4K1-4in

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

A Network-based Analysis of Transcription Factor Activities in S. Cerevisiae

平沼 祐人 *1　山本 泰生 *2　岩沼 宏治 *2
Yuto Hiranuma　Yoshitaka Yamamoto　Koji Iwanuma

*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
Department of Computer Sience and Media Engineering,
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

Recently, molecular interactions between transcription factors and their binding genes have been intensively studied. These interactions are represented as a network, called a Transcription Factor Binding network (TFBN). In this paper, we propose a computational method to estimate the activity of each transcription factor in TFBN using gene expression data. The proposed method solves the problem by formalizing it as an optimization problem. Consequently, we compute transcription factor activities in S. Cerevisiae, and discuss about obtained results.

1. はじめに

分子生物学の発展に伴い, 細胞内の分子間相互作用に関するさまざまな理解が深まっている. 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][2] の分野が注目されているが, システム生物学的アプローチで得られた知識ベースを元に生命機構全体の包括的な解析が可能になってきている [3][4][5].

例えば, マイクロアレイや ChIP-chip 技術により, **転写因子**と**遺伝子発現**の関係について詳細が明らかになっており, 転写因子制御の機構全体がネットワークモデルとして形式化されるようになっている. 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, 束縛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え, 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, 出芽酵母の転写因子制御モデルである**転写因子束縛ネットワーク** [6] を用いて, 実験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, ゲノム創薬等の様々な応用が考えられる. その一方で, 転写因子束縛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解析することは現実的に困難である.

本稿では, はじめに転写因子の活性状態を評価するための 2 つの評価関数を定義する. 最初の評価関数は, 単独の転写因子のみの影響を考慮したものであり, 2 つ目の評価関数は複数の転写因子の影響を考慮したものである. これらの評価関数を用いることで, 転写因子の活性状態を推定する問題を最適化問題へと形式化することが出来る. 最適化問題では, 評価関数の値が最も高くなるような転写因子の活性状態の割り当てを求める. 単独の影響のみを考慮した評価関数は線形分離可能なので, 各転写因子において最も評価が高くなるような状態割り当てを求めるだけで十分である. 一方で, 複数の影響を考慮した問題では, 転写因子活性状態の割り当てのすべての組み合わせを考える必要があるので, 単純な力任せ法では手に負えない. そこで, 本論文では分枝限定法を用いた解法を提案する.

連絡先: 平沼 祐人, 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, 住所: 〒 400-8511 山梨県甲府市武田 4-3-11, E-mail: yyamamoto@yamanashi.ac.jp

2. 準備

本稿で用いる転写因子束縛ネットワークと遺伝子発現の実験データについて説明し, 転写因子の活性状態を定義する.

2.1 転写因子束縛ネットワーク

転写因子束縛ネットワーク (*Transcription Factor Binding Network*: 以下, TFBN と略す) とは, 転写因子とそれが束縛する遺伝子の関係を示すグラフである. TFBN の頂点は, 転写因子または遺伝子であり, 頂点間は有向辺で結ばれている. 転写因子頂点と遺伝子頂点間にパスが存在する場合, その転写因子が遺伝子を制御している.

本研究で用いる TFBN は, 2012 年に Yang らがまとめたものであり [6], 出芽酵母における 112 個の転写因子, 5105 個の遺伝子の関係が次のようなデータ構造でまとめられている.

- TFBN 上の各転写因子 n_i
- n_i からのパスを持つ各遺伝子 n_j
- n_i が n_j へ及ぼす影響を示す $label \in \{activate, inhibit\}$
- n_i から n_j への最短パス上の中間頂点の列

この 4 つの要素から成る情報を TFBN の**最短パス情報**と呼ぶ. 図 1 に TFBN の例を示す. 各頂点は転写因子頂点または遺伝子頂点であり, 頂点間の辺は, その *label* に応じて区別され, → は活性化, ⊣ は阻害の制御をそれぞれ意味する.

2.2 遺伝子発現の実験データ

遺伝子発現の実験データとして, TFBN 上の遺伝子における野生株と変異株での発現量の比 **fold-change** (以下, fc と略す) を利用する. 野生株は自然集団中で最も高頻度に見られる遺伝子型の株であり, 変異株はある特定の遺伝子をノックアウトして得られる株である. ノックアウトとは, 特定の物質の機能を意図的に破壊することでその物質が他に及ぼしていた影響などを調べる生物学的な操作を指す. 遺伝子発現の実験データには, TFBN 上の遺伝子 681 個の fc が記されている. fc はマイクロアレイの実測値から直ちに求まるが, これを元にあるし

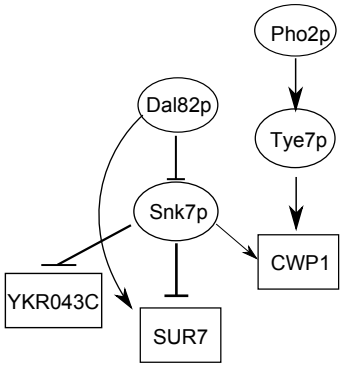

図 1: TFBN の例

きい値 p 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. 本稿では, $p = 2.0$ と設定した.

$$fc = \begin{cases} up & (fc \geq p) \\ down & (fc \leq \frac{1}{p}) \\ stable & (\text{上記以外}) \end{cases}$$

$fc = up$ の場合, 野生株に対して変異株の発現量が有意に増加していることを意味する. 同様に, $fc = down$ の場合は有意に減少, $fc = stable$ 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.

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, その遺伝子を束縛する転写因子の状態を推定することにある.

2.3 転写因子の状態

TFBN 上の転写因子の状態は次のように定義される.

定義 1 TFBN 上の転写因子頂点 n_i の状態を $x_i \in \{up, down, stable\}$ とする. TFBN 上の各転写因子 $n_1, n_2, \ldots, n_k$ の状態割り当てを $A = (x_1, x_2, \ldots, x_k)$ と書く.

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予測できる. その様子を表 1 に示す.

図 1 の転写因子 Snk7p と遺伝子 CWP1, SUR7 を例に説明する. Snk7p は, CWP1 に対して活性化の制御, SUR7 に対して阻害の制御を行っている. Snk7p の状態を $x_{Snk7p} = up$ としたとき, 活性化の制御を受ける CWP1 は $fc = up$ (有意に上昇), 阻害の制御を受ける SUR7 は $fc = down$ (優位に減少) と推定できる. 同様に $x_{Snk7p} = down$ のとき, CWP1 は $fc = down$, SUR7 は $fc = up$ と推定でき, $x_{Snk7p} = down$ のとき, CWP1 は $fc = stable$, SUR7 は $fc = stable$ と推定できる.

表 1: 転写因子状態 x_i としたときの遺伝子変化状態 fc の関係

x_i \ label	$activate$	$inhibit$
up	$fc = up$	$fc = down$
$down$	$fc = down$	$fc = up$
$stable$	$fc = stable$	$fc = stable$

3. 問題設定

3.1 推定問題の形式化

2.3 節で述べたように, TFBN 上で転写因子の状態を与えると制御している遺伝子の発現量を決定することができる. 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 2 A は転写因子の状態割り当て, P は最短パス情報, EX は遺伝子発現の実験データを 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.

$Precision(A, P, EX)$ は最適化問題の評価関数に相当し, こ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評価関数について述べる.

3.2.1 単独の転写因子の影響のみを考慮した評価関数

単独の転写因子のみの影響を考慮するときの各転写因子の推定精度を次のように定義する.

定義 3 $Precision(x_i, P, EX) = \frac{|\ G_p\ |}{|\ EX_p\ |}$

ただし, G_p は発現変化量を x_i と割り当てたことで, その発現変化を正しく予測できた遺伝子の集合, EX_p は転写因子 n_i によって束縛され, かつ観測データを持つ遺伝子の集合を指す.

図 2 では遺伝子 $Gene1$, $Gene3$ に活性化, 遺伝子 $Gene2$ に阻害の制御を行っている転写因子 n_k について, $x_k = up$ としたときの遺伝子発現変化を示す. 遺伝子発現の実験データ $Gene1$, $Gene1$, $Gene3$ の変化状態がそれぞれ, $fc_1 = up$, $fc_2 = up$, $fc_3 = up$ だったとする. このとき, $Gene1$, $Gene3$ の 2 つの遺伝子に関しては正しく予測でき, $Gene2$ に関しては正しく予測できていない. 転写因子 n_i の推定の評価値は, $Precision(x_i, P, EX) = \frac{2}{3}$ となる.

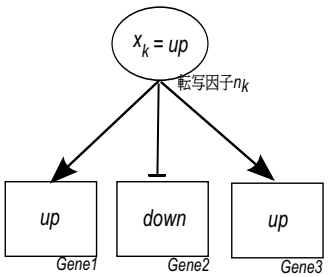

図 2: 転写因子状態 $x_k = up$ の遺伝子発現変化

各転写因子の推定精度を定義 3 のように定義したとき, 単独の転写因子の影響のみを考慮した評価関数 $Precision(A, P, EX)$ は次のように定義される.

定義 4 $Precision(A, P, EX) = \frac{1}{n}\sum_{i=1}^{n} Precision(x_i, P, EX)$

ただし, n は推定の対象となる転写因子の数である.

このとき, $Precision(A, P, EX)$ は線形分離可能であり, 各転写因子ごとの評価関数が最も高くなるような状態の割り当て A を選択すれば, 推定精度は最大となる.

3.2.2 複数の転写因子の影響を考慮した評価関数

複数の転写因子での推定精度を次のように定義する.

定義 5 $Precision(A, P, EX) = \frac{|\ G\ |}{|\ EX\ |}$

ただし, 転写因子の状態割り当て A に対して, G は正しく発現変化を予測できた遺伝子の集合, EX は観測データを持つ全遺伝子の集合である.

図 3 は, 2 つ転写因子頂点 n_1, n_2 と 3 つの遺伝子頂点 $Gene1$, $Gene2$, $Gene3$ で構成される TFBN であり, 転写因子にそれぞれ状態を与えたときの遺伝子発現変化を示す. 遺伝子発現の実験データは, $Gene1 = up$, $Gene2 = up$, $Gene3 = up$ だったとする. このとき, $Gene1$, $Gene3$ の 2 つの遺伝子に関しては正しく推定ができ, $Gene2$ に関しては正しく推定ができていない. 従って複数の転写因子の影響を考慮した推定における推定の評価値は $Precision(A, P, EX) = \frac{2}{3}$ となる.

複数の転写因子の影響を考慮した推定では, 転写因子の活性状態の割り当ての組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} 通りの組合せを調べる必要がある. 従って, すべての組合せを総当たりする力任せ法では組み合わせ爆発により対応できない. そこで, 分枝限定法を用いた探索アルゴリズムにを新たに提案する.

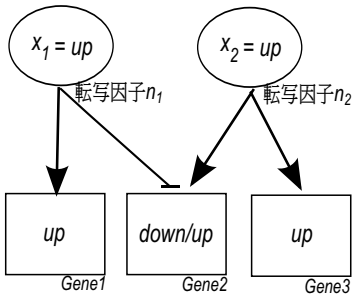

図 3: 転写因子状態 $x_1 = up$, $x_2 = up$ の遺伝子発現変化予測

4. 提案手法

本稿で提案する分枝限定法を用いた探索アルゴリズムと探索順序について説明を行う.

4.1 分枝限定法を用いた探索アルゴリズム

提案手法は, 予測に失敗した遺伝子数が最小となる状態割り当て A を求める. この問題における探索空間は, 各転写因子に割り当てられる 3 つ状態を子頂点に持つ完全三分木であり, この中を深さ優先で探索する. この探索においては, 正しく予測できなかった遺伝子数が単調に増加するので限定操作による枝刈りが可能である. i 番目までの転写因子の状態が定まったとき, 予測に失敗した遺伝子数を i での**暫定解**, そのときの遺伝子数を**暫定値**と呼ぶ. まず, 定義 3 において評価値が最大となる各転写因子の状態を選択する. このとき, 最初に得られる暫定解は, 単独の転写因子の影響のみを考慮した推定における最適な状態割り当てと等しい.

以下, 分枝限定法におけるアルゴリズムを示す.

入力: TFBN, 遺伝子発現の実験データ
出力: TFBN の最適な転写因子の状態割り当て

1. 探索木上の深さ i までの部分状態割り当て A_i と A_i に対する評価値の上界を求める
   暫定値および暫定解を求める
2. バックトラック法により, すべての部分問題が終端するまで以下の 3-5 の処理を繰り返す
3. 探索木の葉に位置するとき
   (a) 暫定値 $>$ A_i の評価値 のとき
   暫定値および暫定解を更新
   現ノードを探索済みとし, 親ノードへ移動
   (b) 暫定値 $<$ A_i の評価値のとき
   現ノードを探索済みとし, 親ノードへ移動
4. 現ノードのすべての子が探索済のとき
   現ノードを探索済みとし, 親ノードへ移動
5. 現ノードの子に未探索のノードがあるとき
   その未探索の子を選択
   (a) 暫定値 $>$ A_i の評価値の上界のとき
   最適値を更新し, 探索を続ける (分枝操作)
   (b) 暫定値 $<$ A_i の評価値の上界のとき
   現ノードを探索済みとし, 親ノードへ移動 (限定操作)

提案手法のアルゴリズム

提案アルゴリズムでは, 深さ優先探索を用いている. 効率良く限定操作による枝刈りを発生させるためには, 探索の順序を工夫する必要がある.

4.2 探索順序

はじめにある特定の状態に決定できるような転写因子がある場合は優先的に探索する. その他の転写因子については, 束縛している遺伝子の数が多い順に探索を行う. これは, 最適解でない状態を選択した際, 評価値の上界が探索木上の浅い段階で暫定値を超えやすくするためである. 実際に探索した転写因子の順序を表 2 に示す. 表中の up, $down$, $stable$ は転写因子の状態を表し, それぞれの状態を与えたときに推定が一致する遺伝子の数である. 表 2 は, TFBN の一部である.

表 2: 転写因子の探索順序

探索順	転写因子名	制御している総遺伝子数	up	$down$	$stable$
1	met28	20	0	11	9
2	gln3	88	31	23	34
3	hms1	77	18	36	23
4	ace2	67	25	23	19
⋮	⋮	⋮	⋮	⋮	⋮

5. 実験結果

単独の転写因子の影響のみを考慮した推定の実験結果と複数の転写因子の影響を考慮した推定の実験結果について示す.

5.1 単独の転写因子の影響のみを考慮した推定実験

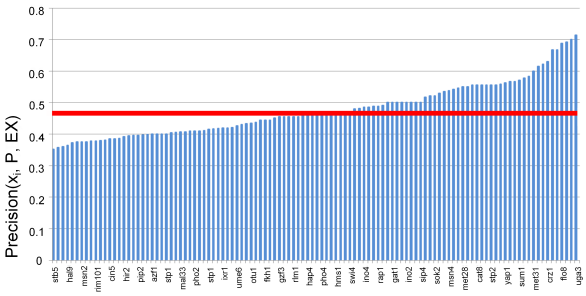

図 4: 単独の転写因子の影響のみを考慮した推定実験

図 4 は, 横軸が各転写因子, 縦軸は各転写因子での推定精度を表す度数分布である. 結果は $Precision(A, P, EX) = 0.48$ となった. この結果の詳細な分析は今後, 生物学の専門家と行う.

5.2 複数の転写因子の影響を考慮した推定の実験結果

任意の転写因子をランダムに抽出して実験を行った. また, 転写因子は制御している遺伝子の数が降順になるようにしている. 実験結果を表 3 に示す.

表 3 には, 転写因子の数が 30, 40, 50, 60 の実験結果が示されている. 探索空間は探索木の頂点数である. 各頂点につき 3 つの状態が存在するため, 限定操作を行わない力任せ法では, $O(3^n)$ である. ただし n は転写因子数を指す. 限定操作による枝刈りで探索空間の削減を行っている. 分枝限定法による探索空間は, 限定操作によって削減された探索空間である. この 2 つに注目すると, 力任せ法の探索空間に比べて, 分枝限定法を用いた探索空間は大幅に削減されていることが分かる.

表 3: 分枝限定法を用いた解法の実験結果

転写因子数	40	50	60
力任せ法による探索空間	1.8×10^{19}	1.1×10^{24}	6.3×10^{28}
分枝限定方による探索空間	8.2×10^{5}	$5.4 \times 10\hat{}6$	2.8×10^{8}
$Precision(A, P, EX)$	0.58	0.58	0.59
計算時間 [s]	208.9	784.5	27993.7

6. 考察及び評価実験

単独の転写因子の影響のみを考慮した推定は, 転写因子 112 個の状態推定を高速に行うことができる. 一方, 複数の転写因子の影響を考慮した推定では, 転写因子 60 個の状態推定にかなりの時間を要していることから, 112 個すべての状態推定を完了するのは困難であることが分かる. しかし, この推定はモデルの意味論により即した評価方法であり, 生物学的見地から, 得られる転写因子状態はより精度の高いものとなっている. 実際, 転写因子 60 個における正しく予測できなかった遺伝子数が, 単独の転写因子の影響のみを考慮した推定方法では 466 個であったのに対し, 複数の転写因子の影響を考慮した推定方法では 361 個と 23% 程度, 予測に失敗する遺伝子数が減少している. このことからも後者の推定の方がより精度の高い転写因子状態を得られていることが分かる.

生命機構モデルに基づきより正確な推定を行うには, 転写因子から遺伝子への制御に加えて別の新たな要素を考慮する必要がある. 例えば, パス長を考慮することが挙げられる. パスの長さは, 転写因子が遺伝子へ及ぼす影響の強さに関わるため, 遺伝子に対してパス長に応じた重み付けを行うができる.

また, 分枝限定法を用いた手法において限定操作による枝刈りが有効にはたらかない理由として, 探索木の浅い位置での限定操作による枝刈り起きていないことを確認した. このため, 探索空間の大幅な削減には至らず計算に非常に時間がかかる. 転写因子の各状態における推定が成功する遺伝子数のばらつきが小さいことが原因の一つだと考えられる. このばらつきを大きくする方法としては, 実験データの遺伝子発現量の比 fc の大きさに基づいたペナルティを設けることが考えられる.

現在, このようなパス長と状態における遺伝子数のばらつきを考慮した予備実験を行ってはいるが, 今のところ十分な高速化には至っていない. この理由としては, TFBN のグラフ構造が非常に密だということが考えられる. どの 2 つの転写因子についても, それらが共通して制御する遺伝子を持つ完全グラフに近いグラフ構造となっている事が分かっており, すなわち非常に転写因子間の依存性が高いグラフであり, これが限定操作による枝刈りを抑制する一因になっていると推測できる. 今後の課題として, これらの知見を踏まえてアルゴリズムや探索順序をさらに工夫する必要がある.

7. おわりに

今回, 遺伝子発現を用いたの転写因子束縛ネットワークの状態推定の手法を提案・実装し, 実験を行った. 転写因子単独の影響のみを考慮した推定では, 高速に求解できた. 複数の影響を考慮した推定では, 60 個程度なら実時間で求解できる. しかし, 転写因子の数がそれ以上になると実時間での求解は困難であるため, 今後更なるアルゴリズムの効率化を進める必要がある.

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. また, 本研究に用いた遺伝子発現の実験データは岡山大学守屋央朗博士より提供して頂いた.

参考文献

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

[2] 江口至洋: 細胞のシステム生物学, pp. 1-245, 共立出版 (2008)

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

[4] 坂本 悠, 山本 泰生, 岩沼 浩二: 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), 3 1 2-2, (2008)

[5] 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, pp. 1-807, 共立出版 (2006)

[6] 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)