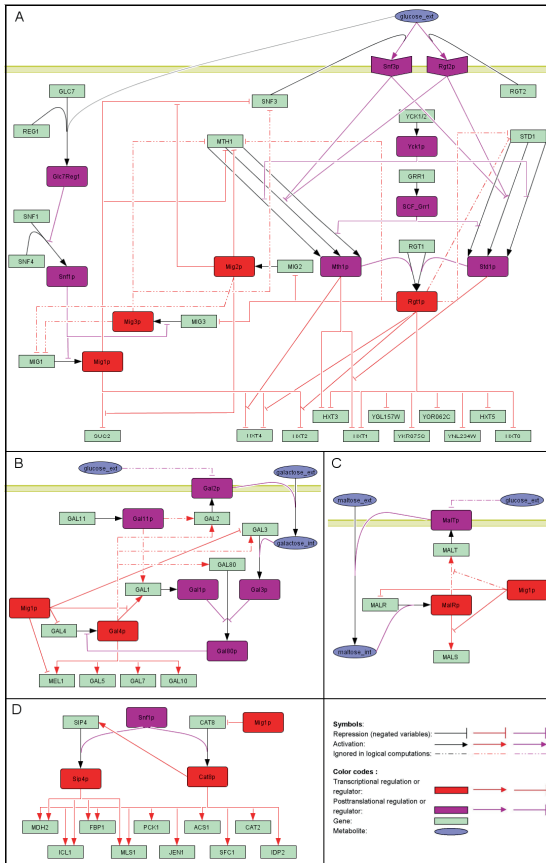


図 2 グルコース抑制の論理モデル

下の通りである。

$$Y_i^{exp} \in \{-1, 0, 1\}.$$

値が-1ならば、野生株と比べ、発現量が有意に減少していることを表し、0ならば、有意に差はないことを表している。また、1ならば野生株と比べ発現量が有意に増加していることを表している。

文献 [4] では、次にモデルによる遺伝子発現の予測 (2 値) を以下のようにして求めている。はじめ全ての外部頂点に 1 を割り当てる。例えば、外部頂点が遺伝子頂点の場合、全て遺伝子発現有と見なすことになる。次にノックアウト遺伝子 x の頂点が遺伝子発現の場合 (変異株 Δx に相当) と遺伝子発現内の場合 (野生株に相当) で、それぞれ各内部頂点の論理定常状態を計算し、その差を得る。 $Y C_i^{mod}$ をこの差、すなわちモデルによる遺伝子 i の発現変化予測とし、 $Y C_i^{exp}$ を実験による遺伝子 i の発現変化とする。このとき、 $Y C_i^{mod}$ と $Y C_i^{exp}$ の定義域は以下の通りとなる。

$$Y C_i^{mod} \in \{-1, 0, 1\}, Y C_i^{exp} \in \{-1, 0, 1\}. \quad (2)$$

(2) より、モデル予測と実験の観測との差は以下のようになる。

$$|Y_i^{mod} - Y_i^{exp}| \in \{0, 1, 2\}.$$

上記の差が 0 であれば、モデルの予測と実験の観測が一

致したことを示し、それ以外は予測失敗だったことを示す。文献 [4] は、この予測精度を整合性の基準として使用している。表 1 に 3 つの変異株に対してモデル検査を行った結果を示す。

表 1 論理モデルの精度

変異株	評価された遺伝子の数	正しく予測できた遺伝子の数	正しく予測できた遺伝子数の割合
$\Delta mig1$	38	24	63%
$\Delta mig1 \Delta mig2$	38	24	63%
$\Delta grr1$	38	15	39%

しかしながら、外部頂点を全て遺伝子発現有とする根拠はない。そこで、観測データから遺伝子発現の有無を推定すると、必ずしもすべての外部頂点に遺伝子発現有と割り当てられていなかった。例えば遺伝子 $RGT1$ は、 $\Delta grr1$ の時に遺伝子発現量が上昇した。そのため、野生株は遺伝子発現無で、変異株で初めて発現したと推測できる。この推測方法では文献 [4] より妥当である。そこで、外部頂点にこのような近似方法で観測データから推測される遺伝子発現の有無を割り当てた。外部頂点に遺伝子発現の有無を割り当てることで、内部頂点の遺伝子発現の有無が推定される。このような近似のもとでモデルが求まる株 $s (s \in \{ \text{野生株}, \Delta grr1, \Delta mig1, \Delta mig1 \Delta mig2 \})$ における遺伝子 i の発現有無を $yc_i^{mod}(s)$ とし、観測データを 2 値に解釈したときの株 $s (s \in \{ \text{野生株}, \Delta grr1, \Delta mig1, \Delta mig1 \Delta mig2 \})$ における遺伝子 i の発現有無を $yc_i^{exp}(s)$ とする。このとき、 $yc_i^{mod}(s)$ と $yc_i^{exp}(s)$ の定義域は以下の通りである。

$$yc_i^{mod}(s) \in \{0, 1\}, yc_i^{exp}(s) \in \{0, 1\}. \quad (3)$$

値が 0 ならば、遺伝子発現していないことを表し、値が 1 ならば、遺伝子発現していることを表している。(3) より、モデルによる予測と観測データとの差は以下のようになる。

$$|yc_i^{mod}(s) - yc_i^{exp}(s)| \in \{0, 1\}.$$

上記の差が 0 であれば、モデルの予測と観測データが一致したことを示し、それ以外は予測失敗だったことを示す。外部頂点の割り当てを修正した後の結果を表 2 に示す。

表 2 2 値を使用した評価結果 (外部頂点の割り当て修正後)

変異株	評価された遺伝子の数	正しく予測できた遺伝子の数	正しく予測できた遺伝子数の割合
$\Delta grr1$	38	22	57%
$\Delta mig1$	38	28	73%
$\Delta mig1 \Delta mig2$	38	29	74%

全ての変異株で、正しく予測できた遺伝子の割合が 100% ならば、モデルと観測データは整合性が取れている

と言える。しかし、 $\Delta grr1$, $\Delta mig1$, $\Delta mig1\Delta mig2$ 全てで100%とならなかった。このことから2値解釈の元ではモデルと観測データの整合性が取れないと結論付けられる。

2.3 モデル検査の問題点

文献 [4] のモデル検査には、観測データの解釈を2値に近似したことで、実際と乖離している。なぜならば、もともとの観測データは、野生株と変異株の遺伝子発現量の変化を3値で表現しているためである。文献 [4] では外部頂点の状態を全て1に近似している。これを観測データを元に修正したものの、表2のように予測精度はまだ十分高いとは言いがたい。そこで、本稿では、3値論理の枠組みを導入することで観測データの本来の意味論のもとでモデルの差異犬種王を行う。

一方、文献 [4] では、ブーリアンネットワーク形式のモデルを用いるため、3値の観測データを2値に近似し、モデル検査をおこなっている。そこで、本論文では、3値論理の枠組みで、文献4のモデルを再検証する。

3. 3値論理を使用した検査

3.1 3値を利用したモデル検査

各遺伝子 i の遺伝子発現量の変化 Y_i^{exp} は、観測データから直ちに求まる。外部頂点に、 Y_i^{exp} の値を代入することで、各内部頂点の遺伝子発現変化をモデルにより予測する。モデルで予測した内部遺伝子 i の値を Y_i^{mod} とする。 Y_i^{mod} と Y_i^{exp} の定義域は、 $Y_i^{mod} \in \{-1, 0, 1\}$, $Y_i^{exp} \in \{-1, 0, 1\}$ となる。

各内部頂点の値を求めるには、ブーリアンネットワークの入力が発現有無の2値ではなく発現変化の3値となるようネットワークの意味論を修正する必要がある。ここでは、3値(増加, 減少, 変化なし)を取る2変数 A と B に対し、(AかつB)と(AまたはB)のとり値について検討する。3値論理として、ウカシェヴィッチの3値論理とクリーネの3値論理 [7][8] に基づく真理値表は生物学的見地から利用することができない。判例として、MTH1 と SNF3 の遺伝子転写調節機構を上げる。2つの機構はそれぞれ次のような論理式で表現される。

$$MTH1 = \neg mig1p \vee \neg mig2p$$

$$SNF3 = \neg mig1p \wedge \neg mig2p$$

ここで、 $mig1p$ と $mig2p$ はそれぞれ2つのタンパク質の活性変化を表すとする。ウカシェヴィッチの3値論理の場合、 $mig1p$ と $mig2p$ がそれぞれ減少と変化なしになると、MTH1 は変化なしになる。SNF3 は減少になる。また、クリーネの3値論理も同様の結果となる。

2つの機構における遺伝子発現量は、文献 [9] で調査されている。 $\Delta mig1$ の場合、MTH1 は増加、SNF3 は変化なし。 $\Delta mig1\Delta mig2$ の場合、MTH1 は増加、SNF3 は増加、

$\Delta mig2\Delta mig3$ の場合は、MTH1 は増加する。SNF3 は変化なし。以上のことから、ウカシェヴィッチの3値論理とクリーネの3値論理もともに使用できない。本稿では、文献 [9] で示された実験結果と矛盾しない3値論理の真理値表(表3)を用いる。

表3 3値論理の真理値表

A	B	AかつB	AまたはB
増加	増加	増加	増加
増加	変化なし	変化なし	増加
増加	減少	不明	不明
変化なし	増加	変化なし	増加
変化なし	変化なし	変化なし	変化なし
変化なし	減少	変化なし	減少
減少	増加	不明	不明
減少	変化なし	変化なし	減少
減少	減少	減少	減少

表3におけるAとBはある遺伝子頂点の発現変化量を表す。また、「不明」は「増加」、「減少」、「変化なし」のいづれも当てはまらないことを表す。すなわち、文献 [9] の実験結果から一意に定まらないことを意味する。もし「不明」と割り当てられた場合その遺伝子はいかなる観測データとも一致していないとする。表3を元に各内部頂点の変化量 Y_i^{mod} を求める観測データと一致するか確認した。このモデル結果を表4に示す。

表4 3値を使用した評価結果

変異株	評価された遺伝子の数	正しく予測できた遺伝子の数	正しく予測できた遺伝子数の割合
$\Delta grr1$	38	17	43%
$\Delta mig1$	38	25	66%
$\Delta mig1\Delta mig2$	38	22	58%

表3と表2を比べたとき、本手法の結果の方が、2値論理ベースの結果より精度が低くなっていることがわかる。本手法では、外部頂点に割り当てられる状態は「変化なし」であるものが多く、それに伴い内部頂点の多くも「変化なし」と割り当てられる。このように「変化なし」と割り当てられた内部頂点において観測データと不一致となる傾向が見られた。

3.2 整合性のとれる割り当ての列挙

2章や3.1節で求めた精度は、ある閾値を設けて変化量を3値に離散化している。閾値とは、その値を境にし、それ以上であれば変化したと認め、それ以下ならば、変化を認めない境界となる値のことである。しかし、各インタラクションに対し同じ閾値を用いなければならない理由はない。例えば、各インタラクションによって、トリガーとなる閾値は異なると考えてもよい。もし、閾値を固定しないなら、各変化量は「変化なし」、もしくは「変化している」(「増加」または「減少」)の2つの状態を分けることが出来

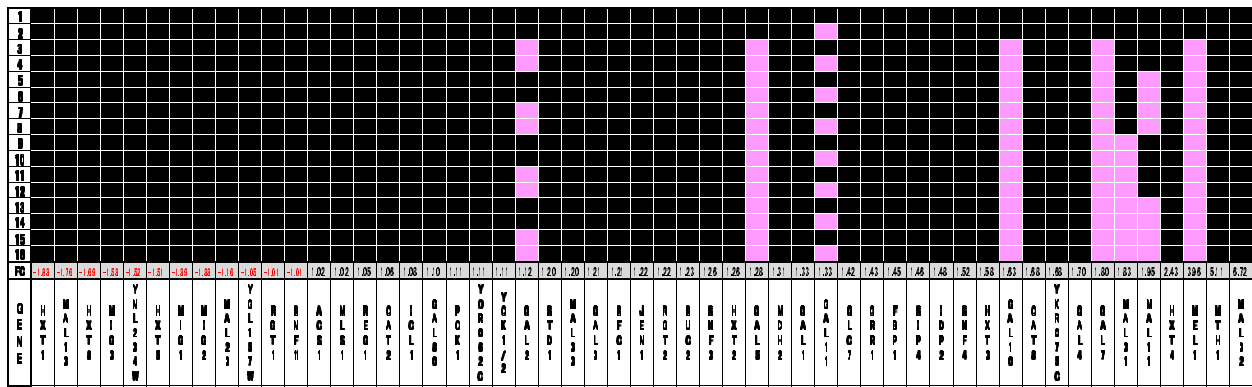


図3 Δmig1の整合性の取れる割り当て

る。ただし、発現量が「増加」しているにも関わらず、「減少」状態に割り当てられることは出来ない。

このように、閾値一意に固定しない場合、3値論理の枠組みでモデルと矛盾しないような変化状態の割り当てが存在するか確かめることが出来る。図3ではΔmig1の観測データにおいて整合性の取れる変化状態の割り当てを列挙している。

図3の黒色で塗っている箇所は、「変化なし」を割り当てたことを表す。また、桃色で塗っている箇所は、「増加」を割り当てたことを表す。ここで、Δmig1は「減少」を割り当てた場合、整合性が取れなかったため、「減少」を表す色は塗られていない。各遺伝子名の上を書いてある数値は、野生株と比べ遺伝子発現量の変化を示している。0以下の数値の場合、野生株と比べ減少していることを示す。また、遺伝子の並びは、発現変化量の順にソートされている。Δgrr1とΔmig1Δmig2の変異株については、全ての内部頂点に「変化なし」が割り当てられたときのみしか、整合性が取れなかった。

今回の結果は2つの解釈が立てられる。1つ目は、もし、各遺伝子ごとに閾値を与えるならばモデルと整合性の取れるような遺伝子発現の割り当てが存在することである。2つ目は、閾値を一意に定める場合、このモデルでは、どんな閾値を与えても整合性を取ることができない、すなわち、もし仮に整合性が取れるならば、閾値以下は黒く塗られ、閾値以上は桃色で塗られるような割り当てが存在するはずである。しかしながら、図3ではいずれの割り当ても黒と桃色が入れ子になって塗られている。以上からこのモデルには何らかの欠陥な因果関係が存在すると推測される。モデルの修正を行う場合は、図3のMTH1とMAL32の遺伝子が候補となる。なぜならば、この2つの遺伝子より発現変化が少ない遺伝子に桃色が塗られているためである。このような遺伝子の因果関係をもう一度調査することで新しい因果関係が発見される可能性がある。

3.3 閾値の変化による精度の変化

次に、閾値の変化によって検証結果に違いが出るか検証する。先行研究[4]では、閾値を1.5と固定しているが、1.5と一意に定める理由は特にない。実際、システム検証を行うユーザーによって、トリガーとなる変化量は異なることが通常である。図4, 5, 6は、各変異株における観測データにおいて、閾値を1.2から6まで変化させたとき、3値論理に基づく提案手法と2値論理に基づく従来手法で予測精度がどのように推移するかを示している。

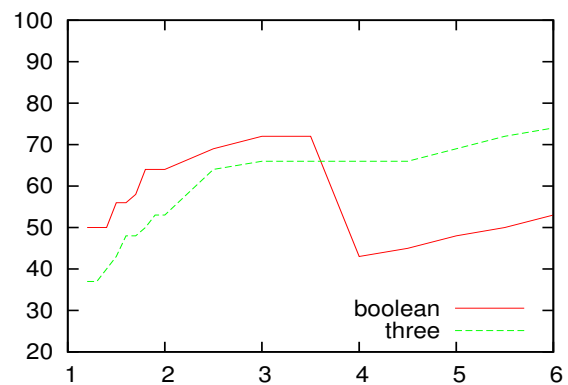


図4 Δgrr1の各閾値による精度

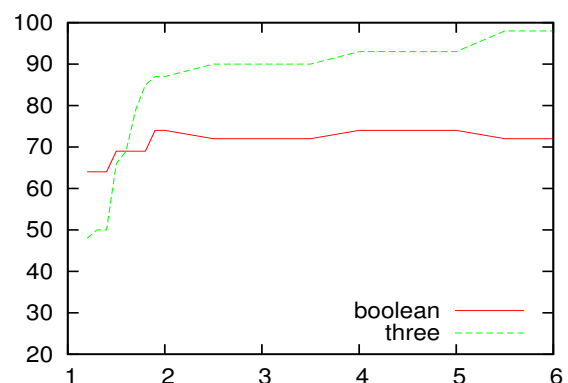


図5 Δmig1の各閾値による精度

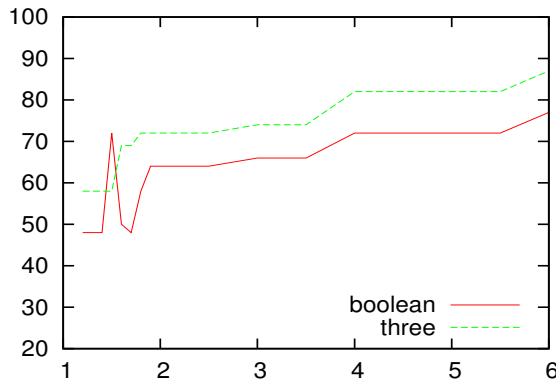


図 6 $\Delta mig1\Delta mig2$ の各閾値による精度

3つの図から、いずれの変異株においても、閾値をあげた場合、本手法でモデル検証を行った方が精度が向上している。 $\Delta mig1$ と $\Delta mig1\Delta mig2$ では、閾値が1.6以上で本手法のほうが従来手法より精度が高くなっており、 $\Delta grr1$ では、閾値を3.0以上とすることで本手法の精度が上回る結果となっている。

実際、閾値を上げた場合、すなわち、変化したと認める境界の値が上昇すると、「変化なし」と割り当てられる遺伝子が相対的に多くなる。一方、本手法によるモデル予測では、内部頂点が「変化なし」となる傾向が見られる。このため、閾値を上げた場合には、本手法の予測精度が向上すると考えられる。

4. まとめと今後の課題

生体システムと観測データの整合性を定性的に検証する場合、定量データのある定性状態に解釈して用いる必要がある。本稿では、酵母菌のグルコース抑制機構のブーリアンネットワークを例にとり、この定量データから定性状態への解釈によって、システム検証の結果が大きく変わることを示した。本稿で取り扱っている遺伝子発現データは、遺伝子ネットワークなどさまざまな生体ネットワークの検証に利用されている。遺伝子発現データを定性状態に解釈する場合、発現の有無ではなく、発現の変化を対象とする必要がある。このためには、少なくとも「増加」、「減少」、「変化なし」の3つの状態を用意し、3値論理に基づいてネットワークを再解釈する必要がある。そこで本稿では、生物学実験の結果に基づいた3値論理の真理値表を用意し、その表により、グルコース抑制機構のモデルを再検証した。さらに定性状態へ解釈する上で必要となる閾値を変動させた場合、検証結果にどのような影響があるか調査した。本稿で示した手法は遺伝子発現データを用いるその他のブーリアンネットワークモデルの検証にも利用することが可能である。

本稿で行なった再検証の結果、グルコース抑制機構の現行モデルにはいくつか欠落している因果関係があることが

予想される。先行研究 [10] では、2値論理の枠組みでこのような因果関係を論理的に推定する研究を行っている。今後の課題として、先行研究の手法を3値論理の枠組みに拡張しモデルを再構成する課題に取り組む。

謝辞

本研究は一部、文科省科学研究費補助金（若手 B：No.22700141）および文科省科学研究費補助金（基盤 C：No.22500127）の援助を受けている。

参考文献

- [1] 宮野悟, 江口至洋, 金久實, 高木利久, 中井謙太. バイオインフォマティクス辞典, 共立出版, 2006.
- [2] H. Kitano. All systems go, *Nature Reviews Drug Discovery* 7, pp. 278-279, 2008.
- [3] 江口至洋. 細胞のシステム生物学, 共立出版, 2008.
- [4] T. S. Christensen, A. P. Oliveira and J. Nielsen. Reconstruction and logical modeling of glucose repression signaling pathways in *Saccharomyces cerevisiae*, *BMC Systems Biology* 3:7, 2009.
- [5] S. L. Westergaard, A. P. Oliveira, C. Bro, J. Nielsen. A systems biology approach to study glucose repression in the yeast *Saccharomyces cerevisiae*, *Biotechnology and Bioengineering* 134-145, 2007.
- [6] 今堀和友, 山川民夫. 生化学辞典第4版, 東京化学同人, 2007.
- [7] M. Bergman. An Introduction to Many-Valued and Fuzzy Logic: Semantics, Algebras, and Derivation Systems. Cambridge University Press, 2008.
- [8] R. Turner, 松田利夫訳, 石本新訳. 人工知能と論理, 共立出版, 1989.
- [9] M. Bergman. An Introduction to Many-Valued and Fuzzy Logic: Semantics, Algebras, and Derivation Systems. Cambridge University Press, 2008.
- [10] A. Kaniak, Z. Xue, D. Macool, JH Kim, and M. Johnston. Regulatory network connecting two glucose signal transduction pathways in *Saccharomyces cerevisiae*. *Eukaryot Cell* 3:221-23, 2004.
- [11] 坂本 悠, 山本 泰生, 岩沼 宏治. 論理モデルによるグルコース抑制機構のパスウェイ補完. 人工知能学会全国大会 (第25回), 2011.