Vol.2016-BIO-45 No.11
2016/3/19

分子ネットワーク上の状態推定とその可視化による知識発見支援

平沼 祐人[1,a)]　山本 泰生[2,3,b)]　守屋央朗[5]　宋 剛秀[4]　岩沼 宏治[2]

概要：本稿では，転写因子結合ネットワーク上の各転写因子の状態をマイクロアレイデータから定性的に推定する手法を提案する．転写因子は遺伝子発現を調整する重要な因子であり，転写因子の状態推定を行うことは，転写機構によって制御される細胞反応のダイナミクスの解明に繋がる．提案手法では，はじめに状態推定問題を最適化問題として定式化した後，推定された状態の評価モデルとして，単独の転写因子による影響のみを考慮したもの（単独推定）と複数因子の相互作用の影響も考慮したもの（複数推定）の２つの問題を考える．一般に後者は解候補の組み合わせ爆発に関わる問題であるが，本研究ではこの組み合わせ最適化問題を制約充足問題に帰着し，高速な SAT 技術を用いて求解している．更に，転写因子の活性状態の全体図を捉えることができる可視化ツールを開発している．このツールにより，生物学者と対話しながら，直感を刺激することで知識発見支援を行う．

Estimating and Visualizing State Transitions of Transcription Factors from Perturbation Experiments

Yuto Hiranuma[1,a)]　Yoshitaka Yamamoto[2,3,b)]　Hisao Moriya[5]　Soh Takehide[4]　Koji Iwanuma[2]

1. はじめに

近年，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] の分野が注目され，このシステム生物学的アプローチにより生命機構全体の形式化及び包括的な解析が可能になってきている [2]．また，マイクロアレイや ChIP-chip 等のハイスループット実験技術により，細胞内の遺伝子転写機構の詳細が明らかになっている．

[1] 山梨大学大学院コンピュータ・メディア工学専攻
Department of Computer Science and Media Engineering, Univ. 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, Univ. of Yamanashi
[3] 独立行政法人科学技術振興機構，さきがけ
JST, Presto
[4] 神戸大学情報基盤センター
Information Science and Technology Center, Kobe Univ.
[5] 岡山大学異分野融合先端研究コア
Research Core for Interdisciplinary Sciences, Okayama Univ.
[a)] g14mk013@yamanashi.ac.jp
[b)] yyamamoto@yamanashi.ac.jp

モデル生物である出芽酵母では転写制御機構のネットワークモデルとして**転写因子結合ネットワーク** (Transcription Factor Binding Network) が構築されている．転写因子とは，遺伝子発現の制御・調節を担うタンパク質であり，結合する各遺伝子に対して，**活性化** (*activate*) と**阻害** (*inhibit*) の制御を行い，活性化された遺伝子の発現量は増加し，阻害された遺伝子の発現量は減少することが知られている．本研究では，転写因子結合ネットワークを用いて，マイクロアレイデータから各転写因子の活性状態を推定する課題に取り組む．遺伝子発現を制御する転写因子の状態推定解析は，大量のデータからどの転写機構が活性化しているかを知ることができ，多彩な細胞反応の表現型の理解に繋がる．その一方で，転写因子結合ネットワークモデルは大規模・複雑化しており，人手でそのモデルを網羅的に解析することは現実的に困難である．また，推定結果からパスウェイの活性情報を知るには深い専門知識や前提知識が必要であり，それらが存在しない場合には調査対象を絞り込むには非常に時間がかかる．そこで本研究では，計算

機により転写因子結合ネットワーク上の全ての転写因子状態を高速に推定し，その解析結果を可視化する知識発見支援ツールを開発する．

2. 準備

本章では，本研究で用いる転写因子結合ネットワークとマイクロアレイデータについて説明し，転写因子の活性状態について定義する．

2.1 転写因子結合ネットワーク

転写因子結合ネットワーク (Transcription Factor Binding Network: 以下，TFBN と略す) とは，転写因子とそれに結合して制御される，またはその制御のカスケード効果により間接的に制御される遺伝子関係を表したネットワークモデルである．TFBN の頂点は，転写因子または遺伝子であり，直接結合関係を持つ有向パスは ChIP-chip データにより実験的に確かめられている．有向パスが存在する場合，転写因子が遺伝子を直接もしくは間接的に制御している．本研究で用いる TFBN は，2012 年に Yang らがまとめたものであり [3]，出芽酵母における 112 個の転写因子，5105 個の遺伝子を含むネットワークである．このとき，転写因子頂点の集合を V_t，遺伝子頂点の集合を V_g とする．転写因子 $t \in V_t$ と遺伝子 $g \in V_g$ 間に有向パスが存在するとき，パスは次のデータ構造によってまとめられている．

- TFBN 上の転写因子 t
- t からの有向パスを持つ遺伝子 g
- t が g へ及ぼす影響を示す $labe_{tg} \in \{1, -1\}$ ただし，$labe_{tg} = 1$ のとき活性化制御，$labe_{tg} = -1$ のとき阻害制御を意味する．
- t から g への最短パス上の中間頂点の列

この 4 つから成るデータを**結合パス**と呼ぶ．このとき，全転写因子-遺伝子間の結合パス集合を P，ある転写因子 t から出発する全結合パス集合を P^t，ある遺伝子 g へ向かう全結合パス集合を P^g と呼ぶ．P の内 $label = 1$ であるものを P_A，そうでないものを P_I と書く．結合パス集合 P が与えられたとき，遺伝子 g と転写因子 t が P 中のあるパスでつながるとき，$(t, g) \in P$ と書く．

2.2 遺伝子発現データ

遺伝子発現の実験データとして，生物学的摂動付加前 ($control$) と付加後 ($sample$) の TFBN 上の遺伝子発現量の比 $fold\text{-}change$ (以下，fc と略す) を利用する．統計検定後，fc はマイクロアレイの実測値から直ちに求まるが，あるしきい値 $threshold$ により離散化し，有意に上昇，あるいは減少したかを判定することが一般的である．fc を離散化した値を示す fc_d は次の式で表される．

$$fc_d = \begin{cases} up & (fc > threshold) \\ down & (fc < \frac{1}{threshold}) \\ stable & (otherwise) \end{cases}$$

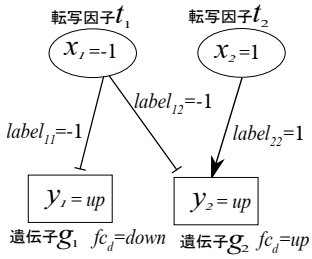


図 1 遺伝子 g_j の発現変化 y_j の導出例

$fc_d = up$ の場合，$control$ に対して $samplel$ の発現量が有意に増加していることを意味する．同様に，$fc_d = down$ の場合は有意に減少，$fc_d = stable$ の場合は発現量が有意に変化していないことを意味する．また，どれだけ fc が増加，減少したかを示す指標として，fc_r を導入する．fc_r を次のように定義する．

定義 1 $$fc_r = \begin{cases} fc & (control < sample) \\ \frac{1}{fc} & (control > sample) \end{cases}$$
ただし，$fc_r \geq 1$ である

TFBN 上にある全遺伝子の観測データ集合を EX とする．

本論文の目的は，実験で得られた各遺伝子の発現変化と TFBN を元に，その遺伝子を制御する転写因子の状態を推定することにある．利用する遺伝子発現データには，TFBN 上の遺伝子 703 個の fc が記されており，移行の実験では，特に断りのない限り $threshold = 2.0$ を基準に設定している．

2.3 転写因子の状態とモデル上の状態伝搬モデル

TFBN 上の転写因子の状態は次のように定義される．

定義 2 TFBN 上の転写因子頂点 t の状態を $x_i \in \{1, -1, 0\}$ とする．$x_i = 1$ は up，$x_i = -1$ は $down$，$x_i = 0$ は $stable$ を意味する．TFBN 上の各転写因子 $t_1, t_2, \ldots, t_k$ の状態割り当てを $A = (x_1, x_2, \ldots, x_k)$ と書く．ただし，k は転写因子の数（ここでは 112 個）とする．

転写因子の状態を与えることで, TFBN 上で制御している遺伝子の発現変化を定性的に導出できる．ある遺伝子 g_j を制御している転写因子 $t_i \in P^g$ に状態 x_i を与えたとき，遺伝子 g_j の遺伝子発現変化を次のように定義する．

定義 3 $t_i \in P^g$ に状態 x_i を与えたときの g の遺伝子発現変化を $y_j \in \{up, down, stable\}$ とする．ただし，$1 \leqq j \leqq l$ であり，l は t が制御している遺伝子数である．

次に，g_j に t_i が与える影響を次のように定義する．

定義 4 g_j に t_i が与える影響を $E^g_{ij} \in \{1, -1, 0\}$ とする．ただし，E^g_{ij} は $E^g_{ij} = x_i \times labe_{tg}$ によって与えられる．

y_j の導出を次のように定義する．

定義 5 $$y_j = \begin{cases} up & (\forall E^g_{ij} = 1) \\ down & (\forall E^g_{ij} = -1) \\ stable & (\forall E^g_{ij} = 0) \end{cases}$$

図 1 に TFBN 上の遺伝子発現変化の導出例を示す．転写因子 t_1，t_2 に対して，$x_1 = -1$，$x_2 = 1$ を与えたとき，遺伝子 g_1 に対して t_1 は $label_{11} = -1$ で制御しており，$E^g_{11} = 1$ である．従って g_1 の発現変化 $y_1 = up$ が導出される．また，g_2 に対して t_1 は $label_{12} = -1$，t_2 は $label_{22} = 1$ で制御しており，$E^g_{12} = 1$，$E^g_{22} = 1$ である．従って g_2 の発現変化 $y_2 = up$ が導出される．

3. 問題設定

本章では，推定問題の最適化問題への定式化と推定精度について説明を行う．

3.1 推定問題の定式化

2.3 節で述べたように，TFBN 上で制御している遺伝子の発現変化を定性的に導出できる．この結果と遺伝子発現の実験データを用いることで，転写因子の状態を推定することができる．

定義 6 転写因子の状態割り当てを A，全結合パスを P，遺伝子発現データを EX とする．このとき，$Precision(A, P, EX)$ を P と EX に関する A の評価関数とする．

$Precision(A, P, EX)$ の設定により推定結果の評価方法，すなわち推定精度が決定する．

3.2 評価関数の設定

はじめに，単独の転写因子の影響のみを考慮した推定（以下，単独推定と呼ぶ）の評価関数について述べ，次に複数の転写因子の影響を考慮した推定（以下，複数推定と呼ぶ）の評価関数について述べる．

3.2.1 単独推定の評価関数

転写因子 t に制御される全遺伝子の観測データ集合を EX_i としたとき，以下を定義する．

定義 7 状態 x_i のもと，t の全結合パス集合 P^t により，遺伝子発現変化が導出された遺伝子集合を I_i と書く．その変化状態が EX_i 中の変化状態と一致する I_i 中の遺伝子集合を $G(x_i, P^t, EX_i)$ とする．

このとき，単独推定における転写因子 t の評価関数を次のように定義する．

定義 8 $Precision(x_i, P^t, EX_i) = \frac{|\ G(x_i, P^t, EX_i)\ |}{|\ EX_i\ |}$

各転写因子の推定精度を定義 3，全転写因子数を n としたとき，単独推定の評価関数 $Precision_s(A, P, EX)$ は次のように定義される．

定義 9 $Precision_s(A, P, EX)$
$= \frac{1}{n} \sum_{i=1}^{k} Precision(x_i, P^t, EX_i)$

$Precision_s(A, P, EX)$ は非負の $Precision(x_i, P^t, EX_i)$ へ線形分離可能であり，各転写因子の評価関数が最も高くなる状態の割り当て A を個々に選択すれば，推定精度は最大となる．

図 1 において g_1 は $fc_d = down$，g_2 は $fc_d = up$ の観測データを持つとき，$x_1 = -1$ の推定結果に対して正しく導出できた遺伝子が g_2，正しく導出できなかった遺伝子が g_1 となる．従って t_1 の単独推定における推定精度は，$Precision(x_1, P^t_1, EX_1) = \frac{1}{2}$ となる．同様にして $x_2 = 1$ の推定結果に対して正しく導出できた遺伝子は g_2 となる．従って，$Precision(x_2, P^t_2, EX_2) = \frac{1}{1}$ となる．これより，$Precision_s(A, P, EX) = \frac{3}{4}$ となる．

3.2.2 複数推定の評価関数

定義 10 状態割り当て A のもと，全結合パス P により，遺伝子発現変化が導出された遺伝子のうち，その変化が EX 中の変化と一致する遺伝子の集合を G(A,P,EX) とする．

このとき，複数推定における推定精度を次のように定義する．

定義 11 $Precision_m(A, P, EX) = \frac{|\ G(A, P, EX)\ |}{|\ EX\ |}$

図 1 において g_1 は $fc_d = down$，g_2 は $fc_d = up$ の観測データを持つとき，それぞれの転写因子に $x_1 = -1$，$x_2 = -1$ を割り当てると，正しく導出できた遺伝子が g_2，正しく導出できなかった遺伝子が g_1 となる．従って t_1 の複数推定における推定精度は，$Precision_m(A, P, EX) = \frac{1}{2}$ となる．

複数の転写因子の影響を考慮した推定では，転写因子の状態の割り当ての組み合わせを考える必要があり，評価関数の値が最も高くなるような 112 個の転写因子の活性状態の割り当て A を求めるには，3^{112} 通りの組み合わせを対象とする必要がある．従って，すべての組み合わせを総当たりする力任せ法は組み合わせ爆発に対応できない．対して関連研究 [4] では，分枝限定法を用いた探索アルゴリズムを提案し，推定解析を行っている．

4. 関連研究

本章では，関連研究 [4] で行った実験結果及びその考察を述べる．

4.1 単独推定及び複数推定の実験結果

単独推定，複数推定を用いて TFBN 上の転写因子の状態を推定した．単独推定ではの 112 個の転写因子の状態推定を 0.01 秒で高速に行うことができた．一方，複数推定では，TFBN 中からランダムに抽出した 60 個の転写因子の状態推定におよそ 28000 秒（7.5 時間）かかることから，分枝限定法を用いた探索アルゴリズムにおいても 112 個の転写因子を実用可能時間内に推定することはできない．

4.2 考察及び評価

複数推定は単独推定に比べて膨大な計算時間を要する．

ただし，複数推定は元ネットワークモデルの意味論により則した評価方法であり，生物学的見地から，より望ましい転写因子状態が得られる可能性がある．実際に，同一の転写因子 60 個に対して単独推定で得られた転写因子状態と複数推定で得られた転写因子状態の双方をよりネットワークの意味論に則している複数推定の評価モデルにより比較したところ，正しく導出できなかった遺伝子数が，単独推定では 466 個であったのに対し，複数推定では 361 個と 23% 程度，導出に失敗する遺伝子数が減少した．このことからも複数推定がよりモデルの意味論に則した転写因子状態を得られていることが分かる．そこで本研究では，最適化問題を制約充足問題に変換し，SAT ソルバを用いることで複数推定を高速に行う．

5. SAT ソルバを用いた推定

本章では，複数推定における最適解を SAT ソルバを用いて求めるために必要な定式化について説明する．

5.1 0-1 変数上の制約最適化問題

制約最適化問題では，与えられた制約を満たす中で目的関数が最適値を取るような解を求める．複数推定問題におけるブール変数，制約，目的関数は以下のように定義される．

ブール変数

- 転写因子: $p_{t,s} \in \{0,1\}$ は転写因子 t の状態が s のときに $p_{t,s} = 1$ となる．
- 遺伝子: $q_{g,s} \in \{0,1\}$ は遺伝子 g の状態が s のときに $q_{g,s} = 1$ となる．
- 補助変数: 各遺伝子 g について矛盾をあらわす変数 b_g を導入する．

制約

- 転写因子が状態を 1 つしかとれないことを表す制約は以下になる．
$$\bigwedge_{t \in V_T} (p_{t,down} + p_{t,stable} + p_{t,up} = 1)$$
- 遺伝子が状態を 1 つしかとれない，もしくは，矛盾を起こすことを表す制約は以下になる．
$$\bigwedge_{g \in V_G} (b_g = 1) \vee (q_{g,down} + q_{g,stable} + q_{g,up} = 1)$$
- 遺伝子 g の fc_d を $fc_d(g)$ と書くと遺伝子の状態が観測データに従うことを表す制約は以下になる．
$$\bigwedge_{g \in V_G} q_g, fc_d(g) = 1$$
- 転写因子の状態割り当てと辺の関係から遺伝子の状態を与えることができる制約は以下になる．
$$\bigwedge_{(t,g) \in P_A} (p_{t,up} = 1 \rightarrow q_{g,up} = 1) \wedge (p_{t,stable} = 1 \rightarrow q_{g,stable} = 1) \wedge (p_{t.down} \rightarrow q_{q,down} = 1)$$
$$\bigwedge_{(t,g) \in P_I} (p_{t,up} = 1 \rightarrow q_{g,down} = 1) \wedge (p_{t,stable} = 1 \rightarrow q_{g,stable} = 1) \wedge (p_{t.down} \rightarrow q_{q,up} = 1)$$

目的関数

$$minimize \sum_{g \in V_G} b_g$$

上記で定義した制約充足問題は SAT 型制約プログラミングシステム scarab[5] を用いて求解できる．実験では，学習節を保持し，逐次制約を追加していくインクリメンタル解法を用いて求解する．探索は二分探索を行い，探索幅があるしきい値以内になった時点で線形探索を行うような混合探索を採用している．SAT ソルバは Sat4j[6] を用いた．

5.2 予備実験及び考察

112 個の転写因子の複数推定は 208 秒で最適解を求めることができた．分枝限定法では転写因子 60 個の推定に 28000 秒かかったことから，少なくとも 40 倍以上の速さで求解できる．その一方で，最適解では全ての転写因子の状態が $stable$ となった．遺伝子発現変化は転写因子の影響を受けて発生するため，転写因子の状態が全て $stable$(変化なし) になることは考えにくい．原因として，推定に用いたマイクロアレイデータ中で $fc_d = stable$ である遺伝子の割合が大きいため，転写因子の状態を全て $stable$ とすることで正しく遺伝子の状態を導出できることにある．

6. 評価関数の再設計

本章では，各観測データの重要度を考慮した評価関数を検討する．

6.1 発現量が有意に変化している遺伝子を対象とする推定

観測データ中の発現量が有意に変化している遺伝子を対象に推定を行う．結果は，112 個の転写因子の推定におよそ 30 分かかり，最適解は転写因子の状態が up，$down$ のいずれかに推定された．ただし，最適解は複数存在し，少なくとも 100 以上の解が存在することが確認できた．

6.2 fc_r に重みをつける推定

マイクロアレイ等の実験データを得たとき，遺伝子発現の変化量 fc_r が大きい遺伝子に対してより強い興味を示す．その興味の強さを表す指標 $Interest$ を導入する．$Interest$ は，観測データ内の各遺伝子に与えられる値であり fc_r に影響を受けるものとする．$fc_r \geq 1$, $threshold \geq 1$ である．このとき，正しく導出された遺伝子の $Interest$ が最大となるような状態を求める問題を考える．すなわち，6.1 節の評価モデルを一般化した推定問題を考える．

図 3 に fc_r に対する 3 種の重み付けモデルを示す．

6.2.1 ステップモデル

ステップモデルは $fc_r \geq threshold$ の遺伝子集合に対して，fc_r の値に関わらず一定の興味度 c を持つモデルである．ステップモデルは次の式によって表される．

$$Interest = \begin{cases} c & (fc_r \geq threshold) \\ 0 & (otherwise) \end{cases}$$

6.2.2 線形モデル

線形モデルでは fc_r の値に対して線形的に興味度が上昇

表 1　生物学的検証

	ステップモデル	線形モデル		指数モデル	
		$threshold = 2$	$threshold = 1$	$threshold = 2$	$threshold = 1$
最適解中に出現する状態する	$up, down$	$up, down$	$up, down, stable$	$up, down$	$up, down$
最適値（$Interest$ の合計値）	420	431	306	425	315
誤り遺伝子数	238	245	211	244	238
計算時間 [s]	90.3	484.5	43963.2	484.5	40246.1

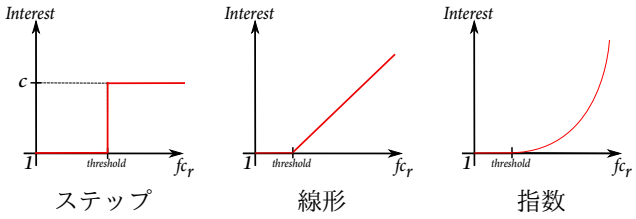
するモデルである．線形モデルは次の式によって表される．

$Interest = a \times (fc_r - threshold)$ (ただし，$a > 0$ の定数)

6.2.3　指数モデル

指数モデルでは fc_r の値に対して指数的に興味でが上昇するモデルである．指数モデルは次の式によって表される．

$Interest = a^{(fc_r - threshold)} - 1$ (ただし，$a > 1$ の定数)



ステップ　線形　指数

図 2　重み付けモデル

6.3　検証実験

まず，線形モデルにおいて $a = 1$，$threshold = 1$ という設定で実験を行った．これは fc_r の値をそのまま $Interest$ 値とするのと同義である．その結果，最適解では，全転写因子の状態が $stable$ になるという問題は回避された．ただし，112 個の転写因子の推定にかかる計算時間は 16 時間と非常に大きい．そこで，各モデルの比較検証においては，遺伝子数を 322 個に縮退したネットワークを用いる．このとき，対象となる遺伝子は，3 つの状態がほぼ均一に出現するようにランダムに抽出している．また，値域幅のばらつきを抑えるため取り得る最大の最適値を統一している．表 1 に検証結果を示す．

結果，いずれのモデルにおいても最適解で，全転写因子の状態が $stable$ になる問題は回避されている．最適値が最も良いのは $threshold = 2$ の線形モデルであり，誤り遺伝子数が最も少なくなったのは $threshold = 1$ の線形モデルであった．計算時間は，$threshold = 1$ の線形，指数モデルが共に 11 時間以上かかっている．これは，全遺伝子の $Interest$ を考慮した推定を行わなければならないためである．一方，ステップモデル及び $threshold = 2$ の線形，指数モデルは $Interest = 0$ となる遺伝子を考慮しないため計算時間が短い．

7.　生物学的検証

TFBN の TF ノックアウトデータ [7]（以下、k データと呼ぶ）を用いて，推定手法の精度を確認した．単独推定と複数推定について、3 章の評価関数 ($stable$ あり) と 6.1 節の評価関数 ($stable$ なし) を利用した推定を行っている．表 2 に結果を示す．表中の値は，それぞれの転写因子に対応する k データを用いて推定を行った際，k データ中の転写因子状態と推定により得られた状態が一致する転写因子数である．k データ中の大部分が $fc_d = stable$ であり，$stable$ ありの推定では単独，複数ともに全ての解が $stable$ になため，全体の 7 割り程の一致数あった．一方，$stable$ なしの推定では単独，複数ともに一致数がかなり低くなった．

ここで，転写因子 $yap6$ に注目したい．$yap6$ は TFBN 上の転写因子であり 63 個の遺伝子を制御している．しきい値を大きく下げ，k データ中の遺伝子が全て有意に変化しているとしたとき，$yap6$ に状態 up を与えると 63 個全ての遺伝子に対して正しく解を導出することができる．$yap6$ はノックアウトされているために，k データ内では僅かではあるが，確かに発現量が減少している．原因は調査中であるが，この結果から新たな知見が得られるのではないかと期待している．

表 2　重み付けモデルの検証結果

単独推定（stable あり）	複数推定（stable あり）	単独推定（stable なし）	複数推定（stable なし）
76/112	76/112	3/112	1/112

8.　可視化ツール

生物学者に対する知識発見支援の方法として，視覚的刺激から直感的なひらめきを誘発させるようなツールの作成を目指す．本章では，既存の可視化ツールを紹介した後，作成した可視化ツールについて説明する．

8.1　既存可視化ツール

既存の可視化ツールの代表的なものに **Cytoscape** [8] がある．Cytoscape はグラフ描画に適しており，巨大なデータを扱うことのできるオープンソースソフトウェアである．また，様々な生物学的なメタ情報を含むネットワークがパッケージとして用意されているため，多くの生物学者が利用している．本研究では，JavaScript 用のライブラリである **Cytoscape.js** [9] を用いて HTML 上で動作する可視化ツールを開発する．

8.2　システム要件

開発する可視化ツール（以下，本システムと呼ぶ）のシステム要件を示す．

8.2.1 本システムの目的

生物学的知識発見プロセスとして仮説駆動型とデータ駆動型と呼ばれるものがある．本システムはこの２つの知識発見プロセスの過程で使用され，生物学者に対しての知識発見支援を行うことを目的とする．

(1) 仮説駆動型では，仮説に基づき実験計画を立て，実験を行い，得られた観測に対して，仮説検証を行うプロセスを繰り返す．本システムはこの実験計画と仮説の立案をサポートすることを目指し，①異なる実験データをもとにリアルタイム推定を行う②ユーザ指定の条件下で再推定する機能を有する．

(2) データ駆動型では，ハイスループット実験により取得した大量データをもとに知識発見を行う．大量データから知識発見の対象を網羅的に探索するのは非常に時間がかかる．また，前提知識を持たない場合には，探索自体が難しい．本システムは，大量のマイクロアレイデータからどの転写パスウェイが活性しているのか，転写機構の状態を俯瞰してユーザに提示する機能を有する．これによってユーザの直感を刺激し，これまでにない新たな気づきを促すことをねらう．

8.2.2 機能要件

本システムには次の機能を実装する．

・基本機能：拡大，縮小，検索，選択

・可視化機能

- 推定結果と推定精度に応じて，頂点の色，大きさを変更する
- fc やパス長といったネットワーク情報の確認できる
- ユーザ指定の条件によって表示範囲を変更する

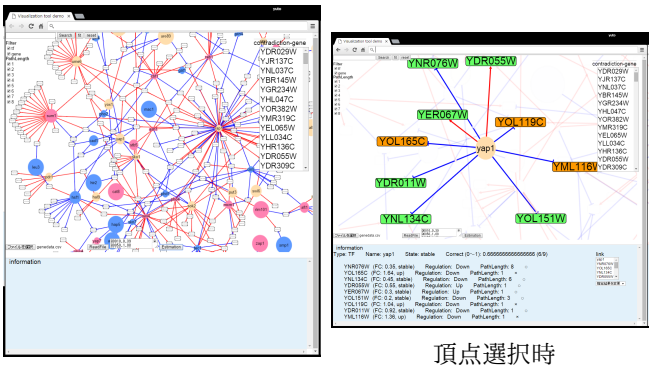
・インタラクション機能

- 新たなマイクロアレイ実験データに対し，リアルタイムで推定し，その結果を可視化する
- パラメータ変更を反映した推定結果をリアルタイムで可視化する

図 3 には作成したツールの全体図と頂点選択時の様子を示す．画面上部にはネットワークの視覚的情報が描画され，画面下部には詳細なネットワーク情報が記述される．インタラクション機能によりリアルタイム推定やパラメータ変更によって，グラフの描画やネットワークの情報が更新される．

9. まとめと今後の課題

本研究では，専門家への知識発見支援として転写因子結合ネットワーク上の転写因子の状態推定問題について最適化問題へ定式化し，単独推定と複数推定の推定を行った．複数推定は，最適化問題を制約最適化問題に変換し，SAT ソルバを用いて高速に求解した．このとき，最適解が全て *stable* になってしまい，これを解決するために，観測データを考慮した新たな評価関数を設計した．評価関数の性能を比較し，生物学的検証を行った．また，知識発見のための対話可能な可視化ツールを作成した．



全体図　　頂点選択時

図 3　可視化ツール

今後の課題として，現行のモデルより生物学的な評価の高い推定モデルを提案する．また，可視化ツールを提供することで知識発見支援プロセスの確認を行う．

謝辞

本研究は一部，文科省科学研究費補助金（若手 B: No.22700141）文科省科学研究費補助金（基盤 C: No.22500127）および JST，さきがけの援助を受けている．

参考文献

[1] Kitano，H.: All systems go，*Nature Reviews Drug Discovery*，Vol. 7，pp. 278-279 (2008)

[2] Ochs，H.: Knowledge-based data analysis comes of age, *Brief Bioinform.*, Vol. 11, No. 1, pp. 30-39 (2010)

[3] Yang, T. H. and Wu, W. S.: Identifying biologically interpretable transcription factor knockout targets by jointly analyzing the transcription factor knockout microarray and the ChIP-chip data, *BMC Systems Biology*, Vol. 6, pp. 102-112 (2012)

[4] 平沼祐人，山本 泰生，岩沼 宏治: 遺伝子発現データを用いた転写因子束縛ネットワークの状態推定，人工知能学会全国大会 (第 28 回)，4K1-4in，(2014)

[5] Takehide Soh, Naoyuki Tamura, and Mutsunori Banbara : Scarab: A Rapid Prototyping Tool for SAT-based Constraint Programming Systems (Tool Paper) , In the Proceedings of the 16th International Conference on Theory and Applications of Satisfiability Testing (SAT 2013), LNCS 7962, pp. 429-436, 2013

[6] Le Berre, D. and Parrain, A.: The Sat4j library, release 2.2, *Journal on Satisfiability, Boolean Modeling and Computation*, Vol. 7, No. 2-3(2010), pp. 59-6.

[7] Reimand J, Vaquerizas JM, Todd AE, Vilo J and Luscombe NM: Comprehensive reanalysis of transcription factor knockout expression data in Saccharomyces cerevisiae reveals many new targets, *Nucleic Acids Res*, Vol. 38, Issue 14, pp. 4768-4777 (2010).

[8] Cytoscape: An Open Source Platform for Complex Network Analysis and Visualization. http://www.cytoscape.org/, (accessed 2016-02-01).

[9] Cytoscape.js, http://js.cytoscape.org/, (accessed 2016-02-01).