統計数理(2006)
第 54 巻 第 2 号 333–356

特集「予測と発見」
[研究詳解]

異種ゲノムデータの統合による遺伝子ネットワーク推定手法

玉田 嘉紀[1] ・井元 清哉[2] ・宮野 悟[2]

(受付 2006 年 4 月 10 日; 改訂 2006 年 6 月 28 日)

要　　旨

細胞内で生成されるタンパクは生物の主要な構成要素であり，遺伝子はその設計図に相当する．遺伝子がタンパクに変換される時期・量の制御も遺伝子の働きによるものであり，生物は遺伝子同士が協調して作用することによって生命を維持している．このような遺伝子間の依存関係を，頂点と枝から構成されるグラフを用いて表現したものを遺伝子ネットワークという．近年のマイクロアレイ技術の発展により，細胞内の遺伝子の活動状態を網羅的に観測できるようになり，遺伝子発現データとして蓄積されている．遺伝子発現データに基づく遺伝子ネットワークの推定問題は，バイオインフォマティクスにおいて最も重要な課題の 1 つと考えられる．遺伝子ネットワークの推定問題は，遺伝子の発現量を確率変数として見なすことにより，グラフィカルモデルの推定問題として定式化される．しかし，ネットワークに含まれる遺伝子は一般に数百以上と多く，そのためモデルに含まれるパラメータの数は膨大となる．したがって，発現データへの過適合を避けるためのモデリングの方法論を構築することが必要不可欠といえる．本稿ではこのような問題を解決するための方法として，著者らが開発した 2 つ の異なるアプローチによる遺伝子ネットワーク推定手法について解説する．1 つは同一の遺伝子から直接の制御を受ける遺伝子の DNA 配列上流領域に共通の制御配列が存在することに着目し，共通配列探索と発現データを組み合わせた方法．他方は 2 種類の異なる生物種の遺伝子ネットワークを，両種に進化的に保存されている情報を互いに利用しながら同時に推定する手法である．両手法は，ベイジアンネットワークを遺伝子ネットワークのモデルとして用い，ネットワークをグラフ構造の事後確率最大化に基づいて推定する．その際，配列情報および進化情報をネットワークの事前確率を構成するために用いることが特徴となっている．開発した手法はシミュレーションおよび実データへの適用を通してその有効性を確認した．

キーワード：遺伝子ネットワーク，遺伝子発現データ解析，制御配列，進化情報，ベイジアンネットワーク．

1. はじめに

遺伝子は細胞内で生成されるタンパクの設計図に相当するもので，細胞の核内に存在する DNA 配列として記録されている．遺伝子に対応する DNA 配列は，転写と呼ばれる作用によっ

[1] 統計数理研究所（現 株式会社ジーエヌアイ GNI 創薬解析センター：〒814–0001 福岡市早良区百道浜 3–8–33–608）
[2] 東京大学 医科学研究所ヒトゲノム解析センター：〒108–8639 東京都港区白金台 4–6–1

て mRNA 配列に写し取られたあと，翻訳と呼ばれる作用により実際のタンパクに変換される．この様に，遺伝子が読み取られタンパクに変換されることを「遺伝子が発現する」と表現する．遺伝子の発現を制御している物質は転写因子と呼ばれるが，転写因子もまた遺伝子から合成されたタンパクの一種である．そのため，ある転写因子が発現すると，その被制御遺伝子の発現制御を促すことにつながる．このように遺伝子の発現は互いに依存しており，協調し複雑に影響しあうことによって生命活動の維持に必要なタンパクを適切に制御している．遺伝子ネットワーク(gene network)は，このような遺伝子間の発現制御の依存関係を，グラフを用いて表したものである．遺伝子ネットワークでは，各遺伝子はグラフ中の頂点として表され，遺伝子間の発現の依存関係は頂点を結ぶ有向枝で表される．

近年，DNA マイクロアレイなどのツールの登場によって，様々な実験的状況下において細胞内の遺伝子の発現状態を網羅的に観測することができるようになった(DeRisi et al., 1997)．それらの遺伝子発現データを用いて，遺伝子ネットワークを推定する研究が盛んにおこなわれている(van Someren, 2002)．遺伝子ネットワーク推定の問題は，遺伝子の発現量を確率変数としてとらえたグラフィカルモデルの推定問題として定式化される．従来のグラフィカルモデル推定問題と比較して，きわめて多くの頂点(遺伝子)からなる構造未知のネットワークを数限られたサンプル(マイクロアレイデータ)から推定することがその特徴であると言える．また，ネットワークの制御や挙動予測よりもネットワークの構造推定に重点が置かれるのもその特徴である．実際の生物における具体的な遺伝子数としては，出芽酵母だと約 6000，ヒトに至っては 20000 以上となる．ネットワーク全体ではなく，興味のある部分に特化した場合においても，1000 程度の遺伝子を含むネットワーク推定をする必要が往々にしてある．一方，同一生物種，同一細胞組織(肝細胞，血管内皮細胞など)のマイクロアレイデータは，コストなどの面からその数は数百のオーダーでしか集まらないのが現状である．つまり，モデルに含まれるパラメータ数がサンプル数よりも多くなる状況が頻繁に生じ，そのためデータへ過適合し，高精度な遺伝子ネットワークを推定することは非常に困難な問題となっている．

著者らは，このような問題に対し 2 つの異なるアプローチによる遺伝子ネットワーク推定手法を開発した．1 つは遺伝子の DNA 配列上流領域に存在する制御配列の情報(Tamada et al., 2003)，他方は異なる生物種間に進化的に保存されている情報(Tamada et al., 2005)をそれぞれ用いた遺伝子ネットワーク推定手法である．前者は，実際の遺伝子発現の仕組みに着目した方法であり，DNA 配列中に存在する制御配列の情報を発現データと組み合わせたものである．同一の転写因子から直接の制御を受ける遺伝子の DNA 配列上には，その転写因子が結合する共通の短い配列が存在する．このような共通の配列を共通配列探索手法を用いて探索し，その結果を発現データと組み合わせることが基本的なアイディアとなっている．後者は，進化的に距離が離れた生物種間であっても，細胞の生命を維持するための基本的な仕組みはよく保存されていることを利用する．また，これら 2 つの手法に共通した特徴としては，ベイジアンネットワークを遺伝子ネットワークのモデルとして用い，遺伝子ネットワークをグラフ構造の事後確率最大化に基づき推定する．その際，配列情報および進化情報をネットワークの事前確率を構成するために用いる．開発した手法は，それぞれ人工データによるシミュレーションおよび実データへの適用を通してその有効性を確認した．

本論文の構成は次の通りである．2 章では，ベイジアンネットワークによる遺伝子ネットワーク推定法を説明し，関連研究としてこれまでに行われてきた遺伝子ネットワーク推定研究の概要を説明する．3 章では制御配列の情報を利用した手法，4 章では進化情報を利用した手法を解説する．

本論文で使用したシミュレーションや実データへの適用の際のデータは，公開可能なものを著者らのウェブページ(http://bonsai.ims.u-tokyo.ac.jp/~tamada/supplement.html)に掲載して

いる．

2. ベイジアンネットワークによる遺伝子ネットワーク推定

2.1 ベイジアンネットワーク

本節では，Imoto et al. (2002)によって提案された B-スプラインを用いたノンパラメトリック回帰に基づくベイジアンネットワークについて説明する．

ベイジアンネットワークは，確率変数間の条件付き独立性を非循環有向グラフを用いて表したものである．グラフの各頂点に確率変数が 1 対 1 で対応付けられ，頂点間に有向枝に沿ったマルコフ連鎖律を仮定することで，同時確率が各確率変数のグラフ上の親変数を所与としたときの条件付き確率の積で表せることが本質的である．今，$X_1,\ldots,X_p$ を p 個の確率変数とし，各変数間の依存関係を表すグラフ構造を G で表すとする．このとき，$X_1,\ldots,X_p$ の同時確率は次のように表せる．

$$\Pr(X_1,\ldots,X_p)=\prod_{j=1}^{p}\Pr(X_j|Pa(X_j)). \tag{2.1}$$

ただし，$Pa(X_j)$ は変数 X_j のネットワークにおける親頂点に対応する変数の集合である．また，親が存在しない頂点の場合は $Pa(X_j)=\emptyset$ とする．図 1 にベイジアンネットワークの具体的例を示す．遺伝子ネットワークでは，各遺伝子の発現量を確率変数としてとらえ，グラフ中の頂点を各変数(遺伝子)に対応させる．遺伝子間の発現の依存関係は各変数間の依存関係として考え，頂点間の有向枝で表す．以降では，変数 X_j とそれに対応するネットワーク中の頂点および遺伝子を明確に区別せず，誤解のない限り単に X_j と表す．

今，p 個の遺伝子 $X_1,\ldots,X_p$ の発現量を N 枚のマイクロアレイにより観測したとする．それらのデータは $N\times p$ 行列 $\boldsymbol{X}$ としてまとめられる．すなわち $\boldsymbol{X}$ の (i,j) 成分 x_{ij} は i 番目のマイクロアレイにおいて観測された遺伝子 X_j の発現量を表す．また，遺伝子ネットワークの構造を $G=(V,E)$ で表すことにする．ただし，$V=\{X_1,\ldots,X_p\}$ は頂点集合，$E\subseteq V\times V$ は有向枝集合である．発現データは連続値であるので，式(2.1)における分解は，同時密度関数の分解として表される．

$$f(\boldsymbol{X}|G,\boldsymbol{\Theta})=\prod_{i}^{N}\prod_{j}^{p}f_j(x_{ij}|\boldsymbol{pa}(x_{ij}),\boldsymbol{\theta}_j).$$

ただし，$\boldsymbol{pa}(x_{ij})$ は i 番目のマイクロアレイにおける遺伝子 X_j の親遺伝子からなる発現データベクトルである．また $\boldsymbol{\Theta}=(\boldsymbol{\theta}_1',\ldots,\boldsymbol{\theta}_p')'$ はパラメータベクトルである．f_j $(1\le j\le p)$ は遺伝子 X_j とその親遺伝子との間の発現の依存関係を表す条件付き密度関数であり，B スプラインを用いたノンパラメトリック加法回帰モデルにより構築される(Imoto et al., 2002)．すなわち，

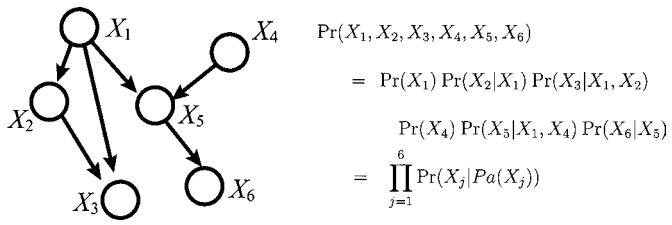


図 1. ベイジアンネットワークの例.

発現データ x_{ij} と $\boldsymbol{pa}(x_{ij})=(p_{i1}^{(j)},\ldots,p_{iq_j}^{(j)})'$ の関係は次のように表される．

$$x_{ij}=m_{j1}(p_{i1}^{(j)})+\cdots+m_{jq_j}(p_{iq_j}^{(j)})+\varepsilon_{ij}.$$

ただし，$m_{jk}(\cdot)$ は滑らかな関数であり，ここでは B スプラインを用いて

$$m_{jk}(p_{ik}^{(j)})=\sum_{m=1}^{M_{jk}}\gamma_{mk}^{(j)}b_{mk}^{(j)}(p_{ik}^{(j)})$$

と表せる．ここで $\{b_{1k}^{(j)}(\cdot),\ldots,b_{M_{jk},k}^{(j)}(\cdot)\}$ はあらかじめ与えられた M_{jk} 個の B スプライン基底関数，また $\{\gamma_{1k}^{(j)},\ldots,\gamma_{M_{jk},k}^{(j)}\}\in\mathbf{R}^{M_{jk}}$ は B スプライン基底関数に対する係数ベクトルである．ε_{ij} はノイズ項で，ここでは平均 0，分散 σ_j^2 の正規分布に従うと仮定する．以上より x_{ij} の条件付き密度関数は次式で与えられる．

$$f_j(x_{ij}|\boldsymbol{pa}(x_{ij}),\boldsymbol{\theta}_j)=\frac{1}{\sqrt{2\pi\sigma_j^2}}\exp\left[-\frac{\{x_{ij}-\sum_{k=1}^{q_j}m_{jk}(p_{ik}^{(j)})\}^2}{2\sigma_j^2}\right].$$

発現データからの遺伝子ネットワークの推定問題は，遺伝子ネットワークのグラフ構造 G の決定と，それに基づくノンパラメトリック回帰の問題に帰着される．グラフ構造 G の決定はモデル選択と見なすことができ，ここではグラフ構造 G の事後確率の最大化に基づいて行う．発現データ $\boldsymbol{X}$ が与えられた下での G の事後確率は次式で表される．

$$\begin{aligned}\pi(G|\boldsymbol{X})&\propto\pi(G)\cdot\pi(\boldsymbol{X}|G)\\&=\pi(G)\int f(\boldsymbol{X}|G,\boldsymbol{\Theta})\pi(\boldsymbol{\Theta}|\boldsymbol{\lambda})d\boldsymbol{\Theta}.\end{aligned}\tag{2.2}$$

ただし $\pi(\boldsymbol{\Theta}|\lambda)$ はハイパーパラメータ λ によって規定されるパラメータ $\boldsymbol{\Theta}$ の事前分布である．従って，遺伝子ネットワークの構造は，MAP 解 $\hat{G}=\mathrm{argmax}_G\pi(G|\boldsymbol{X})$ を求めることで得られる．しかしながら，一般に非循環有向グラフ(Directed Acyclic Graph: DAG)の探索は非常に計算量の多い問題である．具体的には，頂点数が 20 の DAG はおよそ 2.34×10^{72} 通り存在する．そのため発見的な探索法である greedy hill-climbing アルゴリズム(Heckerman et al., 1995)がよく使われる(Friedman et al., 1998; Imoto et al., 2002)．Ott et al. (2004)らは動的計画法に基づく探索アルゴリズムを構築し，30 遺伝子程度の比較的小規模のネットワークであれば最適解を現実的な時間で探索できるアルゴリズムを開発している．しかしながら，本論文ではより多くの遺伝子を含むネットワーク構築を考えるため greedy hill-climbing アルゴリズムを用いてベイジアンネットワークの構造学習を行う．次に greedy hill-climbing アルゴリズムについて説明する．Greedy hill-climbing アルゴリズムはグラフ構造に対して定義されるスコアの最大化に基づいてグラフの局所最適解を探索する．ここでは事後確率をスコアと考え $S(E)$ で表すことにする．

Input: 発現データ $\boldsymbol{X}$，頂点(変数)集合 $V=\{X_1,\ldots,X_p\}$，親候補数 m， 反復回数 T

Output: 非循環有向グラフ $G=(V,E)$

Step 1. $E:=\emptyset$ (G は空グラフ)，$t:=1$.

Step 2. すべての i, j $(1\le i,j\le p,\ i\ne j)$に対し $s_{ij}:=S(\{(X_i,X_j)\})$ を計算.

Step 3. $C_j:=\{s_{\alpha_1^{(j)},\,j},\ldots,s_{\alpha_m^{(j)},\,j}\}$, ただし $\{\alpha_1^{(j)},\ldots,\alpha_m^{(j)}\}$ は $s_{1j},\ldots,s_{pj}$ の上位 m 個の添字集合.

Step 4. $1,\ldots,p$ のランダムな順列 $\pi_1,\ldots,\pi_p$ を発生させる.

Step 5. すべての X_{π_j} $(1\le j\le q)$ に対して X_{π_j} の親候補 $Y\in C_{\pi_j}\cup\{X:(X,X_{\pi_j})\in E\}$

を考え，以下の操作で得られるスコア $S(E')$ を求める．

（a） $E' := E \cup \{(Y, X_{\pi_j})\}$ if $(Y, X_{\pi_j}) \notin E$

（b） $E' := E \setminus \{(Y, X_{\pi_j})\} \cup \{(X_{\pi_j}, Y)\}$ if $(Y, X_{\pi_j}) \in E$

（c） $E' := E \setminus \{(Y, X_{\pi_j})\}$ if $(Y, X_{\pi_j}) \in E$

(a), (b), (c)のうち最もスコアが向上する操作を採用し，$E := E'$ とする．ただし，これらの操作は E' に閉路ができない場合に限る．

Step 6. スコアが向上し続ける限り Step $4 \sim 5$ を繰り返し行う．スコアが向上しなくなった場合，その時点のグラフを G_t とする．$t := t + 1$.

Step 7. $t \leq T$ ならば $E = \emptyset$ とし，Step $4 \sim 6$ を繰り返す．

Step 8. $G_1, \ldots, G_T$ のうち最もスコアの良いネットワークを出力する．

実際にここで示した greedy hill-climbing アルゴリズムを実行するには，ネットワークのスコア $S(E)$ が遺伝子(頂点)毎に以下のように分割できる必要がある．つまり，

$$S(E) = \sum_{X_j \in V} S_j(X_j, Pa(X_j)).$$

事後確率に基づくスコア(式(2.2))はこのような分割が可能である．

2.2 関連研究

遺伝子ネットワークを推定するために，これまで様々な数理モデルとその推定のためのアルゴリズムが提案されてきた．ブーリアンネットワークは遺伝子の発現量を On または Off の 2 値として扱い，遺伝子間の関係を AND, OR, NOT の論理式で表現したモデルである．ブーリアンネットワークを用いた研究としては Liang et al. (1998)や Akutsu et al. (1998)などが挙げられる．前者は時系列の発現データ，後者は遺伝子破壊・強制発現実験などによって得られる発現データを用いてブーリアンネットワークを推定するための具体的な方法を示した論文である．また Akutsu et al. (1999)はブーリアンネットワークを一意に定めるために必要なサンプル数を理論的に求めている．ブーリアンネットワークには，アルゴリズムの計算量やネットワーク推定に必要なマイクロアレイの数などに関して理論的な解析が行えるという利点が挙げられるが，一方でデータに含まれるノイズを考慮していない点や，データを 2 値化するため情報の損失が生じる等の欠点がある．前者の欠点を解決する試みとして Akutsu et al. (2000)や Shmulevich et al. (2002)はノイズを考慮したブーリアンネットワークとその推定法を提案している．確率モデルに基づく遺伝子ネットワーク推定の手法として，Murphy and Mian (1998)や Friedman et al. (1998)はグラフィカルモデルのひとつであるベイジアンネットワークを用いた手法を発表している．前者は時系列の発現データからダイナミックベイジアンネットワークを推定する方法である．しかしながら，ベイジアンネットワークにおいても発現データを離散的に扱うため情報の損失が生じてしまう．発現データはそもそも連続的な量であるため，時系列発現データに対しては D'haeseleer et al. (1999)は線形差分方程式を利用した方法，Chen et al. (1999)，de Hoon et al. (2003)は線形微分方程式を利用した方法をそれぞれ提案している．ベイジアンネットワークを連続値データに適用したものとしては Friedman et al. (2000)が線形モデルに基づく方法，Imoto et al. (2002)はノンパラメトリック回帰モデルに基づく方法を開発している．遺伝子ネットワークを推定する上でベイジアンネットワークの欠点の一つは，そのネットワーク構造が非循環なものに限られる点である．そもそも真の遺伝子ネットワークには，フィードバック制御，セルフループなど循環型の制御が存在する．時系列発現データを用いることでダイナミックベイジアンネットワーク(Murphy and Mian, 1999; Kim et al., 2004)により閉路を含む遺伝子ネットワークを推定可能である．また近年，状態空間モデル(Kitagawa,

1998; Higuchi and Kitagawa, 2000)に基づく遺伝子ネットワーク推定手法(Rangel et al., 2004; Yamaguchi and Higuchi, 2006; Yamaguchi et al., 2006; Hirose et al., 2006)についても研究が進められている．これまで取り上げた手法は，遺伝子間の制御の向きを有向グラフとして推定するモデルである．一方， Toh and Horimoto (2002)によるグラフィカルガウシアンモデルを用いた方法や Basso et al. (2005)による無向グラフを用いた遺伝子ネットワーク推定法も提案されている．本稿で解説する手法は遺伝子ネットワークのモデルとしてノンパラメトリック回帰モデルに基づくベイジアンネットワーク(Imoto et al., 2002)およびダイナミックベイジアンネットワーク(Kim et al., 2004)を用いている．そのため，発現データの離散化などによる情報の損失の問題は生じない．また，遺伝子発現制御の依存関係は一般に線形ではないため，ノンパラメトリック回帰モデルは現実の遺伝子ネットワークを推定するためのより適したモデルと言える．

マイクロアレイにより細胞内で発現している mRNA のほぼ全量を網羅的に計測できるようになった．様々な実験的状況下においてマイクロアレイ実験を繰り返すことにより，真の遺伝子ネットワークに従う遺伝子発現のスナップショットを大量に蓄積することができ，それらの観測データに基づき遺伝子ネットワークを推定することが，バイオインフォマティクスにおけるいわゆるマイクロアレイデータに基づく遺伝子ネットワーク推定であった．しかしながら，マイクロアレイ実験には金銭的コストが必ずかかりデータをいくらでも多く観測するわけにはいかない．また，(1)マイクロアレイによる mRNA 発現データと細胞内で実際に働くタンパクの発現量との相関が遺伝子の種類にもよるが決して高くないこと，(2)ヒトなどの高等生物においては選択的スプライシングにより一つの遺伝子から複数の mRNA およびタンパクが生成されるにもかかわらず，マイクロアレイでは選択的スプライシングによる mRNA のバリエーションが完全には考慮されていないこと，(3)RNA 転写阻害(RNAi)による遺伝子以外の物質が遺伝子転写制御へ介入していることが指摘されている．マイクロアレイによる遺伝子発現データは，転写因子を中心とする遺伝子制御の情報を内在し，その重要性は揺るぎないものであるが，さらに遺伝子ネットワーク推定を生物学における知識発見にまで結びつけるようなより高精度なものにするために，マイクロアレイデータに加えて遺伝子発現制御に関わる生物学的データの併用が研究されている．

Hartemink et al. (2002)は既存の転写因子の結合位置情報に基づいてネットワーク探索の範囲を絞ることによって，既知の生物学的な情報をネットワーク推定に反映させる方法を考案した．Imoto et al. (2004)は，生物学的な知識をネットワーク推定に利用するためのより一般的な枠組みを提案している．本稿で解説する Tamada et al. (2003)はこの枠組みを利用し，制御配列探索と遺伝子ネットワーク推定を同時に行う方法を開発した．配列探索の情報を遺伝子ネットワークの推定に用いるだけでなく，遺伝子ネットワークの推定結果を配列探索にも用い，ネットワーク推定と配列探索の 2 つを交互に行うことが特徴となっている．Imoto et al. (2004)の方法は，ネットワーク中の各枝に対して生物学的な裏付けがあるか・ないかの離散的な情報をネットワーク推定に使用するものであるが，一般に生物学的な情報は連続値をとる場合が多い．この弱点を克服する方法として Bernardo and Hartemink (2005)は連続値として与えられた生物学的知識をネットワーク推定に使用する手法を開発している． Imoto et al. (2004)及び Bernardo and Hartemink (2005)は 1 種類の情報しかネットワーク推定に使用することができないが，Imoto et al. (2006)では離散量・連続量が混在した複数の事前情報を用いるための枠組みを示し，それを利用して薬剤反応経路の同定を行った．本稿で解説するもう一つの方法である Tamada et al. (2005)は，生物学的な情報に基づいて 2 つの遺伝子ネットワークを，互いに情報を補いながら推定することが他にはない特徴となっている．

3. 制御配列情報を利用した遺伝子ネットワーク推定手法

本章では，Tamada et al. (2003)が開発した，遺伝子の DNA 配列上流部位に存在する制御配列の情報を用いた遺伝子ネットワーク推定手法について解説する．まずはじめに，実際の遺伝子発現のしくみをもう少し詳しく見てみることにする．遺伝子の発現制御を行っている転写因子は，遺伝子の DNA 配列上流領域に存在する特定の短い配列に結合することによって，特定の遺伝子の発現を制御している．従って，同一の転写因子から制御を受ける遺伝子の DNA 上流配列には，共通の短い特徴的な DNA 配列が存在する可能性が高い．転写因子とその被制御遺伝子との関係は，遺伝子ネットワーク中では直接の親子関係として表現される．従って，このような共通に存在する制御配列の情報を活用することによって，直接の制御か否かを識別できる可能性がある．図 2 に示したのが，遺伝子ネットワークとそれに対応する遺伝子制御の例である．左図は転写因子 FKH2 が 3 つの遺伝子 ACE2，CLB2，SWI5 を制御していることの有向グラフ表現である．左図の情報をより生物学的に解釈すると右図のようになる．つまり，FKH2 は 3 つの遺伝子 ACE2，CLB2，SWI5 の上流領域にある制御配列 GTAAACA に結合することによりこれらの遺伝子が発現していることを示している．

開発した手法の実行手順を図 3 に示す．まず発現データから，ベイジアンネットワークを用いて遺伝子ネットワークを推定する．次に推定された遺伝子ネットワークから転写因子の候補となる遺伝子を選び出し，その遺伝子に対してネットワーク上で下流にある遺伝子から共通配列を探索する．次に発見した配列の情報に基づきネットワークを修正する．最後に制御配列の

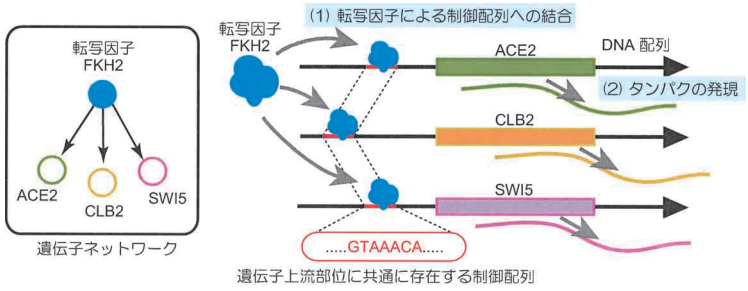


図 2. 遺伝子ネットワークと転写因子による遺伝子発現制御との関係.

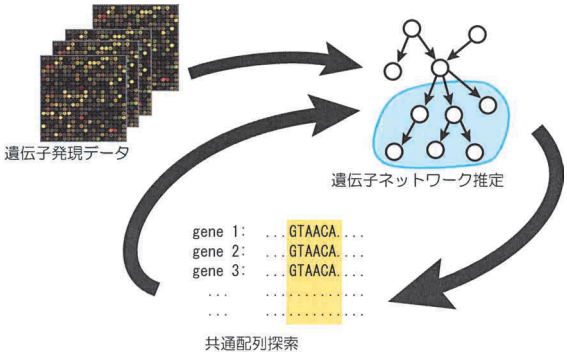


図 3. 発現データと制御配列情報を利用した遺伝子ネットワーク推定の概念図.

情報と発現データを組み合わせて，遺伝子ネットワークを再推定する．制御配列の情報はネットワークの事前確率を構成するために利用される．以上のステップを推定されるネットワークが変化しなくなるまで繰り返すことにより共通配列の探索と遺伝子ネットワークの推定を同時に行う．

遺伝子の DNA 配列上流部位(プロモータ領域と呼ばれる)に存在する制御配列は，すでに述べたように遺伝子発現制御のメカニズムそのものである．しかしながら，それぞれの転写因子が結合する配列に関しては，一部の転写因子を除いて未だ明らかにされていない．正確に制御配列を同定するためには，実際に制御配列の候補部位を実験的に切断し転写因子が結合しなくなるかを確認する必要があり，これはコストと時間のかかる実験である．そこで，効率よく実験を行うために制御配列を予測する問題は，バイオインフォマティクスにおいて重要な問題の一つと考えられている．これまでに制御配列同定のために用いられてきた方法の多くはクラスター分析に基づいている．すなわち，発現パターンの類似している遺伝子群は同様の制御を受けていると考え，発現データの類似性によって遺伝子をクラスタリングし，同じクラスターに属する遺伝子群のプロモーター領域から共通の配列を発見する．本手法は，推定されたグラフ構造を利用して制御配列を探索する遺伝子群を効率的に定義する手法とも言える．また，制御配列探索の対象となる遺伝子群がどの転写因子によって制御されるか，推定されたネットワークによって示唆されている．従って，制御配列に結合する転写因子の候補も同時に推定していることとなる．これはクラスター分析では得ることのできない重要な情報である．

3.1 共通配列探索手法

本節では推定した遺伝子ネットワークの情報を利用して，DNA 配列中に共通に存在する配列を探索する共通配列探索手法について述べる．共通配列探索の対象となる遺伝子は，推定された遺伝子ネットワークの構造に基づいて選択される．まず，転写因子の候補となる遺伝子を選び出し，その転写因子から制御を受けている可能性のある遺伝子を選び出す．具体的には，ある転写因子の候補に対してその直接の子供，および孫からなる遺伝子群を制御配列探索の対象とする．ここで転写因子の候補と書いたのは，実際には転写因子として働くが，転写因子であることが分かっていない遺伝子を考慮したためである．つまり，ネットワーク推定の結果から多くの遺伝子(ここでは 4 つ以上)を制御している遺伝子は転写因子の可能性が高いと考え，そのような遺伝子を転写因子の候補とした．

一般的にクラスター分析に基づいた共通配列探索では，配列探索の対象となる遺伝子すべてに共通の配列が存在することが仮定されている．一方，推定されたベイジアンネットワークに基づき共通配列探索の対象となる遺伝子を選択する本手法では，各遺伝子毎に転写因子候補に対してそれらがどの程度被制御遺伝子として尤もらしいかを尤度として計算することができる．この尤度が高い遺伝子には転写因子候補に対して共通配列が存在する確率も高く，逆に尤度の低い遺伝子には共通配列が存在する確率は低いと考えるのは自然と言えよう．このような考えに基づいた共通配列探索を行えば，推定された遺伝子ネットワークの情報をより有効に活用した生物学的にもっともらしい共通配列を探索できることが期待できる．Bannai et al. (2002)が開発した string pattern regression と呼ばれる共通配列探索手法は，文字列の集合をあるパターンを持つ集合とそうでない集合に分割する．その際，各文字列に割り当てられた数値(属性値と呼ぶ)を利用し，分割後の郡内分散の和が最小になるパターンを探索することがその特徴である．この配列に割り当てる属性値として，2.1 節で定義されたスコアを用いる．すなわち，転写因子候補を X_{TF} と仮定したとき，遺伝子 X_j の配列に割り当てる属性値として $S(\{(X_{\mathrm{TF}}, X_j)\})$ を用いる．これにより，ネットワークの情報に基づく共通配列探索が実現できる．具体的にこの共通配列探索手法を説明する．$D \subset \Sigma^* \times \mathbf{R}$ を共通配列探索の入力データとする．ただし Σ^*

は文字列の集合，R は実数全体の集合とする．つまり D は文字列（ここではプロモータ領域の DNA 配列）と属性値（ここではネットワーク上での枝の尤度）のペアの集合である． String pattern regression は次の評価関数が最小になるパターン v を探索する．

$$MSE(D,v)=\frac{\sum_{(s,r)\in D_v}(\mu(D_v)-r)^2+\sum_{(s,r)\in D_{\bar{v}}}(\mu(D_{\bar{v}})-r)^2}{|D|}.$$

ただし，$|D|$ は集合 D の要素の数，D_v は v にマッチする文字列とその属性値からなる D の部分集合，$D_{\bar{v}}=D\setminus D_v$，$\mu(D')=\frac{\sum_{(s,r)\in D'}r}{|D'|}$ は部分集合 $D'\subset D$ に含まれる属性値の平均値である．また，ここではパターン v のモデルとしてミスマッチを許さない substring pattern を用いる．その場合，最適なパターンを配列の数に対して線形時間で探索可能である（Bannai et al., 2002）．

3.2 配列情報による事前確率

次に共通配列探索の結果に基づきグラフ構造 G の事前確率 $\pi(G)$ を構成する方法について解説する．共通配列探索により，ある転写因子候補 X_{TF} に対し遺伝子 $Y_1,\ldots,Y_n$ の DNA 配列上流領域から共通配列が発見されたとする．つまり，枝 $(X_{\mathrm{TF}},Y_1),\ldots,(X_{\mathrm{TF}},Y_n)$ には共通配列による生物学的な裏付けがあり，このような枝を含むグラフ構造 G にはより高い事前確率を割り当てたい．Imoto et al. (2004)は，このような生物学的な事前知識を事前確率 $\pi(G)$ として利用する枠組みを提案した．具体的には， $G=(V,E)$ に含まれる枝 $X_i\to X_j$ に対し生物学的な裏付けがあれば ζ_1 を，そうで無ければ ζ_2 を割り当てる．ただし，$0<\zeta_1<\zeta_2$ である．この 2 つのハイパーパラメータを用いて事前確率 $\pi(G)$ を次のように定義する．

$$\pi(G)=Z^{-1}\exp\left(-\sum_{(X_i,X_j)\in E}\zeta_{\alpha(i,j)}\right).$$

ただし $\alpha(i,j)$ は $X_i\to X_j$ に生物学的な裏付けがあれば 1 を，そうでなければ 2 を返す関数であり，Z は規格化定数である．

3.3 アルゴリズム

開発した手法のアルゴリズムを以下にまとめる．

Step 1. 発現データからノンパラメトリック回帰モデルとベイジアンネットワークに基づき遺伝子ネットワークを推定する．推定したネットワークを $G=(V,E)$ とする．

Step 2. 各遺伝子 $X\in V$ に対して，D_X をグラフ構造 G における X の子供遺伝子からなる集合とする．$|D_X|\geq 4$ ならば X を転写因子候補とし，D_X と X の孫遺伝子から共通配列を探索する．

Step 3. 各転写因子候補 X_{TF} について，共通配列探索の結果に基づき以下のルールに従ってネットワーク $\overline{G}$ を構成する．ただし $\overline{G}$ は最初は枝を含まない空グラフとする．また以下の操作で $\overline{G}$ に閉路ができる場合は，その操作を行わないものとする．

（a）共通配列探索によって発見された共通配列が存在する X_{TF} の子および孫遺伝子 Y に対し，枝 $X_{\mathrm{TF}}\to Y$ を $\overline{G}$ に追加する．

（b）X_{TF} の親遺伝子 W の DNA 上流配列から Step 2 で発見された共通配列を探索し，その配列が存在すれば，枝 $X_{\mathrm{TF}}\to W$ を $\overline{G}$ に追加する．

Step 4. 共通配列の情報に基づき事前確率 $\pi(G)$ を構成し，$\pi(G)$ と発現データから遺伝子ネットワークを再推定する．その際，$\overline{G}$ を greedy hill-climbing アルゴリズムの初期ネットワークとして用いる． 推定されたネットワークのグラフ構造を G とする．

Step 5. Step 2 から Step 4 を，得られるネットワークに変化がなくなるまで繰り返す．

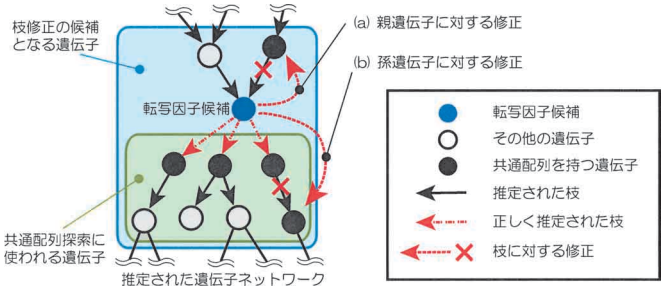


図 4. 配列情報を利用した再推定で期待される枝修正の例.

このアルゴリズムでは得られるネットワークに変化がなくなるまで遺伝子ネットワークの推定と共通配列探索を繰り返し行う．この方法ではアルゴリズムの収束性は保証されないが，現実的には比較的少ない回数で終了する．

図 4 にネットワークの再推定によって期待される枝の修正例を示す．共通配列探索で発見された配列を持つ遺伝子は，転写候補遺伝子に対して直接の子供遺伝子である可能性が高いと言える．従って，配列をもつ転写因子候補の孫遺伝子に対しては図中の(b)で示したような修正を施す．また，転写因子候補の親遺伝子についても発見された配列が存在していれば，向きが間違って推定された枝である可能性が高い．従って，図中の(a)のような修正を施す．

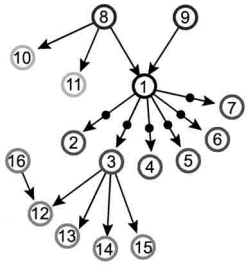
3.4 計算実験

開発した手法の有効性を検証するために，モンテカルロシミュレーションと実データへの適用を行った．

3.4.1 モンテカルロシミュレーション

データ． **図 5 に示す遺伝子ネットワークを設計し，各遺伝子間の関係として図中に示すような関係式を割り当てた．このような設定の下で 100 枚分の疑似マイクロアレイデータを発生させネットワーク推定を行う．疑似マイクロアレイデータには実際の DNA マイクロアレイと同程度の signal-to-noise ratio となるように正規乱数によるノイズを加えている．共通配列探索のための DNA 配列は A，T，C，G の各 4 文字をランダムに発生させる．gene1 を転写因子と仮定し，gene2 から gene7 の配列には共通配列 TATAT を埋め込んだ．また，埋め込んだ共通配列がそれ以外の遺伝子に存在する場合は，発生させた配列中からあらかじめ取り除いた．モンテカルロシミュレーションは 1000 回繰り返し，制御配列探索を組み合わせた提案手法と発現データだけを用いた従来の手法とを比較した．**

結果． **モンテカルロシミュレーションの結果を表 1 にまとめた．行(I)と(II)が共通配列探索と組み合わせた提案手法と発現データだけを用いた従来の手法の結果を表している．列 Sp はいわゆる “specificity” で推定された枝のうち正解の枝の割合を表す．列 Sn は “sensitivity” で正解の枝のうち推定することのできた枝の割合を表す．事前確率 $\pi(G)$ を構成するためのハイパーパラメータは，これら specificity と sensitivity の積の値が最も大きくなるように設定した($\zeta_1 = 1.0, \zeta_2 = 7.0$)．発現データと共通配列の情報を組み合わせることによって，特に Sp の値が大きく上昇していることが分かる(0.384 → 0.540)．正しく推定された枝の数はわずかしか多くなっていないが(10,639→10,768)，間違って推定された枝(不正解数)は著しく減少している(12,727→4,934)．埋め込んだ共通配列 TATAT が正しく推定できた回数は 1,000 回の実験中 433 回であった(表中(III))．図 6 に配列情報ありの場合となしの場合で推定される遺伝子ネッ**



$$
\begin{aligned}
\mathrm{gene1} &= 1.2\ \mathrm{gene8} + 0.8\ \mathrm{gene9} + \varepsilon_1 \\
\mathrm{gene2} &= 0.6\ \mathrm{gene1} + \varepsilon_2 \\
\mathrm{gene3} &= \begin{cases} -1 + \varepsilon_3, & \mathrm{gene1} \le -0.5 \\ \mathrm{gene1} + \varepsilon_3, & -0.5 < \mathrm{gene1} \le 0.5 \\ 1 + \varepsilon_3, & 0.5 < \mathrm{gene1} \end{cases} \\
\mathrm{gene4} &= \begin{cases} 0.4\ \mathrm{gene1} + 1.0 + \varepsilon_4, & \mathrm{gene1} \le -0.3 \\ (\mathrm{gene1} + 1)^2 + \varepsilon_4, & -0.3 < \mathrm{gene1} \le 0.3 \\ 0.4\ \mathrm{gene1} + 1.0 + \varepsilon_4, & 0.3 < \mathrm{gene1} \end{cases} \\
\mathrm{gene5} &= \cos(1.4\ (\mathrm{gene1} + 3.7)) + \varepsilon_5 \\
\mathrm{gene6} &= 0.6\ \mathrm{gene1} + \varepsilon_6 \\
\mathrm{gene7} &= 0.7\ \mathrm{gene1} + \varepsilon_7 \\
\mathrm{gene8} &= \varepsilon_8 \\
\mathrm{gene9} &= \varepsilon_9 \\
\mathrm{gene10} &= \begin{cases} 0.4\ \mathrm{gene8} + 1.0 + \varepsilon_{10}, & \mathrm{gene8} \le -0.3 \\ (\mathrm{gene8} + 1)^2 + \varepsilon_{10}, & -0.3 < \mathrm{gene8} \le 0.3 \\ 0.4\ \mathrm{gene8} + 1.0 + \varepsilon_{10}, & 0.3 < \mathrm{gene8} \end{cases} \\
\mathrm{gene11} &= 1.0/(1 + \exp(-4\mathrm{gene8})) + \varepsilon_{11} \\
\mathrm{gene12} &= 0.8\ \mathrm{gene16} + 0.6(\sin \mathrm{gene3}) + \varepsilon_{12} \\
\mathrm{gene13} &= 1.3\ \mathrm{gene3} + \varepsilon_{13} \\
\mathrm{gene14} &= \begin{cases} 0.4\ \mathrm{gene3} + 1.0 + \varepsilon_{10}, & \mathrm{gene3} \le -0.3 \\ (\mathrm{gene3} + 1)^2 + \varepsilon_{10}, & -0.3 < \mathrm{gene3} \le 0.3 \\ 0.4\ \mathrm{gene3} + 1.0 + \varepsilon_{10}, & 0.3 < \mathrm{gene3} \end{cases} \\
\mathrm{gene15} &= \begin{cases} 0.2\ \mathrm{gene3} - 1.0 + \varepsilon_{15}, & \mathrm{gene3} \le -0.2 \\ 1.4\ \mathrm{gene3} + \varepsilon_{15}, & -0.2 < \mathrm{gene3} \end{cases} \\
\mathrm{gene16} &= \varepsilon_{16}
\end{aligned}
$$

図 5. 設計したシミュレーション用のネットワークと遺伝子間関係式.

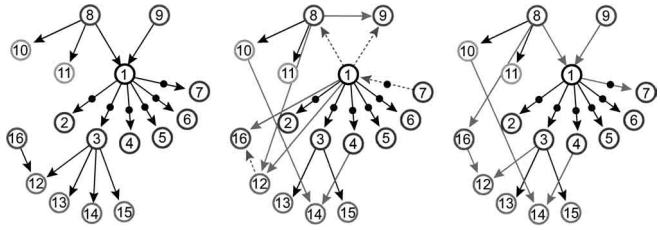
表 1. モンテカルロシミュレーションの結果.

	実験	正解枝	向き違い	不正解枝数	Sn	Sp
(I)	共通配列あり (1000)	10,768	2,086	4,943	0.718	0.540
(II)	共通配列なし (1000)	10,639	2,898	12,727	0.709	0.384
(III)	TATAT detected in (I) (433)	4,785	823	2,118	0.737	0.556
(IV)	TATAT not detected in (I) (567)	5,983	1,263	2,825	0.703	0.528

トワークの典型的な例を示す．以上のことからモンテカルロシミュレーションにおいては，共通配列の情報を用いることにより，より正確な遺伝子ネットワークが推定できることが確認できた.

3.4.2 実データへの適用例

データ．次に開発した手法を実データへ適用した結果を示す．対象とする生物種は出芽酵母であり，遺伝子発現データは遺伝子破壊実験によって得られたマイクロアレイ 120 枚を用いた



(a) 真のネットワーク (b) 共通配列なし (c) 共通配列あり

図 6. 推定されるネットワークの典型的な例.

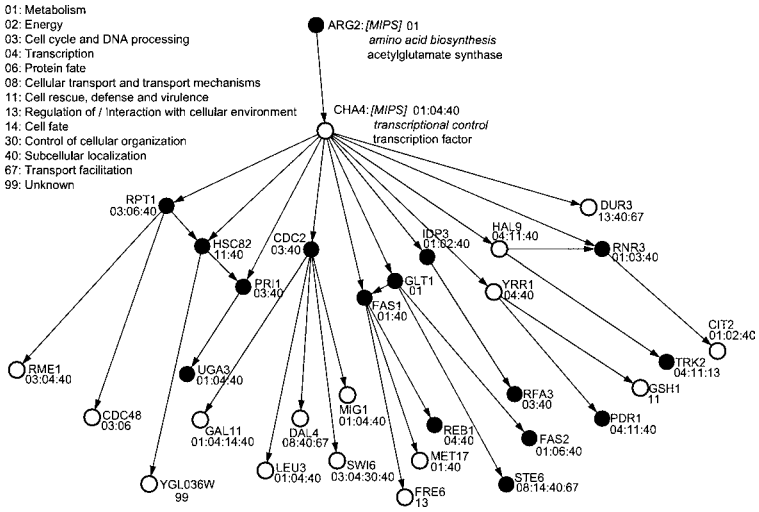


図 7. 配列情報を用いない場合の出芽酵母の遺伝子ネットワーク推定結果.

(Aburatani et al., 2003**)．ネットワーク推定の対象となる遺伝子として，**Imoto et al. **(**2002**)によって推定された出芽酵母の遺伝子ネットワークにまず着目した．その中から，転写因子である** CHA4**，**GAL11**，**SWI6** から距離が** 3 **以下の** 124 **遺伝子を用いた．**DNA **配列は** GenBank **(**http://www.ncbi.nlm.nih.gov/Genbank/index.html**)データベースから上流** 800 bp **を各遺伝子に対して抜き出したものを用いた．**

結果． 本手法をデータに適用した結果，4 **回の繰り返しで推定されるネットワークが変化しなくなりプログラムが終了した．遺伝子** CHA4 **が転写因子候補として選択されたため，その遺伝子に着目して結果を解析した．図** 7 **と図** 8 **に配列情報を使わなかった場合と，使った場合の** CHA4 **周りの部分ネットワークをそれぞれ示す．図中の遺伝子に付けられた番号は** MIPS functional category **における機能分類を表す．本手法を適用することによって** GAL11 **と** GAL2 **の関係が正しく推定されるようになった．**GAL11 **は** GAL2 **を制御していることが知られているが(**Suzuki et al., 1988**)，配列情報を使わない場合** GAL2 **は** ARG2 **の親として現れ，既知の制御における上下関係とは逆である．一方で本手法を適用したネットワークには** GAL11 **が** GAL2 **の上流に現れ，既存の知識と一致する．**

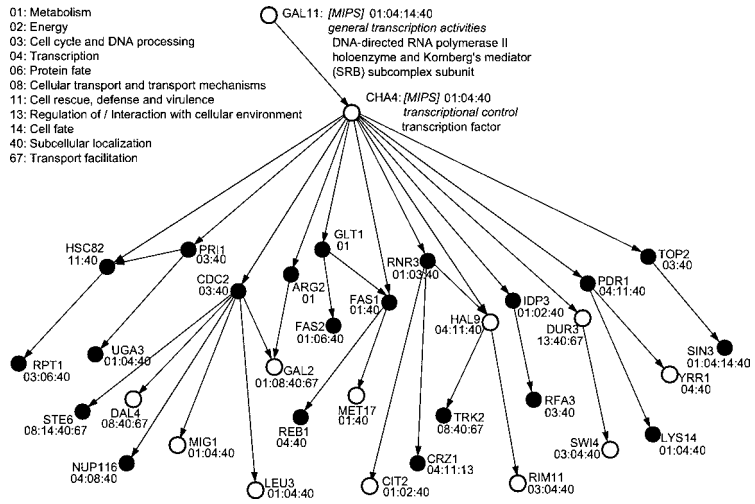


図 8. 提案する手法による出芽酵母の遺伝子ネットワーク推定結果.

表 2. 既知の MCM1-FKH2 **結合配列（上段）と新たに見つかった配列（下段の** REB1 **と** ARG2**）.**

	MCM1	FKH2
	CCY-WWWNN-RG	RYMAAYA
ACE2	CtC-AAAA-CGGcaaaat-GTAAACAttggc	
HOF1	tCC-TcTT-TGGgcaagttGTAAACAataaa	
ALK1	CCC-TTTT-TGGtaaaa-cGTAAACAaaata	
SUR7	CCC-AATCG-GGaaaa-ttGTAAACAttagc	
BUD4	CCC-gATTT-GGaaaaa-gGTAAACAacaat	
SWI5	CCT-gTTTA-GGaaaaa-gGTAAACAataac	
CLB2	CC-GAATCA-GGaaaa--gGTCAACAacgaa	
REB1	CCaaccTAA-AGtaaataaATAAACAtcatc	
ARG2	CCagTTccACGGcaactcacTAAACctatcc	

Y = C or T, W = A or T, R = A or G, M = A or C

4 **回の繰り返しでそれぞれ探索された共通配列は** AAAGA**,** AAACG **(**2 **回),** TAAAC **であった．このうち** TAAAC **は転写因子** FKH2 **の結合配列として既知の配列であった(**Boros et al., 2003**)．**FKH2 **は** MCM1 **という別の転写因子とタンパク複合体を作り** ACE2 **や** SWI5 **などの遺伝子を制御することが知られている．**MCM1 **の結合配列も既知であるため，**FKH2 **の結合配列** TAAAC **を持つとされた遺伝子(図中で黒丸で示した遺伝子)の上流から** MCM1 **の結合配列が見つかれば，それらの遺伝子は** MCM1-FKH2 **複合体によって制御されている可能性が高いと言える．実際に，それらの遺伝子の** DNA **配列から** MCM1 **結合配列を探索したところ** REB1 **と** ARG2 **からよく似た配列が発見された(表** 2**)．**FKH2 **と** MCM1 **はネットワーク推定の対象遺伝子に含まれていなかったが，**CHA4 **は** FKH2 **との関連が示唆されている(**Yang et al., 2005**)．また** FKH2 **は細胞周期に関与する転写因子であるが，図** 2 **のネットワークで** CHA4 **下流には細胞周期に関与しているものも多い(**8 **個)．今回** MCM1-FKH2 **の結合配列を持つとされた** REB1

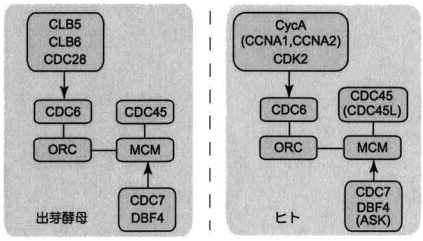
と ARG2 は MCM1-FKH2 転写因子から制御を受けていることは知られていないが，以上のことから新しい MCM1-FKH2 の制御遺伝子である可能性がある．

4. 異種生物種間に進化的に保存された情報を活用した遺伝子ネットワーク推定手法

本章では Tamada et al. (2005) が開発した進化の情報を利用した遺伝子ネットワーク推定手法について解説する．細胞周期など，細胞が生きていくために必要な基本的な機能は，異なる生物種においてもよく保存されており，それらに関わる遺伝子から生成されるタンパク配列もよく類似している．このような遺伝子は，進化の過程において同一の遺伝子を起源としてもつと考えられており，そのような遺伝子同士を ortholog 遺伝子という．Ortholog 遺伝子間の制御関係もまた異なる種においてもよく保存されていることが分かっている．従って，そのような遺伝子間に進化的に保存された情報を活用することによって，推定される遺伝子ネットワークの精度を高められることが期待できる．たとえば，2 つの生物種 A と B において X_a, X_b を A の遺伝子，Y_a，Y_b を B の遺伝子とし，X_a と Y_a, X_b と Y_b をそれぞれ ortholog 遺伝子同士とする．すなわち X_a と Y_a, X_b と Y_b はそれぞれ同一の起源から進化した遺伝子であり，タンパクの配列が類似している．仮に生物種 A において $X_a \to X_b$ という制御関係があった場合，生物種 B においても対応する遺伝子同士に同じような関係，すなわち $Y_a \to Y_b$ という制御関係があることが期待できる．また X_a と X_b の間になんら関係が無い場合，Y_a と Y_b の間においても関係がないことが予想される．図 9 は KEGG データベース (Kanehisa and Goto, 2000, http://www.genome.ad.jp/kegg/) に登録されている出芽酵母とヒトの細胞周期におけるネットワークを部分的に取り出したものである．右の表に，ネットワーク中の ortholog 遺伝子のリストとそのタンパク配列の類似度を示した．表中の BLAST E-value は BLAST (Altschul et al., 1990) によって計算されるタンパク配列の類似度を表す指標で，値が小さいほど配列が類似していることを表す．この図から類似した遺伝子同士の関係もまたよく類似しており，これらのネットワークは進化的によく保存されていると言える．このような進化的に保存された情報を用いて，2 つの生物種の遺伝子ネットワークを同時に推定を行うのが本章で解説する手法の基本的な考えである．

4.1 進化情報を用いた遺伝子ネットワーク推定

今，2 つの生物種 A と B の遺伝子ネットワークをそれぞれ $G_A = (V_A, E_A)$，$G_B = (V_B, E_B)$ で表す．ただし V_A, V_B はそれぞれ生物種 A, B の遺伝子集合，E_A, E_B はグラフ中の有向枝の集合を表す．生物種 A と B の間に保存されている進化情報を H_{AB} で表し，以下で定義され



出芽酵母	ヒト	BLAST E-value
CLB5	CCNA1	4.03×10^{-25}
CLB6	CCNA2	2.73×10^{-25}
CDC28	CDK2	1.65×10^{-85}
CDC6	CDC6	2.31×10^{-32}
CDC45	CDC45L	1.91×10^{-32}
CDC7	CDC7	4.35×10^{-16}
DBF4	ASK	3.10×10^{-6}
MCM	MCM	2.00×10^{-156}
ORC	ORC	4.70×10^{-11}

図 9. (左) KEGG データベースに登録されている出芽酵母とヒトの細胞周期における遺伝子ネットワークの一部．(右) 対応する遺伝子同士のタンパク配列の類似度 (BLAST E-value)．ただし MCM と ORC はこれらのタンパク複合体中でもっとも大きい遺伝子間の E-value を載せている．

る集合とする．

$$\boldsymbol{H}_{AB}=\{(X,Y)\mid \mathrm{e}(X,Y)<\delta,\ X\in V_A,\ Y\in V_B\}.$$

ただし，$\mathrm{e}(X,Y)$ は遺伝子 X と Y のタンパク配列の類似度を表す BLAST E-value であり，δ は閾値である．つまり生物種 A と B のすべての遺伝子の組を考え，E-value が δ 未満のものを ortholog 遺伝子と見なし ortholog 遺伝子ペアの集合を進化情報 $\boldsymbol{H}_{AB}$ と定義する．

生物種 A,B の発現データをそれぞれ $\boldsymbol{X}_A,\boldsymbol{X}_B$ で表す．2 つの遺伝子ネットワークのグラフ構造 G_A,G_B の推定は次の事後確率の最大化に基づいて行う．

$$\pi(G_A,G_B|\boldsymbol{X}_A,\boldsymbol{X}_B,\boldsymbol{H}_{AB}). \tag{4.1}$$

ここで $\boldsymbol{X}_A$ と $\boldsymbol{X}_B$ は独立と仮定すると，式(4.1)は次のように分解することができる．

$$\begin{aligned}\pi(G_A,G_B|\boldsymbol{X}_A,\boldsymbol{X}_B,\boldsymbol{H}_{AB}) &\propto \pi(\boldsymbol{X}_A,\boldsymbol{X}_B|G_A,G_B,\boldsymbol{H}_{AB})\pi(G_A,G_B,\boldsymbol{H}_{AB})\\ &=\pi(\boldsymbol{X}_A|G_A)\pi(\boldsymbol{X}_B|G_B)\pi(\boldsymbol{H}_{AB}|G_A,G_B)\pi(G_A,G_B)\\ &=\pi(\boldsymbol{X}_A|G_A)\pi(\boldsymbol{X}_B|G_B)\pi(\boldsymbol{H}_{AB}|G_A,G_B)\pi(G_A)\pi(G_B).\end{aligned} \tag{4.2}$$

ただし，$\pi(\boldsymbol{X}_A|G_A,G_B,\boldsymbol{H}_{AB})=\pi(\boldsymbol{X}_A|G_A)$，$\pi(\boldsymbol{X}_B|G_A,G_B,\boldsymbol{H}_{AB})=\pi(\boldsymbol{X}_B|G_B)$ とする．式 (4.2)の $\pi(\boldsymbol{X}_A|G_A)$ と $\pi(\boldsymbol{X}_B|G_B)$ はグラフ構造 G_A,G_B が所与の下での発現データの尤度を表し，ベイジアンネットワークを用いて計算する．$\pi(\boldsymbol{H}_{AB}|G_A,G_B)$ は，G_A と G_B の 2 つのグラフ構造が与えられた下での進化情報 $\boldsymbol{H}_{AB}$ の事後確率．また，$\pi(G_A,G_B)=\pi(G_A)\pi(G_B)$ は G_A および G_B に対する事前確率である．$\pi(\boldsymbol{H}_{AB}|G_A,G_B)$ のモデリングが進化情報を利用した遺伝子ネットワーク推定で本質的となる．事前確率 $\pi(G_A)$ および $\pi(G_B)$ に関しては，特に事前の知識などを仮定しないのであれば定数として考える．次に $\pi(\boldsymbol{H}_{AB}|G_A,G_B)$ をどのように構成するかについて説明する．

4.2 進化情報のモデリング

前述したとおり，事後確率 $\pi(\boldsymbol{H}_{AB}|G_A,G_B)$ は進化情報 $\boldsymbol{H}_{AB}$ をキーとした 2 つのネットワーク G_A,G_B の類似度に基づき構成する．ここでは(I)ortholog 遺伝子間で共通に枝のある遺伝子ペアの数 n_{P}，(II)ortholog 遺伝子同士で共通に遺伝子間の関係がない遺伝子ペアの数 n_{N} に基づいて構成する．図 10 に n_{P} と n_{N} の数え方の具体例を示す．右図において点線で示したのは，遺伝子間に枝が存在しないことを表す．左図の 2 つのネットワークにおいて ortholog 遺伝子同士のペアは全部で 9 通りあるが，そのうち G_A,G_B 共通に枝があるものは 3 通りである．従って $n_{\mathrm{P}}=3$ である．同様に共通に枝のないものは 4 通りである($n_{\mathrm{N}}=4$)．一方で互い

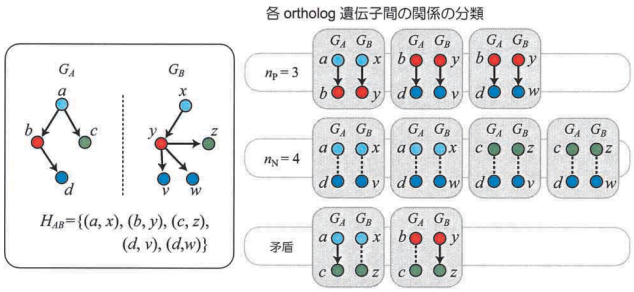


図 10. $\pi(\boldsymbol{H}_{AB}|G_A,G_B)$ を計算する際の n_{P} と n_{N} の数え方の例．

に矛盾している枝は 2 通り存在している．n_{P} と n_{N} を用いて事後確率 $\pi(\boldsymbol{H}_{AB}|G_A,G_B)$ を

$$\pi(\boldsymbol{H}_{AB}|G_A,G_B)=Z^{-1}\exp(\zeta_{\mathrm{P}}n_{\mathrm{P}}+\zeta_{\mathrm{N}}n_{\mathrm{N}})$$

と定義する．ただし Z は規格化定数，ζ_{P} と ζ_{N} (>0) はハイパーパラメータである．

4.3 推定アルゴリズム

次に，事後確率(4.2)式の最大化に基づいて遺伝子ネットワーク G_A,G_B を同時に探索するアルゴリズムについて述べる．2 章で述べたように，最適な遺伝子ネットワークの推定は非常に計算量の多い問題である．さらに本手法の場合，同時に 2 つのネットワークを探索する必要があるため，既存の探索方法を使用することができない．ここでは greedy hill-climbing アルゴリズムを利用した次のネットワーク探索法を用いた．

Step 1. $G_A=(V_A,E_A)$ と $G_B=(V_B,E_B)$ を従来の方法を用いて $\boldsymbol{X}_A$ および $\boldsymbol{X}_B$ からそれぞれ別々に推定する．これらのネットワークは以降の各 step での初期ネットワークとして使用する．

Step 2. G_A および G_B からランダムに遺伝子を選択する．以下，これを X_a で表す．すなわち，$X_a\in V_A\cup V_B$ である．

Step 3. X_b を X_a と同じ生物種の遺伝子とする．可能なすべての X_b に対して以下の操作を考える．

(a) 枝 $X_b\to X_a$ が存在しなければ，X_a の親として X_b を付け加える．

(b) 枝 $X_b\to X_a$ が存在すれば，X_b を X_a の親から取り除く．

スコアを最も向上させる X_b とその操作を実際に G_A または G_B に対して行い，ネットワークを更新する．

Step 4. Step 2 と 3 をスコアが向上しなくなるまで繰り返す．得られたネットワークを出力ネットワークの候補とする．

Step 5. Step 1 から Step 4 を規定回数繰り返す．得られた候補ネットワークのうち最もスコアの良かった遺伝子ネットワークのペア G_A, G_B を探索結果として出力する．

前述したとおり，比較的少ない遺伝子数であれば最適なネットワークを探索可能なので，それを Step 1 として適用可能である．

4.4 計算実験

開発した手法を実データに適用し評価を行った．使用したデータは出芽酵母とヒトの細胞周期における時系列マイクロアレイデータである．マイクロアレイの枚数は出芽酵母 77 枚(Spellman et al., 1998)，ヒト 114 枚(Whitfield et al., 2002)で両方とも公開されている発現データを利用した．この計算実験では時系列データを使用するので，ベイジアンネットワークの代わりに，ダイナミックベイジアンネットワークを用いた(Kim et al., 2004)．次にダイナミックベイジアンネットワークを簡単に解説する．今，$\boldsymbol{X}$ を生物種の区別無しに発現データの行列とする．$\boldsymbol{X}$ の (i,j) 成分 x_{ij} は i 番目に観測されたマイクロアレイデータにおける遺伝子 X_j の発現量を表す．ただし，$i=1,\ldots,N$, $j=1,\ldots,p$，また N は時系列の観測点数(マイクロアレイ数)，p は遺伝子数である． G を推定したい遺伝子ネットワークとすると，ダイナミックベイジアンネットワークでは $\pi(\boldsymbol{X}|G)$ は次のように表される．

$$\pi(\boldsymbol{X}|G)=\pi(G)\int\prod_{i=1}^{N}\prod_{j=1}^{p}f_j(x_{ij}|\boldsymbol{pa}(x_{i-1,j}),\boldsymbol{\theta}_j)\pi(\boldsymbol{\Theta}|\boldsymbol{\lambda})d\boldsymbol{\Theta}.$$

ただし，$\boldsymbol{pa}(x_{i-1,j})$ は遺伝子 X_j の親遺伝子における $(i-1)$ 番目のマイクロアレイの発現デー

タベクトル，$pa(x_{0j}) = \emptyset$，$\pi(\Theta|\lambda)$ はハイパーパラメータ λ で規定されるパラメータ Θ の事前分布である．

4.4.1 計算実験 I

データセット． 開発した手法を評価するために，まず細胞周期において G_1/S 期に特異的に活動する遺伝子に注目した．これは KEGG データベース(http://www.genome.jp/kegg/pathway/hsa/hsa04110.html)において出芽酵母とヒトの G_1/S 期 におけるネットワークがよく類似しており，手法の効果を確かめることに適しているからである．ネットワーク推定の対象となる遺伝子として，Spellman et al. (1998)および Whitfield et al. (2002)において細胞周期に関与しているとされた遺伝子を選び，それらに対して ortholog 遺伝子を加えた出芽酵母 17 遺伝子，ヒト 19 遺伝子を使用した(表 3)．進化情報 H_{AB} を構成する際の BLAST E-value の閾値 δ は，一般的に使われている $\delta = 10^{-5}$ を使用した．ただし，データセット中に含まれる遺伝子が少ないため，1 つの遺伝子が ortholog 遺伝子を最大で 4 つまでしか持たないという制約を加えた．

結果． 推定されたネットワークを図 11 に示す．ダイナミックベイジアンネットワークではネットワーク推定の対象となる遺伝子数が少ない場合，最適なネットワークの探索が容易なため，4.3 節のアルゴリズム Step 1 として最適解のネットワークを使用した．この最適解ネットワークと提案手法で推定したネットワークを KEGG データベースに登録されてある情報と比較した結果を表 4 にまとめた．表中の Sn は sensitivity で KEGG に登録されている枝が推定できた割合，Sp は specificity で推定された枝が KEGG データベースに登録されている割合を表す．進化情報を使った場合，出芽酵母ではあまり推定率は向上しなかったが，ヒトにおいては特に sensitivity が大きく向上していることが分かる(0.244 → 0.478)．ネットワーク推定時のハイパーパラメータは， KEGG データベースに対して sensitivity と specificity の積の値がもっとも大きくなるように設定した($\zeta_P = 1.09, \zeta_N = 0.31$)．

次に提案する手法によって新たに推定された枝に着目し評価を行った．その結果いくつかの枝に関して，生物学的な知識と一致する結果が得られた．図 12 に例を示す．図中で点線は発現データのみから推定された枝，実線は進化情報を用いることによって新たに推定された枝を表す．2 つの生物種間にまたがる線は，つながった遺伝子が ortholog 遺伝子同士であることを表す．図中(a)において，点線で示した CLB6 → CDC6 という出芽酵母における関係が発現データのみから推定された．ヒトにおいてこれに対応するのは CCNA1 → CDC6 であるが発現データのみからでは推定されなかった．これらの関係は KEGG データベースに登録されている関係である．提案する手法を適用することによって，ヒトにおいて正しい枝が推定されるようになった． 図中(b)では発現データのみから出芽酵母において，CDC6 と MCM 複合体を構成する遺伝子間に枝が推定された．これらは KEGG データベースに登録されていない関係であったが，Schepers and Diffley (2001)および Méndez and Stillman (2000)により出芽酵母とヒト両方において直接の相互作用を起こすことが知られている．出芽酵母のみからではヒトにおいてこれらの関係は推定されなかったが，提案手法により推定されるようになった．

表 3. 計算実験 I で使用する遺伝子リスト．

出芽酵母 (17 genes)	ヒト (19 genes)
CDC20, CDC28, CDC45, CDC46, CDC47, CDC54, CDC6, CDC7, CLB5, CLB6, CLN1, CLN2, MCM2, MCM3, MCM6, ORC1, SIC1	CCNA1, CCNA2, CCND1, CCNE1, CCNE2, CDC25A, CDC45L, CDC6, CDC7, CDK7, CDKN1B, MCM2, MCM3, MCM4, MCM5, MCM6, MCM7, ORC1L, ORC3L

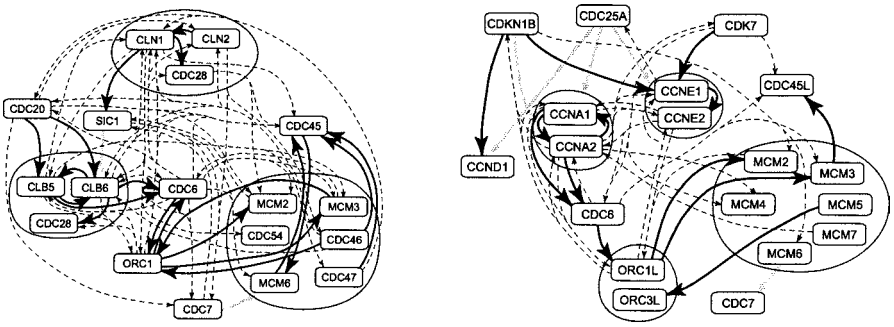


図 11. **計算実験 I で推定された出芽酵母とヒトの遺伝子ネットワーク．実線は KEGG データベースに登録されていた枝，点線は登録されていなかった枝，灰色は登録されていたが推定されなかった枝をそれぞれ表す．ただし MCM および ORC タンパク複合体を構成する遺伝子間の枝は省略した．**

表 4. **計算実験 I の比較結果．データベースに登録されている枝の推定率で比較した．**

	出芽酵母		ヒト	
	Sn	Sp	Sn	Sp
開発した手法	0.682	0.500	0.478	0.571
従来の手法	0.540	0.482	0.244	0.440

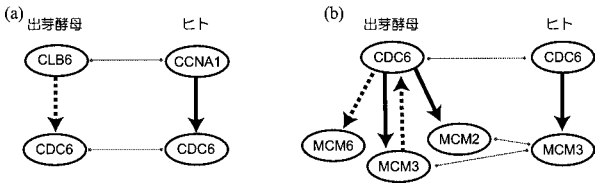


図 12. **計算実験 I における進化情報によって修正された枝の例．**

4.4.2 計算実験 II

データセット． **計算実験 I は KEGG データベースに登録されている遺伝子を中心にネットワーク推定を行ったが，計算実験 II としてより大きいデータセットに対して開発した手法を適用した．選択した遺伝子は出芽酵母 53 遺伝子，ヒト 62 遺伝子である（表 5）．これらの遺伝子は，まず KEGG の細胞周期ネットワークから Spellman et al.（1998）および Whitfield et al.（2002）において細胞周期に関与しているとされた遺伝子をすべて選び（酵母 44 遺伝子，ヒト 38 遺伝子），それらに対して ortholog かつ細胞周期に関与しているとされる酵母 9 遺伝子およびヒト 25 遺伝子を加えたものである．表中で，下線が引かれている遺伝子が KEGG データベースの細胞周期ネットワークに登録されていない遺伝子を表す．**

結果． **KEGG データベースに現れない遺伝子が数多くあるため，計算実験 I のように KEGG データベースとの比較により評価することは難しい．そこでまず Gene Ontology（GO）（Ashburner et al., 2000）による評価を行った．GO は様々な生物種の遺伝子に対して共通の用語を用いて機能を割り当てたもので，仮に 2 つの遺伝子が関連したものであれば，それらの遺伝子**

表 5. 計算実験 II で使用した遺伝子リスト.

出芽酵母 (53 genes)
APC1, BUB1, BUB2, CDC20, CDC45, CDC46, CDC47, CDC5, CDC54, CDC6, CLB1, CLB2, CLB4, CLB5, CLB6, CLN1, CLN2, CLN3, , DAM1, DBF2, DBF20, DUN1, ELM1, FAR1, GIN4, HSL1, HSL7, KIN3, MCD1, MCM2, MCM3, MCM6, MEC3, MOB1, ORC1, PCL1, PCL2, PDS1, PHO5, POL30, PRR1, RAD53, SCC3, SCH9, SIC1, SLT2, SMC1, SMC3, SWE1, SWI4, TEM1, VHS1, YCK1
ヒト (62 genes)
BUB1, BUB1B, BUB3, CCNA2, CCNB1, CCNB2, CCND1, CCNE1, CCNE2, CCNF, CDC16, CDC20, CDC25A, CDC25B, CDC25C, CDC27, CDC42, CDC45L, CDC6, CDC7, CDK7, CDKN1B, CDKN2C, CDKN2D, CENPE, CENPF, CIT, DKFZP434C245, DSP, E2F1, E2F5, FZR1, GADD45A, HDAC3, MAD2L1, MAP2K6, MAP3K2, MAPK13, MCM2, MCM4, MCM5, MCM6, MDM2, MPHOSPH1, NEK2, NKTR, ODF2, ORC1L, ORC3L, PCNA, PKMYT1, PLK, PTTG1, RAB23, RAB3A, RAD21, RAN, ROCK1, SGK, SMC4L1, STK17B, TTK

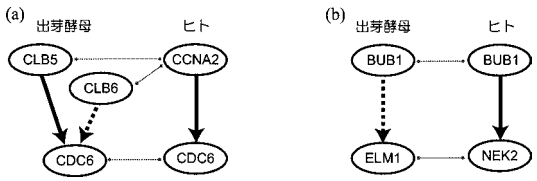


図 13. 計算実験 II において進化情報によって修正された枝の例.

の GO は同じものである場合が多い．発現データのみから出芽酵母，ヒトを別々に推定した遺伝子ネットワークでは，枝で結ばれた 2 つ の遺伝子に共通の GO が出現する数の平均は出芽酵母で 1.45，ヒトで 1.67 であった．一方，開発した手法ではそれぞれ 2.20, 2.31 であった．従って，進化情報を利用して推定した遺伝子ネットワークの方が，GO の遺伝子機能分類に対して一致していると言える． ネットワーク推定時のハイパーパラメータは，推定の対象となる遺伝子のいくつかが KEGG データベースに含まれるため，計算実験 I と同様の方法で決定した($\zeta_P = 2.37$，$\zeta_N = 0.57$)．

次に計算実験 I と同様に，新たに推定された枝に着目して評価を行い，いくつかの枝に関して生物学的な知識と一致する結果が得られた(図 13)．図中(a)において，点線で示した CLB6 → CDC6 という関係が発現データのみから推定された．CLB5/CLB6 → CDC6 および CCNA2 → CDC6 は共に KEGG データベースの細胞周期ネットワークに登録されている既知の関係である．進化情報を用いることにより，これら全ての枝が正しく推定されるようになった．図中(b)では，出芽酵母において BUB1 → BLM1 が発現データから推定された．一方のヒトでは，対応する BUB1 → NEK2 は推定されなかった．NEK2 は MAD1 と相互作用を起こすことが知られており(Lou et al., 2004)，KEGG データベースの細胞周期ネットワークでは BUB1 → MAD1 が存在する．また，NEK1，MAD1，BUB1 は細胞周期の M 期において機能がお互いに関連していることも分かっている(Lou et al., 2004; Brady and Hardwick, 2000)．提案手法を用いることによってヒトにおいて BUB1 → NEK2 が新たに推定されたが，MAD1 がネットワーク推定の対象として含まれていなかったことを考慮すると，この関係は正しいものといえる．一方，出芽酵母において BUB1 → ELM1 は発現データのみから推定されたものの，それらに関係があることは知られていない．ELM1 は NEK2 と同様，細胞周期の M 期に機能する

遺伝子であり，NEK2 と類似したタンパク配列であることから，BUB1 と ELM1 は関連する可能性がある．

5. おわりに

本稿では著者らが開発した異種ゲノムデータを統合した遺伝子ネットワーク推定法として，(1)遺伝子 DNA 配列上流部位に存在する制御配列の情報を利用した方法，(2)異なる生物種間に進化的に保存されている情報を利用した方法を解説した．前者では，モンテカルロシミュレーションへの適用と出芽酵母の発現データを用いた実データ解析により評価を行い，特に実データ解析においては既知の制御配列が探索され，新たな生物学的な関連が示唆される結果が得られた．後者では，出芽酵母とヒトの細胞周期における時系列発現データを用いて評価を行い，発現データだけからでは推定できなかった関係が提案手法により推定されることが確認できた．

今後の課題として以下のことが挙げられる．配列情報を利用した遺伝子ネットワーク推定法では，1 つの転写因子候補に対して共通配列を 1 つしか想定していない．しかしながら一般的に複数の転写因子が組み合わさって遺伝子の制御をする場合が多いため，プロモータ領域には複数の制御配列が存在する可能性がある．従って，このような場合を考慮したネットワーク推定法が必要である．進化情報を利用した手法では，近年急速に近縁種のゲノムデータが蓄積されており，それらを比較分析する研究が盛んに行われている．従って，そのようなデータを活用し 2 種以上の生物種のネットワークを推定する手法へ提案した手法を拡張することによって，より高精度な遺伝子ネットワークを推定できるものと考えている．

遺伝子制御の仕組みはまだ完全には分かっておらず，常に新しい仕組みが発見・報告されている．そのような生物学的な仕組みをモデルに反映させることによって，より高精度な遺伝子ネットワーク推定手法を今後も開発する必要がある．また細胞内の微少な物質を観測する技術は急速に進歩しており，遺伝子発現にかかわる新たなデータも今後出現すると思われる．

謝　辞

本論文について有益なご指摘を下さいました査読者に深く感謝いたします．共通配列探索手法に関しては，九州大学大学院システム情報科学研究院の坂内英夫先生からプログラムの提供および有益な助言をいただきました．また，出芽酵母遺伝子破壊実験による発現データは，九州大学大学院生物資源環境科学研究府の久原哲先生および田代康介先生より提供いただきました．進化情報を利用したネットワーク推定手法の研究では KEGG データおよび実データ解析について東京大学医科学研究所の片山俊明先生から有益な助言をいただきました．ここに深く感謝いたします．本研究は，東京大学医科学研究所ヒトゲノム解析センター・スーパーコンピュータシステムの計算機を利用して行いました．

参 考 文 献

Aburatani, S., Tashiro, K., Savoie, C. J., Nishizawa, M., Hayashi, K., Ito, Y., Muta, S., Yamamoto, K., Ogawa, M., Enomoto, A., Masaki, M., Watanabe, S., Maki, Y., Takahashi, Y., Eguchi, Y., Sakaki, Y. and Kuhara, S. (2003). Discovery of novel transcription control relationships with gene regulatory networks generated from multiple-disruption full genome expression libraries, *DNA Research*, **10**, 1–8.

Akutsu, T., Kuhara, S., Maruyama, O. and Miyano, S. (1998). Identification of gene regulatory networks by strategic gene disruptions and gene overexpressions, *Proceedings of the Ninth*

Annual ACM-SIAM Symposium on Discrete Algorithms, 695–702.

Akutsu, T., Miyano, S. and Kuhara, S. (1999). Identification of genetic networks from a small number of gene expression patterns under the Boolean network model, *Pacific Symposium on Biocomputing*, **4**, 17–28.

Akutsu, T., Miyano, S. and Kuhara, S. (2000). Inferring qualitative relations in genetic networks and metabolic pathways, *Bioinformatics*, **16**, 727–734.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990). Basic local alignment search tool, *Journal of Molecular Biology*, **215**, 403–410.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M. and Sherlock, G. (2000). Gene ontology: Tool for the unification of biology. The gene ontology consortium, *Nature Genetics*, **25**, 25–29.

Bannai, H., Inenaga, S., Shinohara, A., Takeda, M. and Miyano, S. (2002). A string pattern regression algorithm and its application to pattern discovery in long introns, *Genome Informatics*, **13**, 3–11.

Basso, K., Margolin, A. A., Stolovitzky, G., Klein, U., Dalla-Favera, R. and Califano, A. (2005). Reverse engineering of regulatory networks in human B cells, *Nature Genetics*, **37**, 382–390.

Bernard, A. and Hartemink, A. J. (2005). Informative structure priors: Joint learning of dynamic regulatory networks from multiple types of data, *Pacific Symposium on Biocomputing*, **10**, 459–470.

Boros, J., Lim, F. L., Darieva, Z., Pic-Taylor, A., Harman, R., Morgan, A. B. and Sharrocks, D. A. (2003). Molecular determinants of the cell-cycle regulated Mcm1p-Fkh2p transcription factor complex, *Nucleic Acids Research*, **31**, 2279–2288.

Brady, D. M. and Hardwick, K. G. (2000). Complex formation between Mad1p, Bub1p and Bub3p is crucial for spindle checkpoint function, *Current Biology*, **10**, 675–678.

Chen, T., He, H. L. and Church, G. M. (1999). Modeling gene expression with differential equations, *Pacific Symposium on Biocomputing*, **4**, 29–40.

D'haeseleer, P., Wen, X., Fuhrman, S. and Somogyi, R. (1999). Linear modeling of mRNA expression levels during CNS development and injury, *Pacific Symposium on Biocomputing*, **4**, 41–52.

De Hoon, M. J. L., Imoto, S., Kobayashi, K., Ogasawara, N. and Miyano, S. (2003). Inferring gene regulatory networks from time-ordered gene expression data of bacillus subtilis using differential equations, *Pacific Symposium on Biocomputing*, **8**, 17–28.

DeRisi, J. L., Lyer, V. R. and Brown, P. O. (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale, *Science*, **278**, 680–686.

Friedman, N., Murphy, K. and Russell, S. (1998). Learning the structure of dynamic probabilistic networks, *Proceedings of the Fourteenth Conference on the Uncertainty in Artificial Intelligence*, 139–147.

Friedman, N., Linial, M., Nachmann, I. and Pe'er, D. (2000). Using Bayesian network to analyze expression data, *Journal of Computational Biology*, **7**, 601–620.

Hartemink, A. J., Gifford, D. K., Jaakkola, T. S. and Young, R. A. (2002). Combining location and expression data for principled discovery of genetic regulatory network models, *Pacific Symposium on Biocomputing*, **7**, 437–449.

Heckerman, D., Geiger, D. and Chickering, D. M. (1995). Learning Bayesian networks: The combination of knowledge and statistical data, *Machine Learning*, **20**, 197–243.

Higuchi, T. and Kitagawa, G. (2000). Knowledge discovery and self-organizing state space model, *IEICE Transactions on Information and Systems*, **E83-D**(1), 36–43.

Hirose, O., Yoshida, R., Imoto, S., Yamaguchi, R., Higuchi, T. and Miyano, S. (2006). Construction of large gene networks from short time courses of gene expression profiles by state space models (submitted).

Imoto, S., Goto, T. and Miyano, S. (2002). Estimation of genetic networks and functional structures between genes by using Bayesian network and nonparametric regression, *Pacific Symposium on Biocomputing*, **7**, 175–186.

Imoto, S., Higuchi, T., Goto, T., Tashiro, K., Kuhara, S. and Miyano, S. (2004). Combining microarrays and biological knowledge for estimating gene networks via Bayesian networks, *Journal of Bioinformatics and Computational Biology*, **2**, 77–98.

Imoto, S., Tamada, Y., Araki, H., Yasuda, K., Print, C. G., Charnock-Jones, S. D., Sanders, D., Savoie, C. J., Tashiro, K., Kuhara, S. and Miyano, S. (2006). Computational strategy for discovering druggable gene networks from genome-wide DNA expression profiles, *Pacific Symposium on Biocomputing*, **11**, 559–571.

Kanehisa, M. and Goto, S. (2000). KEGG: Kyoto encyclopedia of genes and genomes, *Nucleic Acids Research*, **28**, 27–30.

Kim, S., Imoto, S. and Miyano, S. (2004). Dynamic Bayesian network and nonparametric regression for nonlinear modeling of gene networks from time series gene expression data, *Biosystems*, **75**, 57–65.

Kitagawa, G. (1998). Self-organizing state space model, *Journal of the American Statistical Association*, **93**, 1203–1215.

Liang, S., Fuhrman, S. and Somogyi, R. (1998). REVEAL, a general reverse engineering algorithm for inference of genetic network architectures, *Pacific Symposium on Biocomputing*, **3**, 18–29.

Lou, Y., Yao, J., Zereshki, A., Dou, Z., Ahmed, K., Wang, H., Hu, J., Wang, Y. and Yao, X. (2004). NEK2A interacts with MAD1 and possibly functions as a novel integrator of the spindle checkpoint signaling, *Journal of Biological Chemistry*, **279**, 20049–20057.

Méndez, J. and Stillman, B. (2000). Chromatin association of human origin recognition complex, CDC6, and minichromosome maintenance proteins during the cell cycle: Assembly of prereplication complexes in late mitosis, *Molecular and Cellular Biology*, **20**, 8602–8612.

Murphy, K. and Mian, S. (1999). Modelling gene expression data using dynamic Bayesian networks, Tech. Report, Computer Science Division, University of California, Berkeley, California.

Ott, S., Imoto, S. and Miyano, S. (2004). Finding optimal models for small gene networks, *Pacific Symposium on Biocomputing*, **9**, 557–567.

Rangel, C., Angus, J., Ghahramani, Z., Lioumi, M., Sotheran, E., Gaiba, A., Wild, D. L. and Falciani, F. (2004). Modeling T-cell activation using gene expression profiling and state-space models, *Bioinformatics*, **20**, 1361–1372.

Schepers, A. and Diffley, J. F. (2001). Mutational analysis of conserved sequence motifs in the budding yeast Cdc6 protein, *Journal of Molecular Biology*, **308**, 597–608.

Shmulevich, I., Dougherty, E. R., Kim, S. and Zhang, W. (2002). Probabilistic Boolean networks: A rule-based uncertainty model for gene regulatory networks, *Bioinformatics*, **18**, 261–274.

Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstien, D. and Futcher, B. (1998). Comprehensive identification of cell cycle-regulated genes of the Yeast *Saccharomyces cerevisiae* by microarray hybridization, *Molecular Biology of the Cell*, **9**, 3273–3297.

Suzuki, Y., Nogi, Y., Abe, A. and Fukasawa, T. (1988). GAL11 protein, an auxiliary transcription activator for genes encoding galactose-metabolizing enzymes in *Saccharomyces cerevisiae*, *Molecular and Cellular Biology*, **8**, 4991–4999.

Tamada, Y., Kim, S., Bannai, H., Imoto, S., Tashiro, K., Kuhara, S. and Miyano, S. (2003). Estimat-

ing gene networks from gene expression data by combining Bayesian network model with promoter element detection, *Bioinformatics*, **19**, ii227–ii236.

Tamada, Y., Bannai, H., Imoto, S., Katayama, T., Kanehisa, M. and Miyano, S. (2005). Utilizing evolutionary information and gene expression data for estimating gene networks with Bayesian network models, *Journal of Bioinformatics and Computation Biology*, **3**, 1295–1313.

Toh, H. and Horimoto, K. (2002). Inference of a genetic network by a combined approach of cluster analysis and graphical Gaussian modeling, *Bioinformatics*, **18**, 287–297.

van Someren, E. P. (2002). Genetic network modeling, *Pharmacogenomics*, **3**, 507–525.

Yamaguchi, R. and Higuchi, T. (2006). State-space approach with the maximum likelihood principle to identify the system-generating time course gene expression date of yeast, *International Journal of Data Mining and Bioinformatics*, **1**, 77–87.

Yang, Y.-L., Suen, J., Brynildsen, M. P., Galbraith, S. J. and Liao, J. C. (2005). Inferring yeast cell cycle regulators and interactions using transcription factor activities, *BMC Genomics*, **6**, 90.

Whitfield, M. L., Sherlock, G., Saldanha, A. J., Murray, J. I., Ball, C. A., Alexander, K. E., Matese, J. C., Perou, C. M., Hurt, M. M., Brown, P. O. and Botstein, D. (2002). Identification of genes periodically expressed in the human cell cycle and their expression in tumors, *Molecular Biology of the Cell*, **13**, 1977–2000.

356 Proceedings of the Institute of Statistical Mathematics Vol. 54, No. 2, 333–356 (2006)

Utilizing Heterogeneous Genomic Data to Estimate Gene Networks

Yoshinori Tamada[1], Seiya Imoto[2] and Satoru Miyano[2]

[1]The Institute of Statistical Mathematics
[2]Human Genome Center, Institute of Medical Science, The University of Tokyo

We describe statistical methods for estimating gene networks from gene expression data and other biological information. Since information contained in gene expression data is limited, it is very difficult to accurately estimate gene networks from microarray data alone. This paper introduces two methods for overcoming this limitation. One is to estimate gene networks along with promoter element detection. The other is to estimate gene networks of two distinct organisms utilizing evolutionarily conserved relationships between genes in the two organisms. The former method tries to detect consensus motifs from a set of genes according to the network estimation, then to re-estimate the network along with the detected motifs embedded in a prior probability. The latter method simultaneously estimates two gene networks of two distinct organisms from gene expression data with the evolutionary information. The evolutionary information is defined according to the similarity of the protein sequences of the genes. The both methods use Bayesian networks as models for gene networks and estimate them from the maximization of the posterior probability of the networks. The prior probabilities are constructed based on promoter element detection and evolutionary information, respectively. We evaluate these methods through Monte Carlo simulations and real data analyses. We thus confirm that our methods can estimate gene networks more accurately than previously proposed methods.

Key words: Gene networks, gene expression data, promoter detection, evolutionary information.