日本統計学会誌
第37巻，第1号，2007年9月
55頁～69頁

マイクロアレイ遺伝子発現データからの遺伝子間因果に関する知識発見

井 元 清 哉*

Knowledge Discovery of Causal Relations among Genes from Microarray Gene Expression Data

Seiya Imoto*

マイクロアレイ技術の発展に伴い，遺伝子の発現状態を様々な実験的状況下においてゲノムワイドに知ることができるようになった．そのようなマイクロアレイデータに基づくバイオインフォマティクス，システムバイオロジーにおける最もチャレンジングな話題の一つが遺伝子ネットワークの推定である．本稿では，マイクロアレイデータからの遺伝子ネットワーク推定という問題に対して，我々の研究室が2001年より取り組んできたベイジアンネットワークに基づく手法について説明する．

Along with the development of microarray technology, genome-wide profiling of gene expressions can be done in various experimental conditions. Based on such microarray data, the estimation of gene networks is one of the most challenging problems in bioinformatics and systems biology. In this paper, we introduce the research activity of our group from 2001 for addressing this problem by using Bayesian networks.

Key Words and Phrases: Bayesian networks, nonparametric regression, Bayesian approach, microarray gene expression data

1. はじめに

マイクロアレイ技術の出現は，遺伝子の発現状態をゲノムワイドなスケールで観測可能とし新たな生物学の第一歩であったと同時に，配列解析中心であったバイオインフォマティクスに新たな，そして大きな課題を提示した．はじめにマイクロアレイデータが脚光を浴びたのはクラスター分析である．遺伝子をクラスタリングすることにより遺伝子の機能予測を行う．癌などの患者のマイクロアレイデータをクラスタリングすることにより，ゲノムレベルでの病理診断・予後予測を行う．マイクロアレイデータのクラスター分析（Eisen et al., 1998; Golub et al., 1999; Tamayo et al., 1999）というパラダイムは，医学・生物学に対して大きなインパクトを与えた．また，教師なしクラスター分析のみならず，例えば抗癌剤に対する感受性をクラスラベルとした判別分析に対してもサポートベクトルマシン，ニューラルネットワーク，Adaブーストなど多様な統計手法が適用され多くの成果をあげている（Furey et al., 2000; Takenouchi et al., 2007）．このように，マイクロアレイを用いた医学・生物学のパラダイムシフトに対して，

* 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丁目6番1号
E-mail:imoto@ims.u-tokyo.ac.jp

統計科学の果たした役割は大きい.

遺伝子破壊実験，強制発現実験，化学物質投与など様々な実験的状況下における遺伝子の発現状況のスナップショットをマイクロアレイにより観測することで，極めて広範囲な遺伝子発現に関するデータを得ることができる．システムバイオロジーの大きな目的の一つは，細胞内で行われている遺伝子発現のメカニズムを解明し，システムとして生命を理解することである．細胞内において，遺伝子は互いに独立に働いているのではなく，制御・被制御の関係が複雑に絡み合ったネットワークを構成し生命を維持している．この遺伝子発現に関するネットワークを遺伝子ネットワークとよぶ．マイクロアレイ遺伝子発現データに内在する遺伝子発現制御に関する情報は，遺伝子ネットワークを明らかにする可能性があると考えられ，そのための研究が統計科学的・情報科学的アプローチを用いバイオインフォマティクスにおいて精力的に推進されている.

本稿では，これまでに我々の研究室で行ってきたベイジアンネットワーク（Bayesian network）に基づく遺伝子ネットワーク推定手法の開発研究について紹介する．本稿の構成は次の通りである．2章において，マイクロアレイデータについて基本的な説明を行う．3章において，我々の提案したベイジアンネットワークにB-スプラインに基づくノンパラメトリック回帰モデルを組み合わせた手法を紹介する．4章では，より高度な生物における遺伝子ネットワーク推定を目指した手法としてマイクロアレイデータに他の情報を加えた遺伝子ネットワーク推定について説明する．我々の研究室において，推定した遺伝子ネットワークのアプリケーションの一つとして位置づけている薬剤ターゲット遺伝子の同定について5章にその成果をまとめる．6章では，ベイジアンネットワーク以外の遺伝子ネットワーク推定へのアプローチについて述べる.

2. マイクロアレイデータ

“遺伝子が発現する”とは本来“遺伝子を元にタンパク質が生合成される”ことを意味する．遺伝子を基にしてタンパク質が合成されるプロセスは簡略化すると図1のように表せる．すなわち，ある遺伝子が働いているか否かを知るためには，その遺伝子を基にしてタンパク質が合成されたかどうかを調べればよい．マイクロアレイは，タンパク質を直接観測する代わりに，

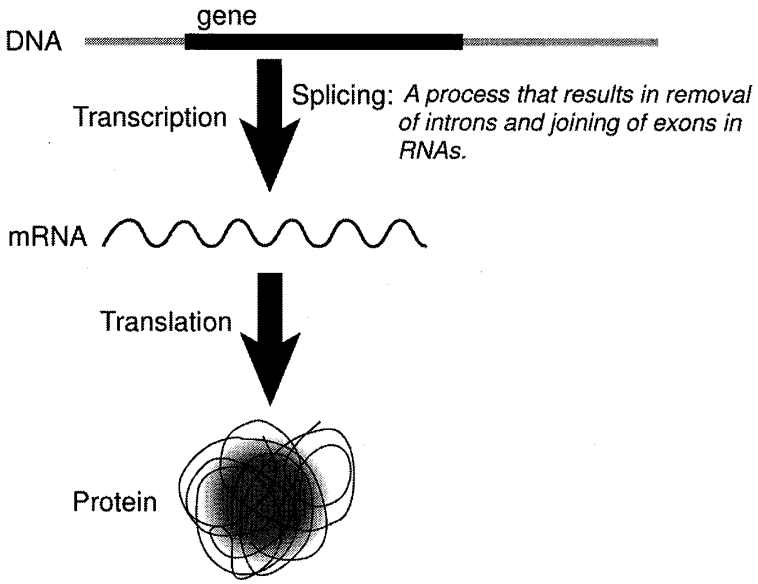

図1 DNAからタンパク質への変換

その前状態である mRNA の量を調べるツールである．

マイクロアレイは，Affymetrix 社の S. P. Fodor らによる光リソグラフィー法に基づくマイクロアレイ（商品名 GeneChip®）と Stanford 大学の P. O. Brown らによるスポット法に基づくマイクロアレイ（いわゆる cDNA マイクロアレイ）に分けられる．cDNA マイクロアレイは研究目的にあわせてマイクロアレイをデザインでき，その簡便さから広く用いられている．一方，GeneChip® は cDNA マイクロアレイの 20 倍近い密度を持ち，規格製品であるため入手も比較的容易であるが研究目的に合わせた注文生産は困難となる．したがって，本稿では研究室レベルでの生産が可能な cDNA マイクロアレイについて取り上げる．

cDNA マイクロアレイにより遺伝子の発現量を測る際は 2 種類の細胞を用意する（理由については後述する）．一般的には，ひとつは通常細胞であり，他方は癌細胞や実験的に処理を施した細胞（ここではサンプル細胞と呼ぶ）である．実験的な処理としては，遺伝子破壊実験（gene disruption），過剰発現実験（overexpression）や Heat shock，Cold shock などのショックが一般的である．図 2 は cDNA マイクロアレイの概念図である．まず，正常細胞とサンプル細胞から全遺伝子に関して mRNA を抽出し，それを鋳型として cDNA を生成する．正常細胞，サンプル細胞から生成した cDNA をそれぞれ蛍光色素 Cy3，Cy5 によって蛍光し，マイクロアレイ上にスポットされた cDNA に対してハイブリダイゼーションする．ここで，ハイブリダイゼーションとは，一本鎖の DNA，RNA を組み合わせることで，二本鎖分子の DNA-DNA，DNA-RNA，RNA-RNA を形成させることである．マイクロアレイ上の 1 つのスポットには 1 つの遺伝子から生成された cDNA が貼り付けられており，蛍光された cDNA と相補的な配列となっているため 2 つの cDNA は水素結合により電気的に引き合う．ハイブリダイゼーション後，マイクロアレイからスキャナーにより各スポットにおける Cy3 と Cy5 の色強度（インテンシティ）がそれぞれ計測される．蛍光色素 Cy3，Cy5 はそれ自体には色は付いていないが，スキャン後ソフトウェア的にインテンシティの大きさに応じてそれぞれ緑色，赤色がコンピュータモニター上で着色される．したがって，ある遺伝子が正常細胞ではほとんど発現せず，サンプル細胞では過剰に発現しているような場合，その遺伝子に対応するスポットは赤

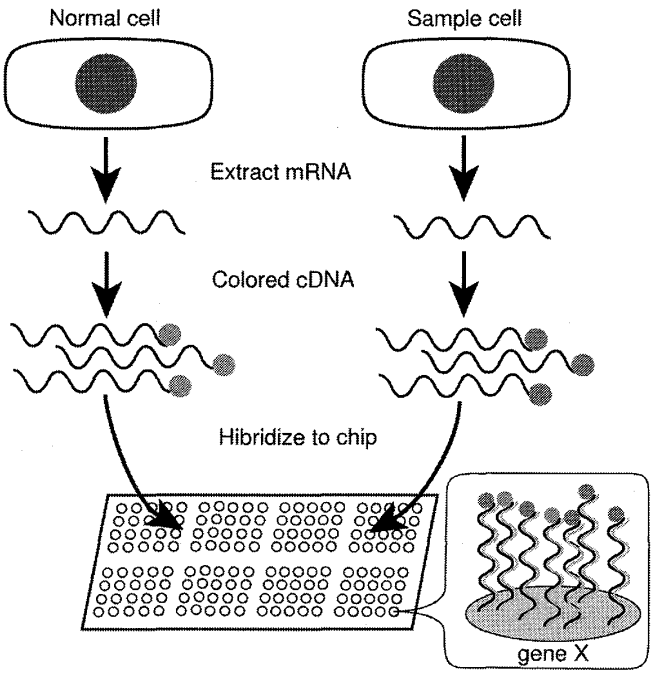

図 2　cDNA マイクロアレイ

色に見えることになる．緑に見えるスポットは正常細胞でのみ発現している遺伝子であり，黄色に見えるスポットは2つの細胞で共に発現しているものである．両方の細胞で共に発現がない遺伝子に関しては色が見えない，つまり黒色に見えることになる．

マイクロアレイ上の各スポットをスキャンし，Cy3とCy5のインテンシティを計測する．得られたCy3，Cy5のインテンシティからバックグランドの Cy3，Cy5のインテンシティをそれぞれ引いたものを観測値とする．ここで，バックグランドのインテンシティとは，マイクロアレイ自体が持っている Cy3，Cy5の色のシグナルであり，cDNAをスポットしない状態で観測される．つまり，1枚のマイクロアレイによる1つの遺伝子の発現状態は，正常細胞のmRNAとサンプル細胞のmRNAの2次元データとして得られる．しかしながら，これら2つのインテンシティは単独では定量性を持たない．つまり，スポット間においてインテンシティの大小の比較はできない．これは，マイクロアレイに貼り付けるcDNAを一定量に保つことが非常に困難であることなどが原因となる．そこで，同一スポットにおける正常細胞とサンプル細胞のインテンシティの比を考える．つまり，i番目の遺伝子に対するCy3, Cy5のインテンシティをそれぞれG_i, R_iとすると正常細胞を基準とした比G_i/R_iをデータとする．通常，データの分布を対称とするために底を2とする対数比を取り解析に用いることが多い．

2色蛍光のcDNAマイクロアレイとは別に，Affymetrix社の製品であるGeneChipなどのオリゴヌクレオチドに基づくマイクロアレイもある．これは，各スポット上に遺伝子から生成されるmRNAとは相補的な配列をした短いRNA断片を光リソグラフィー法により合成し，各遺伝子のmRNAの量を計測するものである．短いRNA断片と書いたが，長さはGeneChipだと25文字である．この25文字のRNA断片を各遺伝子についてそのmRNAの配列から11カ所選び，それらを11個のスポット上にそれぞれ合成し，各遺伝子から生成されるmRNAの量を計測する．各スポット上のRNA断片は実験により合成することになるので，その数はコントロールすることができる．したがって，GeneChipでは2種の細胞を用いることなく各遺伝子から生成されるmRNAの絶対量を計測することができるという仕組みになっている．GeneChipの場合，ある一つの遺伝子が生成するmRNAの量を11個のスポット（11種の配列）により計測するため，これらの値を一つの値にまとめる作業が必要となる．計測したいmRNAと完全に相補的な配列のスポットをパーフェクトマッチ（PM）とよび，意図的に25文字の真ん中の1文字だけ異なる配列のものはミスマッチ（MM）とよばれる．このミスマッチのスポットは，目的以外のmRNAがハイブリするクロスハイブリダイゼーションの影響を除くために設計されたものである．このPMとMMの利用方法や，スポットに固定するのは25文字の短い配列であるため，A，Tよりも結合力の強いG，Cの含有量の違いによるバイアスを除いたりすることでインテンシティの要約方法に対して様々な工夫がなされている．代表的なものとしては，MAS5.0（Affymetrix, 2002; Hubbell et al., 2002），dChip（Li and Wong, 2001），GCRMA（Irizarry et al., 2003），PLIER（Affymetrix, 2005）などの方法があげられる．また，オリゴヌクレオチドに基づく単色のマイクロアレイにおいても，スキャナの設定によりインテンシティはチップ間で比較できないため，チップ間を比較可能とするための正規化が必要となる．また，Kanno et al.（2006）は遺伝子が生成しているmRNAのコピー数を計測することが可能となる技術を開発している．

遺伝子発現データはmRNAの量を計測していたが，mRNAはヒトなどの高等生物では遺伝子をコードしているDNA領域に存在するいくつかのエクソンとよばれる領域が結合して形成されている．DNAからmRNAが形成されるシステムはスプライシングとよばれ，ヒトでは多くの場合，一つの遺伝子から複数のスプライシングパターンによって複数のmRNAが形成され，これらは異なるタンパクに翻訳されることが知られている．これはヒト遺伝子が約3万程

度しかないにもかかわらず，細胞を形成しているタンパク質は10万種を超えることの理由の一つである．この一つの遺伝子から複数種のmRNAを形成する機構は選択的スプライシングとよばれ，異常な選択的スプライシングは疾患に関連することもある．この選択的スプライシングを観測するためにはmRNAよりもより解像度を上げ，エクソンレベルでの発現量の計測が必要となる．そのためのシステムがAffymetrixから2005年12月に発売された．ヒトの約100万のエクソン発現量を計測するためのGeneChip Human Exon 1.0 ST Arrayである．エクソン発現データの解析についてはYoshida et al.（2006, 2007）を参照されたい．

3. ベイジアンネットワーク

我々は，ベイジアンネットワークを用いて各遺伝子の発現量を確率変数と見なした有向グラフを発現データから推定することで，遺伝子ネットワークを推定するアプローチについてこれまでに研究を進めてきた．今，$\mathcal{X}=\{X_1,\ldots, X_p\}$ を確率変数の集合とし，それらの間に非閉路有向グラフ G で表される依存関係があるとする．確率変数間に有向枝に沿ったマルコフ連鎖率を仮定することで非閉路有向グラフ G により確率変数間の条件付き独立が表され，同時確率の分解

$$\Pr(\mathcal{X})=\prod_{j=1}^{p}\Pr(X_j|Pa(X_j)) \tag{3.1}$$

を得る．ただし，$Pa(X_j)$ は確率変数 X_j の G 上での直接の親に対応する確率変数の集合である．

我々の研究は，Friedman et al.（2000）が大きなモチベーションとなっている．Friedman et al.（2000）は離散データに対するベイジアンネットワークを用いるためにマイクロアレイデータを3値に離散化する方法を提案している．その生物学的意味としては，cDNAマイクロアレイによって観測された発現データにおいて，i 番目の遺伝子の発現値 x_i は実験的な処理をしたサンプル細胞での発現量 s_i とコントロール細胞（通常は正常細胞）での発現量 c_i を用いて対数比 $x_i=\log(s_i/c_i)$ として得られていることから，x_i を離散化した値 $\{+1, 0, -1\}$ は，“$+1$＝コントロールに比べて発現している”，“0＝コントロールと有意な差はない”，“-1＝コントロールに比べて発現していない”と見なしていると考えることができる．

Friedman et al.（2000）では，このアプローチを線形回帰モデルを基に構成したベイジアンネットワークと比較することでその優位性を主張している．その優位性の一つのファクターとしては遺伝子間の非線形関係が取り上げられている．つまり，離散データに対するベイジアンネットワークでは非線形な関係が捉えられているが，線形回帰に基づく方法では遺伝子間の関係は線形に限られ高精度な遺伝子ネットワーク推定は難しいというものである．しかしながら，そもそもマイクロアレイデータは連続的な量として観測されるため，離散化による情報の損失は無視できないと考え，我々は B-スプラインに基づくノンパラメトリック回帰を用い遺伝子間の非線形関係を捉えることのできるベイジアンネットワークをImoto et al.（2002）において発表した．また，Imoto et al.（2003a）では，マイクロアレイデータの分散不均一性に対応するための不等分散ノンパラメトリック回帰モデルの利用を提案した．Imoto et al.（2002）のモデルについて以下に説明する．

いま，$\mathcal{X}=\{X_1,\ldots, X_p\}$ に関して n 個の独立な観測値 $\boldsymbol{X}=(\boldsymbol{x}_1,\ldots, \boldsymbol{x}_n)'$ が得られたとする．ここで，$\boldsymbol{x}_i=(x_{i1},\ldots, x_{ip})'$ であり，x_{ij} は，i 番目のマイクロアレイによって計測された j 番目の遺伝子の発現データである．マイクロアレイデータは連続変数であるため，（1）式の分解は密度関数を用いて

$$f(\boldsymbol{X}|\boldsymbol{\Theta}, G)=\prod_{n=1}^{n}\prod_{j=1}^{p}f_j(x_{ij}|\boldsymbol{pa}_{ij}, \boldsymbol{\theta}_j) \tag{3.2}$$

と表すことができる．ただし，$\boldsymbol{\Theta}=(\boldsymbol{\theta}_1',\ldots, \boldsymbol{\theta}_p')'$ はパラメータベクトル，$\boldsymbol{pa}_{ij}$ は i 番目のマイクロアレイによって観測された $Pa(X_j)$ の発現データベクトルである．したがって，ネットワーク G の構築は密度関数 f_j の構成に帰着される．ここでは，f_j に対して正規分布を仮定し，その平均構造に対してはノンパラメトリック回帰

$$x_{ij}=m_{j1}(p_{i1}^{(j)})+\cdots+m_{jq_j}(p_{iq_j}^{(j)})+\varepsilon_{ij} \tag{3.3}$$

により構成する．ここで，$\varepsilon_{ij}(i=1,\ldots, n)$ は互いに独立に平均0，分散 σ_j^2 の正規分布に従う確率変数，$\boldsymbol{pa}_{ij}=(p_{i,1}^{(j)},\ldots, p_{i,q_j}^{(j)})'$ は X_j の親変数に関するデータであり $q_j=|Pa(X_j)|$ とおいた．また，$m_{jk}(\cdot)$ $(k=1,\ldots, q_j)$ は滑らかな回帰関数である．回帰関数 $m_{jk}(\cdot)$ は B-スプラインを用いた基底関数展開によって構成する．つまり，回帰関数 $m_{jk}(\cdot)$ は M_{jk} 個の B-スプライン $\{b_{1k}^{(j)}(\cdot),\ldots, b_{M_{jk}k}^{(j)}(\cdot)\}$ によって

$$m_{jk}(p)=\sum_{s=1}^{M_{jk}}\gamma_{sk}^{(j)}b_{sk}^{(j)}(p) \tag{3.4}$$

と表す．ここで，$\gamma_{sk}^{(j)}(s=1,\ldots, M_{jk})$ はパラメータである．以上をまとめると，ノンパラメトリック回帰を用いたベイジアンネットワークモデル

$$f(\boldsymbol{X}|\boldsymbol{\Theta}, G)=\prod_{i=1}^{n}\prod_{j=1}^{p}\frac{1}{\sqrt{2\pi\sigma_j^2}}\exp\left[-\frac{\{x_{ij}-\sum_k\sum_s\gamma_{sk}^{(j)}b_{sk}^{(j)}(p_{ik}^{(j)})\}^2}{2\sigma_j^2}\right].$$

を得る．ただし，$\boldsymbol{\Theta}=(\boldsymbol{\theta}_1',\ldots, \boldsymbol{\theta}_p')'$ に対して，$\boldsymbol{\theta}_j=(\{\gamma_{sk}^{(j)}\}_{s,t}, \sigma_j^2)'$ である．B-スプラインに基づくノンパラメトリック回帰については，de Boor (1978), Imoto and Konishi (2003) を参照されたい．

なお，B-スプラインによるノンパラメトリック回帰においては，(4) 式の係数パラメータに加えて，各 B-スプライン $b_{sk}^{(j)}$ を規定する節点の位置もパラメータと考えることができ，それらの最適化問題を一般的には解く必要がある（田辺・田中，1983）．しかしながら，この最適化には多くの計算時間が必要となる．そこで，B-スプラインを規定する節点はここでは等間隔に配置し，ある程度多くの B-スプラインを用い，その係数パラメータに曲線の滑らかさに関する事前分布を仮定したベイズモデルを用いることにする．この方法により，節点の位置，B-スプラインの個数の選択が，平滑化事前分布に含まれるハイパーパラメータの選択に置き換えることができ，計算量を大きく減少させることができる．この方法は Eilers and Marx (1996) の提案した P-スプラインと等価な方法である．

ベイジアンネットワークとノンパラメトリック回帰を組み合わせたことにより，ベイジアンネットワークのモデル識別性の問題についても一つの解を得ている．すなわち，ベイジアンネットワークには，同じ条件付き独立性を示す異なるグラフ構造が存在する．図3は全て B 所与の元で A と C が条件付き独立になるグラフの例である．簡単な計算により，各グラフ構造に従い式（1）によって得られる分解から $\Pr(A, C|B)=\Pr(A|B)\Pr(C|B)$ が導ける．この場合，3つのグラフ構造から計算される尤度は，離散型や正規線形回帰モデルに基づくベイジアンネットワークでは等しくなり，同数のパラメータを持つ場合 AIC などの情報量規準によるモデル選択はできない．この問題に対しては，非正規，非線形の2つのアプローチが考えられる（狩野・宮村，2006）．Imoto et al. (2002) は非線形モデルを構築することで，同じ条件付

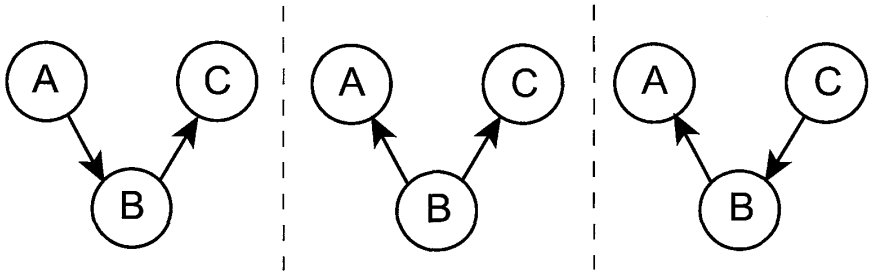

図3 B所与の元でAとCが条件付き独立となるグラフの例.

き独立性を表すグラフを識別したものであると考えることができる．この非線形性を制御関係のキーとしたアプローチは，遺伝子間（mRNA 間）の関係ではないが，遺伝子産物であるタンパク質とそのタンパク質が誘導する遺伝子から転写される mRNA の量との関係は，ミカエリス－メンテン式のような非線形関係が知られていることからの意味づけもできる．

上述した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問題は，確率変数の集合$\mathcal{X}=\{X_1,\ldots, X_p\}$が与えられた元で確率変数間の依存関係をモデリングするという立場で説明した．しかしながら，例えば出芽酵母の遺伝子ネットワークを推定する場合，出芽酵母には約6,000 個の遺伝子が存在するため，6,000 個の確率変数を含むネットワークを推定することになる．しかしながら，後述するように，ベイジアンネットワークの構造決定の問題は NP-困難の問題であるため膨大な計算量を減らすために様々な工夫がなされる．また，遺伝子ネットワーク推定に用いるマイクロアレイは，数百枚にすぎない．もちろん，例えば出芽酵母において公表されているマイクロアレイデータを集めると数千というデータが利用できる．しかしながら，それらは Affymetrix の GeneChip で計測されていたり，Agilent 社のチップであったり，cDNA の研究室独自に作成したマイクロアレイデータであったりと，そのプラットフォーム，実験条件はさまざまである．ヒトの遺伝子ネットワーク推定においては，細胞組織も，肝臓，脳，心臓，腎臓などさまざまであり，SNP などによる個体差，癌などの細胞条件の違いもあり，均一にコントロールされたマイクロアレイデータのライブラリとしてはそれほど多くは集まらない．したがって，全遺伝子でネットワーク推定がたとえ計算機上では可能であっても，その精度はモデルに含まれるパラメータ数と，推定に用いることのできるサンプル数のバランスを考えると決して高くないことが分かる．したがって，今得られているマイクロアレイデータによって推定することの可能な遺伝子セットの適切な定義が必要となる．この点には，現在のところ生物学的な知識に頼るところが大きく，ネットワーク推定を前提とした変数選択法はその研究の進展が期待される．現時点では，興味ある現象に関わるマイクロアレイデータにおいて発現状態が変化した遺伝子から，擬陽性をある程度許容するような緩い基準を用いて抽出するような方法が用いられる．

ベイジアンネットワークの構造推定は極めて計算量が多く，グラフ構造を評価するためにいかなるスコア関数を用いてもその最適化は NP-困難であることが知られている．従って，greedy アルゴリズムや焼きなまし法など発見的な学習手法を採用せざるを得ず，より最適解に近い解を得るためには学習ステップを十分に踏むか，または遺伝子数を減らす必要がある．Greedy アルゴリズムの詳細，計算量を減らすための具体的な工夫については，井元（2007）に詳しい説明がある．なお，比較的少数，具体的には 30 程度の遺伝子からなるネットワークであれば，Ott et al.（2004）は動的計画法をベースとしたアルゴリズムを提案し最適解が現実的な時間で求まることを示した．また，マイクロアレイデータに含まれるはずれ値の取り扱いはやはり遺伝子ネットワーク推定においても重要であり，Imoto et al.（2004b）では移動中央

値平滑化と残差ブートストラップを利用したよりロバストな手法を提案した．

また，はずれ値と同様に，欠損値の取り扱いはマイクロアレイデータの解析にとって重要な問題の一つである．例えば，マイクロアレイ上のゴミが欠損の原因となったり，2章で説明したcDNAのマイクロアレイデータでは，ある遺伝子のCy3とCy5のインテンシティを用いて対数比のデータを生成するが，片方，もしくは両方の色のインテンシティがスキャナの検出限界感度以下であるとその対数比は計算できなくなり欠損値となる．このような場合，欠損がランダムに発生するとは限らず，統計的に厳密な取り扱いは極めて困難となる．マイクロアレイデータでの欠損値の取り扱いについては，統計的に十分な議論を行っているものは現時点では少なく，厳密には問題があるがランダムな欠損と考えてk近傍法などを利用した欠損値補完法を用いて解析前にあらかじめ欠損値なしのデータを生成しておく方法が多くは用いられる．

出芽酵母は約6,000の遺伝子を有する．Imoto et al.（2002）では細胞周期に関わる遺伝子に絞り，細胞周期に関わる遺伝子が影響を受ける実験によって得られた77枚のマイクロアレイデータ（Spellman et al., 1998）を用いてネットワーク推定を行った．ベイジアンネットワークを用いた遺伝子ネットワーク推定において，このような遺伝子の絞り込みは，前述したように計算量を減らすという観点のみならず，データへの過適合を可能な限り避けるためにも必要不可欠といえる．また，Imoto et al.（2003a）では遺伝子破壊実験によって得られたマイクロアレイデータを用いた．遺伝子破壊実験によるマイクロアレイデータとは，ある遺伝子の発現を強制的に止めて観測したマイクロアレイデータであり，発現を止められた遺伝子の下流に位置する遺伝子たちが影響を受ける．ネットワーク推定に用いたマイクロアレイデータは，1枚のマイクロアレイに対して1個の遺伝子の発現を止めた一遺伝子破壊実験によるものである．破壊した100個の遺伝子の大多数は転写因子である．ネットワーク推定に用いた521個の遺伝子は，これら破壊した転写因子から制御を受けていると報告されている遺伝子である．破壊した転写因子は比較的研究の進んだものが多く，推定したネットワークの評価は既知の情報を基に行うことが可能である．逆に，ほとんど研究されていない遺伝子のネットワークを推定しても，その評価は困難なものとなる．いかにして推定したネットワークを評価し，情報を抽出するかは重要な課題である．

4. 他の生物学的データの併用

前述した解析例ではネットワークに含まれる遺伝子数に比べ，推定に用いることのできるマイクロアレイデータ数（サンプル数）は決して十分な数とは言えず，データへの過適合は否めない．マイクロアレイ実験は今でこそコストもかなり下がってきてはいるが，やはり数千というオーダーでデータを揃えるためには相当な資金力が必要となる．また，前述した研究は出芽酵母の遺伝子ネットワーク推定を目的としており，ヒトなどの高等生物のネットワーク推定においては，そのネットワークの複雑さからさらにデータへ過適合する可能性もある．そこで，マイクロアレイデータへの過適合を避け，高等生物の遺伝子ネットワーク推定にも耐えうる汎化能力の高い手法を開発する必要がある．今，ここで考えている問題は遺伝子ネットワークの推定であり，遺伝子ネットワークを特徴づける観測データはマイクロアレイによる遺伝子発現データだけではないと言うことに最初に着目したのはHartemink et al.（2002）であった．

Hartemink et al.（2002）は，マイクロアレイデータに加え，結合位置データを遺伝子ネットワーク推定に用いた．結合位置データとは，ある転写因子がターゲットの遺伝子を制御するか否かを実験的に観測したデータである．Hartemink et al.（2002）は結合位置データをベイジアンネットワークの構造推定における探索範囲の制約として用いることで，マイクロアレイデータへの過適合を避ける手法を提案した．しかしながら，Hartemink et al.（2002）の手法は結合

位置データという極めて特殊な形式で与えられる転写因子に特異的な情報に限っていること，結合位置データに含まれるノイズを考慮していないことが問題点としてあげれらる．そこで，Imoto et al.（2004a）は結合位置データはもとより，タンパク質間相互作用データ，結合配列データ，文献情報など様々な生物学的情報をマイクロアレイデータと共に遺伝子ネットワーク推定に利用できる枠組みを提案した．その本質は，生物学的情報を用いたネットワークの事前確率の構築にある．ネットワークの構造に対して informative な事前分布を構成することでマイクロアレイデータへの過適合を避けより精度の高い遺伝子ネットワークを推定することを目的としている．Bernard and Hartemink（2005）は，Imoto et al.（2004a）の枠組みに従って Hartemink et al.（2002）の手法を再構成したものである．

Imoto et al.（2004a）の枠組みを利用し特定の生物学的情報をネットワークの事前分布に利用した例として，我々の研究室では以下の研究を行った．Tamada et al.（2003）は制御配列の利用に着目し，制御配列探索手法とネットワーク推定手法を組み合わせた手法を提案した．この手法によって，転写因子が結合する制御配列を予測することができ，同時にその情報とマイクロアレイデータによりネットワークを推定することができる．本手法は，“クラスター分析⇨制御配列探索”というスタンダード化された解析ステップに対して“ネットワーク推定＋制御配列探索”という新たなパラダイムを提案した．Nariai et al.（2004）はタンパク質間相互作用データをマイクロアレイデータと共に遺伝子ネットワーク推定に利用する研究を行った．その特色は，タンパク質間相互作用データを単に利用するだけでなく，ネットワーク推定と同時にタンパク質複合体の予測を行うことにある．ベイジアンネットワークに基づく遺伝子ネットワーク推定に対して制御配列，タンパク質間相互作用を組み合わせた解析については Imoto et al.（2005）にもまとめられている．

また，Nariai et al.（2005）では，Imoto et al.（2004a）の枠組みをさらに発展させ，無向グラフにより構築されるタンパク質相互作用ネットワークと有向グラフにより構築される遺伝子ネットワークを同時に推定するという新しい問題を定義し，マルコフ確率場（Markov random field）とベイジアンネットワークを組み合わせたモデルを用いて一つの解を示した．Tamada et al.（2005）では，2種類の生物種間に共通に保存されている細胞周期など生命の維持に本質的なネットワークの情報を利用して，2つの生物種の遺伝子ネットワークを同時に推定する手法を提案し，出芽酵母とヒトの細胞周期に関するネットワーク推定に適用した．その結果，独立に推定した結果と比べより多く既知の関係を捉えていることを確認した．Imoto et al.（2006a）では，ネットワークの事前分布構築に使用された生物学的情報に含まれる誤りに着目し，その誤りを訂正できる自己組織型の遺伝子ネットワーク推定手法を提案した．Imoto et al.（2006b）では，Imoto et al.（2004a）が離散的な生物学的情報（例えば，既知 or 未知）を用いてネットワークの事前分布を構築していたのに対し，離散型，連続型が混在する複数の生物学的情報に基づく事前分布の構成法を提案し，薬剤によって影響を受けた遺伝子パスウェイの推定を行った．解析の詳細は次節で述べる．

5. 創薬ターゲット遺伝子のイン・シリコ探索

創薬に至るプロセスは大きく分けて薬剤のターゲットとなる遺伝子の同定と薬剤自身の設計に分けることができると考えられる．後者については計算化学を用いた既に分かっているターゲットに対して最適な化合物を設計するアプローチなど洗練された技術が既に確立されつつある．しかしながら，前者の薬剤のターゲットとなる遺伝子をゲノムワイドな情報からシステマティックに同定するための手法に関しては，その研究はまだ始まったばかりである．Imoto et al.（2003b）は，遺伝子ネットワークが薬剤ターゲット遺伝子の同定に対して本質的な情報を

与えることを指摘し，ネットワーク推定技術を用いた薬剤ターゲット遺伝子の同定法を提案した最初の論文である．Imoto et al.（2003b）は，出芽酵母に薬剤（griseofulvin：抗真菌薬）を投与したマイクロアレイデータと遺伝子破壊実験によるマイクロアレイデータを用い，薬剤ターゲット遺伝子の同定を2つのプロセスに分けることを提案し，そのための計算科学的手法を提示した．その最初のステップは，薬剤によって直接影響を受けた遺伝子（drug-affected genes）の同定である．この目的ために，薬剤に相当する仮想的な遺伝子をルートノードとする多層有向グラフを構成するための仮想遺伝子法を提案した．また，次のステップとしては，薬剤投与によって直接影響を受けた遺伝子の上流に（理想的には直接の親に）受容体のような薬剤のターゲットとなりうる遺伝子（druggable genes）があるかどうか探索する．このためには，遺伝子ネットワーク推定技術が必要不可欠である．また，Savoie et al.（2003）は仮想遺伝子法によりgriseofulvinのターゲットとして*CIK1*を同定した．さらに，*CIK1*を破壊することによってgriseofulvinを投与したときと同様の表現型を得ることを示した．

Imoto et al.（2003b）の示したdrug-affected genesからdruggable genesへ至るcomputational drug target gene discoveryにおいて，追加すべき重要な情報の一つは，薬剤によって影響を受ける遺伝子パスウェイ（drug-affected pathways）の同定である．もし，このdrug-affected pathwayに副作用に関する遺伝子が乗っているのならば，その影響を事前に知ることができ，副作用を避ける手段を探す手がかりとなりうる．そのための研究がTamada et al.（2005）によってなされた．これら3つの研究はImoto et al.（2006c）にもまとめられている．

Imoto et al.（2006b）ではヒト血管内皮細胞を用いて高脂血症薬であるfenofibrateの反応パスウェイを推定した．その中で，druggable genesをより発展させdruggable gene networkという概念を提案した．Druggable gene networkとは，ある特定の薬剤に対して反応性のあるパスウェイとして定義され，既知のターゲット遺伝子や新規のターゲット遺伝子候補を含むことからその名前が付けられた．図4（a）はfenofibrateに関与すると推定された1192遺伝子からなる推定されたネットワークの全体像を表している．このように極めて多くのノードからなる複雑グラフから有用な情報を抽出することは，それ自体が新たな研究課題となる．ネットワーク推定の結果に対してはいくつかの評価方法が提案されているが，その多くはデータベースに蓄積された知識との整合性からの評価である．遺伝子の機能に関連した評価方法としてはGene Ontologyコンソーシアムによるアノテーション（GO term）に基づくGO::TermFinderによる推定されたサブグラフに特徴的な機能を持つ遺伝子が有意に存在するか否かを判定するなどの方法が用いられる．Gene Ontologyに基づくGO termの評価法としてはAl-Shahrour et al.（2004），Gupta et al.（2007）なども参照されたい．

*PPAR-α*はfenofibrateのターゲット遺伝子であることが知られており，推定された反応パスウェイにおいては，*PPAR-α*は多数の遺伝子を制御しており，まさにそのパスウェイのトリガー的役割を担っていた．また，*PPAR-α*に関連していくつかの既知の情報と整合性のある関係が得られた．図4（b）は*PPAR-α*の下流ネットワークの一部である．Imoto et al.（2006b）の推定したネットワークは，多数の既知ターゲット遺伝子を含み，そのほとんどがネットワーク上で多数の遺伝子を制御している，いわゆるhub遺伝子となっていた．高脂血症に関連する脂質代謝遺伝子において，fenofibrateのターゲット遺伝子である*PPAR-α*よりも多くの遺伝子を制御していたものは17個あり，そのうち6個の遺伝子は既存薬のターゲット遺伝子であった．それらのうちいくつかを紹介すると，*HMGCR*は多くの製薬会社がターゲットとしており，三共の高脂血症治療剤であるHMG-CoA還元酵素阻害剤メバロチンのターゲット遺伝子でもある．他にも*LIPG*，*LSS*などが発見されるが，興味深いのはその中に*COX2*が含まれていることである．

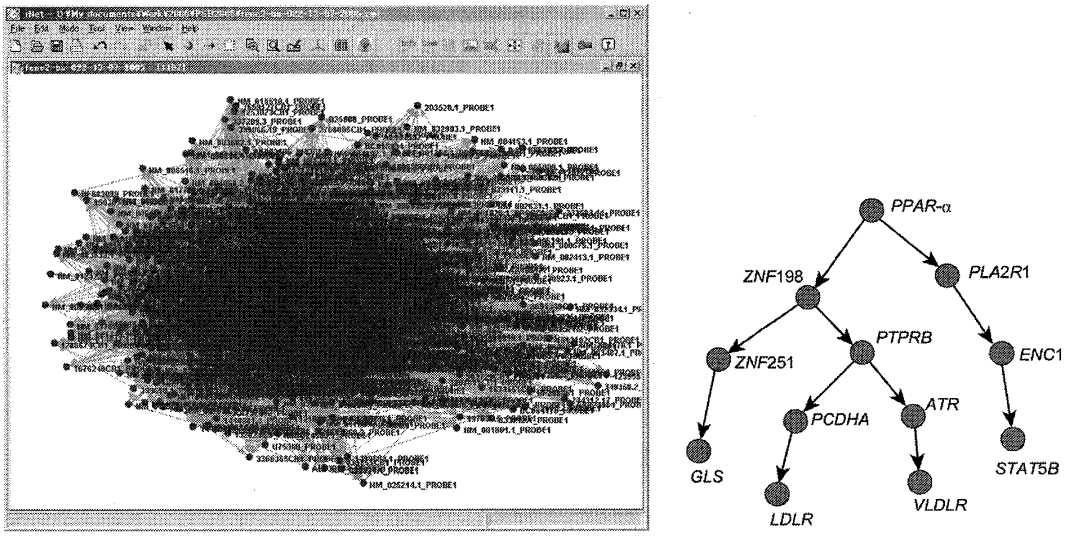

図4　(a) Fenofibrate によって影響を受けたと推定された 1192 遺伝子のネットワーク．(b) *PPAR-α* の下流パスウェイ．

COX2 は関節炎治療薬であり，ここでターゲットとしてる高脂血症との関連性は明らかではないが，多くの製薬会社がターゲットとしている遺伝子である．メルクのバイオックスも *COX2* 選択的阻害剤であるが，2005 年のバイオックスの副作用訴訟は記憶に新しい．これは，*COX2* を抑制したことによる心筋梗塞が副作用として現れた結果だと推測されている．推定したネットワークにおいて *COX2* の周りの情報を見てみると，*COX2* が直接制御していると推定された遺伝子の一つが JAK/STAT とよばれるパスウェイに上にある遺伝子であった．JAK/STAT パスウェイは心筋梗塞と関連があることが知られているため，推定したネットワークはその副作用メカニズムを解明するための重要な手がかりになる可能性がある．

6. 遺伝子ネットワーク推定のための諸手法

本稿では，ベイジアンネットワークに基づく遺伝子ネットワーク推定に焦点を絞って，我々の研究室でこれまでに行ってきた研究について紹介した．しかしながら，マイクロアレイデータに基づいて遺伝子ネットワークを推定するための手法はもちろんベイジアンネットワークだけではない．多種多様な手法がこれまでに提案され様々な成果を上げている．本節では部分的ではあるがそれらの中で興味深いものに関して紹介する．

ブーリアンネットワーク

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のための計算科学的手法として早くから注目を集めたのはブーリアンネットワーク（Boolean network）である．1994 年に Affymetrix 社が GeneChip の販売を開始してから 4 年，1996 年に Patrik Brown の研究室から初めて cDNA マイクロアレイにより観測された遺伝子発現データが公開されてから 2 年を経て，Akutsu et al. (1998)，Liang et al. (1998) はそれぞれ遺伝子破壊実験・強制発現実験によるマイクロアレイデータ，時系列に観測されたマイクロアレイデータを用いたブーリアンネットワークによる遺伝子ネットワーク推定について発表した．

ブーリアンネットワークによる遺伝子ネットワーク推定においては，マイクロアレイデータは発現が誘導されているか抑制されているかという 2 値に離散化され，ネットワーク上での親子関係はそれらの AND や OR といったブール関数により記述される．シンプルなモデルであ

る利点を生かし，ネットワーク推定における計算量の評価や，ネットワークを一意に決定するために必要となるマイクロアレイの枚数の下限など情報科学的理論面を中心に研究は進められた．詳しくは，Akutsu et al.（1999; 2000a, b）の一連の研究を参照されたい．また，決定論的モデルであるブーリアンネットワークを確率モデルとして拡張した確率的ブーリアンネットワーク（probabilistic Boolean network）の研究としては，Shmulevich et al.（2002a, b）がある．

微分方程式モデル

遺伝子発現データは遺伝子破壊実験のような実験後の定常状態において観測されるものだけではない．ダイナミックな遺伝子発現の挙動を捉えるために，何らかのショックを与えてからの発現の時系列変化を追跡する実験を行うことができる．時系列に観測されたマイクロアレイデータに対して，微分方程式（ordinary differential equation）を用いて遺伝子ネットワークを推定する試みとしてはChen et al.（1999），de Hoon et al.（2003）などがあげられる．これらの手法は線形モデルに基づく手法であるが，化学反応に関する一般化質量作用則の近似モデルであるS-system（Savageau, 1969）を利用した研究も行われている．詳しくは，岡本（2000），Kikuchi et al.（2003），Kimura et al.（2005）を参照されたい．

ダイナミックベイジアンネットワーク

ダイナミックベイジアンネットワーク（dynamic Bayesian network）は時系列データ解析のためのベイジアンネットワークの一つの拡張と捉えることができる．時系列マイクロアレイデータに対して，ある時刻のマイクロアレイデータは1時点前のマイクロアレイデータにのみ依存し，遺伝子間の依存関係は各時点間で不変と仮定したモデルは微分方程式に基づくモデルと等価であり，後述する状態空間モデルのスペシャルケースとなっている．詳しくは，Husmeier（2003），Kim et al.（2003, 2004）を参照されたい．

グラフィカル・ガウシアンモデル

遺伝子ネットワークはその情報の性質から有向グラフを用いて表現することが自然ではあるが，枝に向きの付いていない無向グラフによるモデリングも試みられている．Toh and Horimoto（2002）はグラフィカル・ガウシアンモデル（graphical Gaussian model）を用い遺伝子ネットワークの構築を試みた．しかしながら，グラフィカルガウシアンモデルの構築に関しては偏相関行列を求める必要があり，数千の遺伝子（確率変数）に対して数百枚のマイクロアレイデータ（サンプル）では相関行列の逆行列計算が不安定となる．そこで，Toh and Horimoto（2002）では，まず遺伝子をクラスター分析によって数十のクラスタに分割し，そのクラスター間のネットワークを推定している．また，遺伝子をクラスターにまとめずにグラフィカル・ガウシアンモデルを推定する手法としては正則化法を用いる方法が考えられる．Shimamura et al.（2007）では，lasso型の推定量を用いてグラフィカル・ガウシアンモデルにより遺伝子ネットワークを推定するための方法が提案されている．

状態空間モデル

近年，遺伝子発現は転写モジュール（transcriptional module）という単位で理解されることが多い（Segal et al., 2003）．ここで，転写モジュールとは，機能的に関連しかつ共発現する遺伝子の集合と定義される．このようなシステムは状態空間モデル（state space model）によって自然にモデリングできる．すなわち，状態空間モデルにおける観測データはある時刻における遺伝子の発現データが並んだベクトルであり，状態変数はその観測データが縮約された転写モジュールを表していると解釈できる．この性質に着目し，転写モジュール間のネットワーク推定をYamaguchi et al.（2006）は行った．また，Yoshida et al.（2005）は時刻によって構造の変化を許容した遺伝子ネットワーク推定法を状態空間モデルにマルコフスイッチングを組み合わせて構築した．

時系列マイクロアレイデータの大きな特徴の一つとして，その観測時点数の少なさがあげられる．Imoto et al.（2002）で用いた出芽酵母のマイクロアレイデータは短い4つの時系列マイクロアレイデータからなり，最も時点数の多いものでも24時点である．さらには，Imoto et al.（2006b）で用いたfenofibrate投与後の発現データは時系列に観測されているが，その時点数は6である．しかしながら，そのような短い時系列データに対しては，クオリティコントロールのため繰り返し計測を行う場合が多い．Hirose et al.（2006）はその繰り返し情報を利用し，極めて短い時系列マイクロアレイデータから遺伝子ネットワークを推定する方法を研究している．

謝　辞

査読者の方には丁寧に本稿を読んでいただき，多数の貴重なコメントを頂きました．感謝申し上げます．本稿で紹介した研究は，著者単独の研究成果ではなく，多くの共同研究者の方々とのコラボレーションによる成果です．ここに記して深く感謝いたします．

参 考 文 献

Affymetrix Inc.（2002）Statistical algorithms description document.（Affymetrix社のWebサイト上で公開）.

Affymetrix Inc.（2005）Technical note: guide to probe logarithmic intensity error（PLIER）estimation.（Affymetrix社のWebサイト上で公開）.

Akutsu, T., Kuhara, S., Maruyama, O. and Miyano, S.（1998）. A system for identifying genetic networks from gene expression patterns produced by gene disruptions and overexpressions, *Genome Informatics*, **9**, 151–160.

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, *Pac. Symp. Biocomput.*, **4**, 17–28.

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

Akutsu, T., Miyano, S. and Kuhara, S.（2000b）. Algorithms for identifying Boolean networks and related biological networks based on matrix multiplication and fingerprint function, *J. Comp. Biol.*, **7**, 331–344.

Al-Shahrour, F., Diaz-Uriarte, R. and Dopazo, J.（2004）. FatiGO: A web tool for finding significant associations of Gene Ontology terms with groups of genes, *Bioinformatics*, **20**, 578–580.

Bernard, A. and Hartemink, A. J.（2005）. Informative structure priors: joint learning of dynamic regulatory networks from multiple types of data, *Pac. Symp. Biocomput.*, **10**, 459–470.

Chen, T., He, H. L. and Church, G. M.（1999）. Modeling gene expression with differential equations, *Pac. Symp. Biocomput.*, **4**, 29–40.

De Boor, C.（1978）. *A Practical Guide to Splines*, Springer, Berlin.

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, *Pac. Symp. Biocomput.*, **8**, 17–28.

Eilers, P. and Marx, B.（1996）Flexible smoothing with *B*-splines and penalties（with discussion）, *Statistical Science*, **11**, 89–121.

Eisen, M. B., Spellman, P. T., Brown, P. O. and Botstein, D.（1998）. Cluster analysis and display of genome-wide expression patterns, *Proc. Natl. Acad. Sci. USA*, **95**, 14863–14868.

Friedman, N., Linial, M., Nachman, I. and Pe'er, D.（2000）. Using Bayesian networks to analyze expression data, *J. Comp. Biol.*, **7**, 601–620.

Furey, T. S., Cristianini, N., Duffy, N., Bednarski, D. W., Schummer, M. and Haussler, D.（2000）Support vector machine classification and validation of cancer tissue samples using microarray expression data, *Bioinformatics*, **16**, 906–914.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M.L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D. and Lander, E. S.（1999）. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, *Science*, **285**, 531–537.

Gupta, P. K., Yoshida, R., Imoto, S., Yamaguchi, R. and Miyano, S.（2007）. Statistical absolute evaluation of gene ontology terms with gene expression data. *Proc. 3rd International Symposium on Bioinformatics Research and Applications*, Lecture Note in Bioinformatics, Springer-Verlag, **4463**, 146–157.

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, *Pac. Symp. Biocomput.*, **7**, 437–449.

Hirose, O., Yoshida, R., Imoto, S., Yamaguchi, R., Higuchi, T. and Miyano, S. (2007). Construction of large gene networks from short time courses of gene expression profiles by state space models. *submitted for publication.*

Hubbell, E., Liu, W.-M. and Mei, R. (2002) Robust estimators for expression analysis, *Bioinformatics*, **18**, 1585–1592.

Husmeier, D. (2003). Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks, *Bioinformatics*, **19**, 2271–2282.

Imoto, S., Goto, T. and Miyano, S. (2002). Estimation of genetic networks and functional structures between genes by using Bayesian networks and nonparametric regression, *Pac. Symp. Biocomput.*, **7**, 175–186.

Imoto, S., Kim, S., Goto, T., Aburatani, S., Tashiro, K., Kuhara, S. and Miyano, S. (2003a). Bayesian network and nonparametric heteroscedastic regression for nonlinear modeling of genetic network, *J. Bioinform. Comput. Biol.*, **1**, 231–252.

Imoto, S. and Konishi, S. (2003). Selection of smoothing parameters in B-spline nonparametric regression models using information criteria. *Annals of the Institute of Statistical Mathematics*, **55**, 671–687.

Imoto, S., Savoie, C. J., Aburatani, S., Kim, S., Tashiro, K., Kuhara, S. and Miyano, S. (2003b). Use of gene networks for identifying and validating drug targets, *J. Bioinform. Comput. Biol.*, **1**, 459–474.

Imoto, S., Higuchi, T., Goto, T., Tashiro, K., Kuhara, S. and Miyano, S. (2004a). Combining microarrays and biological knowledge for estimating gene networks via Bayesian networks, *J. Bioinform. Comput. Biol.*, **2**, 77–98.

Imoto, S., Higuchi, T., Kim, S., Jeong, E. and Miyano, S. (2004b). Residual bootstrapping and median filtering for robust estimation of gene networks from microarray data, *Lecture Note in Bioinformatics*, **3082**, 149–160, Springer-Verlag.

Imoto, S., Matsuno, H. and Miyano, S. (2005). Gene networks: estimation, modeling and simulation. in R. Eils and A. Kriete (Eds.), *Computational Systems Biology*, Academic Press, 205–228.

Imoto, S., Higuchi, T., Goto, T. and Miyano, S. (2006a). Error tolerant model for incorporating biological knowledge with expression data in estimating gene networks, *Statistical Methodology*, **3**, 1–16.

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. (2006b). Computational strategy for discovering druggable gene networks from genome-wide RNA expression profiles, *Pac. Symp. Biocomput.*, **11**,559–571.

Imoto, S., Tamada, Y., Savoie, C. J. and Miyano, S. (2006c). Analysis of gene networks for drug target discovery and validation. in J. Walker and M. Sioud (Eds.), *Target Discovery and Validation*, Volume 1, pp. 33–56 (a volume of "Methods in Molecular Biology" series), Humana Press, USA.

井元清哉 (2007)「マイクロアレイデータによる遺伝子ネットワーク推定」. 樋口知之(編)『統計数理は隠された未来をあらわにする—ベイジアンモデリングによる実世界イノベーション』, pp. 85–117), 東京電機大学出版局.

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J. Scherf, U. and Speed, T. P. (2003) Exploration, normalization and summaries of high density oligonucleotide array probe level data, *Biostatistics*, **4**, 249–264.

狩野 裕, 宮村 理 (2006). 統計的因果推論と因果探索. 第1回データマイニングと統計数理研究, SIG-DMSM-A601.

Kikuchi, S., Tominaga, D., Arita, M., Takahashi, K. and Tomita, M. (2003). Dynamic modeling of genetic networks using genetic algorithm and S-system, *Bioinformatics*, **19**, 643–650.

Kim, S., Imoto, S. and Miyano, S. (2003). Inferring gene networks from time series microarray data using dynamic Bayesian networks, *Briefings in Bioinformatics*, **4**, 228–235.

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.

Kimura, S., Ide, K., Kashihara, A., Kano, M., Hatakeyama, M., Masui, R., Nakagawa, N., Yokoyama, S., Kuramitsu, S. and Konagaya, A. (2005). Inference of S-system models of genetic networks using a cooperative coevolutionary algorithm, *Bioinformatics*, **21**, 1154–1163.

Li, C. and Wong, W. H. (2001) Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection, *Proc. Natl Acad. Sci.* USA, **98**, 31–36.

Liang, S., Fuhrman, S. and Somogyi, R. (1998). REVEAL, a general reverse en-gineering algorithm for inference of genetic network architectures, *Pac. Symp. Biocomput.*, **3**, 18–29.

Nariai, N., Kim, S., Imoto, S. and Miyano, S. (2004). Using protein-protein interactions for refining gene networks

estimated from microarray data by Bayesian networks, *Pac. Symp. Biocomput.*, **9**, 336–347.

Nariai, N., Tamada, Y., Imoto, S. and Miyano, S. (2005). Estimating gene regulatory networks and protein-protein interactions of Saccharomyces cerevisiae from multiple genome-wide data, *Bioinformatics*, 21, Suppl.2, ii206–ii212.

岡本正宏 (2000).「S-system による遺伝子の相互作用推定」. 高木・富田(編)『ゲノム情報生物学』165–188.

Ott, S., Imoto, S. and Miyano, S. (2004). Finding optimal models for small gene networks, *Pac. Symp. Biocomput.*, **9**, 557–567.

Savageau, M. A. (1969). Biochemical systems analysis II: the steady state solution for an n-pool system using a power law approximation, *J. Theoret. Biol.*, **25**, 370–379.

Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D. and Friedman, N. (2003). Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data, *Nat. Genet.*, **34**, 166–176.

Shimamura, T., Imoto, S. Yamaguchi, R. and Miyano, S. (2007) Weighted lasso in graphical Gaussian modeling for large gene network estimation based on microarray data, Geneome Informatics, **18**(2), in press.

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

Shmulevich, I., Dougherty, E. R. and Zhang, W. (2002b) Gene perturbation and intervention in probabilistic boolean networks, *Bioinformatics*, **18**, 1319–1331.

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

Takenouchi, T., Ushijima, M. and Eguchi, S., (2007) GroupAdaBoost: accurate prediction and selection of important genes, *IPSJ Transactions on Bioinformatics*, **3**, 1–8.

Tamada, Y., Kim, S., Bannai, H., Imoto, S., Tashiro, K., Kuhara, S. and Miyano, S. (2003). Estimating gene networks from gene expression data by combining Bayesian network model with promoter element detection, *Bioinformatics*, **19**, Suppl. 2, ii227–ii236.

Tamada, Y., Imoto, S., Tashiro, K., Kuhara, S. and Miyano, S. (2005). Identifying drug active pathways from gene networks estimated by gene expression data, *Genome Informatics*, **16**(1), 182–191.

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E. S. and Gloub, T. R. (1999). Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation, *Proc. Natl. Acad. Sci. USA*, **96**, 2907–2912.

田辺國士, 田中輝雄 (1983)「ベイズモデルによる曲線・曲面の当てはめ」,『地球』, **5**, 179–186.

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.

Yamaguchi, R., Higuchi, T., Yoshida, R., Imoto, S. and Miyano, S. (2006). Finding module-based gene networks with state-space models — Mining hight-dimensional and short time-course gene expression data, *IEEE Signal Processing Magazine*, **24**(1), 37–46.

Yoshida, R., Imoto, S. and Higuchi, T. (2005). Estimating time-dependent gene networks from time series microarray data by dynamic linear models with Markov switching, *Proc. IEEE 4th Computational Systems Bioinformatics*, 289–298.

Yoshida, R., Numata, K., Imoto, S., Nagasaki, M., Doi, A., Ueno, K. and Miyano, S. (2006). A statistical framework for genome-wide discovery of biomarker splice variations with GeneChip Human Exon 1.0 ST Arrays, *Genome Informatics*, **17**(1), 88–99.

Yoshida, R., Numata, K., Imoto, S., Nagasaki, M., Doi, A., Ueno, K. and Miyano, S. (2007). Computational genome-wide discovery of aberrant splice variations with exon expression profiles, *Proc. IEEE 7th Internatinal Symposium on Bioinformatics and Bioengineering*, in press.