ベイジアンネットワークによる遺伝子ネットワークの推定

井元　清哉

1. マイクロアレイデータ

ヒトのDNAには遺伝子とよばれる領域が約30,000箇所存在する．この遺伝子とよばれる領域は，その塩基配列にしたがい様々なタンパク質を生成し，我々の体を形成している．各遺伝子は，独立にタンパク質を生成しているわけではなく，必要なときに必要なタンパク質を必要な量生成するためのプログラムを有している．このプログラムは，遺伝子間の相互作用によって形成されるネットワークという形で実現され，このネットワーク上での情報のやりとりに従い，各遺伝子はタンパク質を生成する．細胞内に存在するそのようなネットワークとしては，エネルギーを生成するための代謝パスウェイ，細胞外からのシグナルを細胞内に伝えるシグナル伝達パスウェイ，遺伝子間の転写制御から成る転写制御ネットワークなどさまざまなものが存在する．

各遺伝子がタンパク質を生成する過程は，大きくはDNA上の遺伝子コード領域からメッセンジャーRNAへの転写，メッセンジャーRNAからタンパク質への翻訳とよばれる2つのステップから成る．細胞内である遺伝子が活動しているか否かを知るためには，その遺伝子がタンパク質を生成しているか否かを調べればよい．しかしながら，タンパク質の量を直接計測するのは困難のため，タンパク質の前段階であるメッセンジャーRNAを遺伝子がどの程度生成しているかをゲノムワイドに計測するための機器がマイクロアレイである．本稿では，遺伝子がメッセンジャーRNAを生成していることを発現しているとよぶ．マイクロアレイを用いると，ヒトの場合だと約30,000個の遺伝子の発現状態を同時に計測することができる．

いもと　せいや
東京大学 医科学研究所ヒトゲノム解析センター
〒108-8639 港区白金台 4-6-1

このマイクロアレイデータを用いることで，様々な情報を得ることができる．例えば，様々な癌細胞からマイクロアレイにより遺伝子発現状態を計測することで，今まで病理診断に基づき決定していた癌の種類を遺伝子発現状態に基づき診断し，薬効や副作用，予後を予測できる可能性がある．このような研究に対しては，マイクロアレイデータのクラスタ解析などが用いられている．本稿で取り扱うのは，より基盤情報である遺伝子間の制御関係を遺伝子制御ネットワークとしてマイクロアレイデータにより推定する方法である．この問題は，現在バイオインフォマティクスにおいて集中的に研究が推進されており，生命をシステムとして理解することを目的としたシステムバイオロジーに必要不可欠な情報である．

2. 遺伝子発現制御

前述したように，各遺伝子は独立に働いているのではなく，相互作用を通して互いに情報を伝達するシステムを有している．遺伝子の中には転写因子とよばれる遺伝子があり，他の多くの遺伝子の発現を制御していることが知られている．具体的には，遺伝子コード領域の上流部分（プロモータ領域とよばれる）にはある特定の転写因子が結合する配列があり，転写因子がその部位に結合することで下流の遺伝子コード領域の転写が開始されメッセンジャーRNAが生成される．図1は遺伝子間の転写制御の模式図である．この図は，転写因子が活性因子として働く状況を説明したものであるが，逆に，結合することによりターゲット遺伝子の発現を止める，抑制因子として働くタンパク質も存在する．マイクロアレイでは，各遺伝子が生成するタンパク質の量は計測しないものの，その前段階であるメッセンジャーRNAの量をゲノムワイドに計測することができるため，マイクロアレイデータを用いて未知の遺伝子発現制御関係を発見することが期待される．

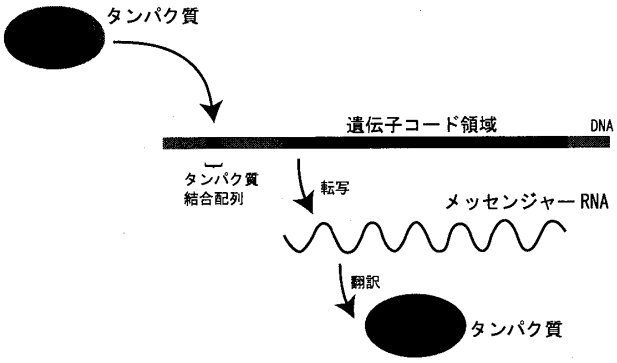

図1　遺伝子発現制御の模式図

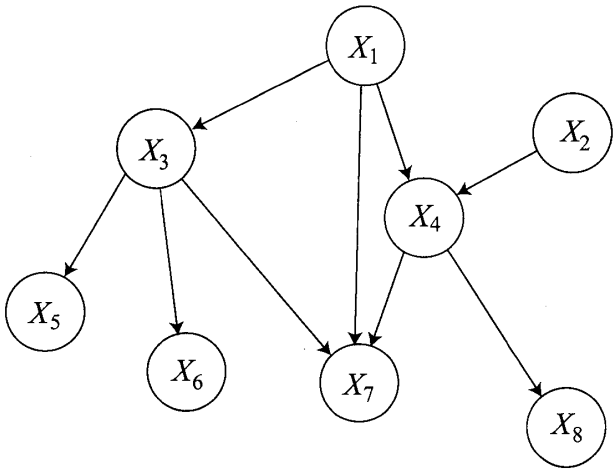

図2　非閉路有向グラフの例

3. ベイジアンネットワーク

ベイジアンネットワークは，多数の確率変数間の依存関係を記述するグラフィカル・モデルの1つである．確率変数 $\chi=\{X_1, \cdots, X_p\}$ を考える．確率変数間の依存関係は，各確率変数 X_j を1つのノードとする非閉路有向グラフ G によって与えられているとし，各ノード間にはマルコフ連鎖律を仮定する．非閉路有向グラフとは，矢印の向きにノードを辿ったとき，自分自身に戻ってくるパスのない有向グラフである．図2に非閉路有向グラフの例をあげる．マルコフ連鎖率により，各ノード（確率変数）は，直接の親ノードのみにその状態は依存し，直接の親確率変数を与えた元では非子孫の確率変数と独立となる．確率変数間に非閉路有向グラフによる依存関係とマルコフ連鎖律を仮定したことにより，確率変数の同時確率の分解を得る．

$$Pr(\chi)=\prod_{j=1}^{p} Pr(X_j|Pa(X_j)). \tag{1}$$

ただし，$Pa(X_j)$ は G 上での X_j の直接の親確率変数の集合である．例えば，図2においては $Pa(X_7)=\{X_1, X_3, X_4\}$ となる．

今，χ に従う n 個の実現値 $x_1, \cdots, x_n$ を得たとする．ここで，x_i の第 j 成分 x_{ij} は i 番目のマイクロアレイによって計測された j 番目の遺伝子からのメッセンジャーRNAの量に相当する．いま，ベイジアンネットワークの枠組みにおいて，X_j を j 番目の遺伝子が生成するメッセンジャーRNAの量を表す確率変数とするとき，X_j は連続型の確率変数となる．そこで，(1)式は密度関数を用いて

$$f(x_i|\Theta)=\prod_{j=1}^{p} f_j(x_{ij}|pa(X_j)_i, \theta_j) \tag{2}$$

と書き直すことができる．ここで，$pa(X_j)_i$ 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 G 上での直接の親遺伝子の発現データベクトルである．図2では，例えば，$pa(X_7)_i=(x_{i1}, x_{i3}, x_{i4})^t$ となる．ここで a^t はベクトル a の転置を表す．また，$\Theta=(\theta_1^t, \cdots, \theta_p^t)^t$ はパラメータである．(2)式から分かるように，非閉路有向グラフ G による密度関数の分解が与えられると次のタスクは各条件付き密度関数 f_j の統計的モデリングである．この問題は，G 所与のものでは，$Pa(X_j)$ を与えた下での X_j の条件付き分布の推定であり，いわゆる回帰モデルの推定問題に他ならない．いま，$pa(X_j)_i=(pa_{i1}^{(j)}, \cdots, pa_{iq_j}^{(j)})^t$ とすると，もっとも基本的なモデルは線形回帰モデル

$$x_{ij}=\beta+\beta_{j1}pa_{i1}^{(j)}+\cdots+\beta_{jq_j}pa_{iq_j}^{(j)}+\varepsilon_{ij}$$

であろう[5]．しかしながら，遺伝子間の関係は線形である保証はない．そこで，遺伝子間の線形性の仮定を緩め，ノンパラメトリック回帰に基づき遺伝子間の非線形関係を捉えるためのモデル

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

の利用が提案された[7][8]．

回帰関数 $m_{jk}(k=1, \cdots, q_j)$ の構成に対しては，フーリエ級数，スプライン，カーネル関数など様々な方法が考えられるが，ここでは B-スプラインに基づく基底関数展開法により m_{jk} を構成する．すなわち，$\{b_{k1}^{(j)}, \cdots, b_{kM_{jk}}^{(j)}\}$ をあらかじめ与えられた M_{jk} 個の B-スプラインとすると回帰関数は

$$m_{jk}(x)=\sum_{r=1}^{M_{jk}} \gamma_{kr}^{(j)} b_{kr}^{(j)}(x)$$

とかける．ここで，$\gamma_{kr}^{(j)}(r=1, \cdots, M_{jk})$ はパラメータである．B-スプラインの構成法，B-スプラインによるノンパラメトリック回帰モデルの詳細は，[3][4][9]などを参照されたい．

離散データに基づくベイジアンネットワーク，正規線形回帰モデルに基づくベイジアンネットワークをマイクロアレイデータから構築することも可能である．

しかしながら，これらのモデルには，次節のテーマであるネットワーク構造の推定に対してモデル識別性の問題がある．詳しくは，[1][13]を参照されたい．

4. ネットワーク構造推定

前節では，遺伝子間の依存関係を表す非閉路有向グラフ G が与えられた元での遺伝子ネットワークのモデリングについて述べた．しかしながら，遺伝子ネットワークの構造は大部分がいまだ決定されていない．そこで，データに基づいて遺伝子ネットワークの構造 G を推定する必要がある．つまり，ベイジアンネットワークの構造推定の問題である．統計学的に捉えると，どのネットワーク構造がより良いかを判定するモデル評価の問題とみなすことができる．この問題に対しては，情報量，ベイズなどのアプローチが研究されている．ここでは，マイクロアレイデータ D が与えられた元でのネットワーク構造 G の事後確率に基づきネットワーク構造を選択するベイズアプローチについて述べる．事後確率は $Pr(G|D)=Pr(G)p(D|G)/p(D)\propto Pr(G)p(D|G)$ と書き直すことができる．ここで，$Pr(G)$ はネットワーク構造に関する事前確率，$p(D|G)$ は周辺尤度であり，$p(D)$ は規格化定数である．離散型，正規線形回帰，ノンパラメトリック回帰それぞれに基づくベイジアンネットワークに対し，この事後確率に基づく遺伝子ネットワーク構造推定のための基準として，BDe[2]，BIC[18]，BNRC[7]が提案されている．

ベイジアンネットワークの構造推定のためのスコア関数最適化は，NP-困難であることが知られている．そこで，発見的なアルゴリズムによりネットワーク構造の推定を行うのが標準的である．発見的な方法としては，Greedy アルゴリズム[5][7]，焼きなまし法[6]などが用いられる．推定されるネットワーク構造の最適性が保証されるアルゴリズムとして，[16]は離散型ベイジアンネットワークの構造推定を MDL に基づいて行う際に，Branch and Bound アルゴリズムを利用した方法を提案している．より一般のモデルについては，スコア関数の各ノードに関する局所スコアへの分解条件：$s(G)=\sum_{j=1}^{p}s_j$ のもとでノード数が 30 程度のネットワーク構造を最適化するアルゴリズムを[15]が提案している．参考までに，ノード数が 30 個の非閉路有向グラフは約 2×10^{158} 個存在し，ナイーブに全探索はできない．

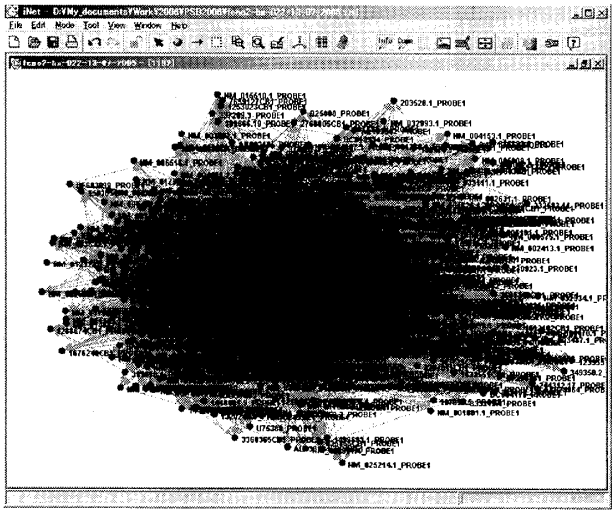

図3 推定された fenofibrate 関連遺伝子ネットワーク

5. 他の生物学的情報の利用

前節までに説明した手法は，マイクロアレイデータのみを用いた遺伝子ネットワーク推定のための手法である．しかしながら，図1からも分かるように，遺伝子ネットワークの表す転写制御にはタンパク質が介在する．また，図1にあるタンパク質結合配列の情報は転写制御に関して極めて重要な情報である．そこで，マイクロアレイデータと他の生物学的情報を併用することで，遺伝子ネットワークをより高精度に推定するための方法の研究が精力的になされている．詳しくは，[6][10][14][17]などを参照されたい．

6. 事例紹介

ヒト血管内皮細胞において高脂血症薬 fenofibrate に影響を受ける遺伝子ネットワーク推定[11]の例をあげる．[11]では，ヒト血管内皮細胞に対して fenofibrate を暴露し計測した時系列マイクロアレイデータと，siRNA による遺伝子ノックダウンマイクロアレイデータを組み合わせ，fenofibrate によって影響を受ける遺伝子ネットワークの同定を行った．その方法はまず，(手順1) 薬剤 (fenofibrate) 応答の時系列マイクロアレイデータから fenofibrate によって影響を受けると予測される遺伝子セットを同定し，(手順2) その遺伝子間のネットワークをノックダウンマイクロアレイによって推定するというものである．手順1により，1192 遺伝子を fenofibrate 関連遺伝子として同定し，推定した遺伝子ネットワークは図3である．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.

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3の推定されたネットワークは，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 *PPRA-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．例えば，*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．このようなある特定の薬剤に影響を受ける遺伝子ネットワークを推定・解析することにより，(1)新たな薬剤標的遺伝子，(2)副作用や薬効のメカニズムの解明，(3)より効果的な化合物の組み合わせに関して重要な情報が得られると考えられている．

参考文献

[1] D. M. Chickering: Learning equivalence classes of Bayesian-network structures. *Journal of Machine Learning Research*, 2 (2002), 445-498.

[2] G. Cooper and E. Herskovits: A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9 (1992), 309-347.

[3] C. De Boor: *A Practical Guide to Splines*. Springer-Verlag Berlin, 1978.

[4] P. H. C. Eilers and B. Marx: Flexible smoothing with *B*-splines and penalties (with discussion). *Statistical Science*, 11 (1996), 89-121.

[5] N. Friedman, M. Linial, I. Nachman and D. Pe'er: Using Bayesian network to analyze expression data. *J. Comp. Biol.*, 7 (2000), 601-620.

[6] A. J. Hartemink, D. K. Gifford, T. S. Jaakkola and R. A. Young: Combining location and expression data for principled discovery of genetic regulatory network models. *Pac. Symp. Biocomput.*, 7 (2002), 437-449.

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

[8] S. Imoto, S. Kim, T. Goto, S. Aburatani, K. Tashiro, S. Kuhara and S. Miyano: Bayesian network and nonparametric heteroscedastic regression for non-linear modeling of genetic network. *J. Bioinform. Comp. Biol.*, 1 (2003), 231-252.

[9] S. Imoto and S. Konishi: Selection of smoothing parameters in B-spline nonparametric regression models using information criteria. *Ann. Inst. Stat. Math.*, 55 (2003), 671-687.

[10] S. Imoto, T. Higuchi, T. Goto, K. Tashiro, S. Kuhara and S. Miyano: Combining microarrays and biological knowledge for estimating gene networks via Bayesian networks. *J. Bioinform. Comp. Biol.*, 2 (2004), 77-98.

[11] S. Imoto, Y. Tamada, H. Araki, K. Yasuda, C. G. Print, S. D. Charnock-Jones, D. Sanders, C. J. Savoie, K. Tashiro, S. Kuhara and S. Miyano: Computational strategy for discovering druggable gene networks from genome-wide RNA expression profiles. *Pac. Symp. Biocomput.*, 11 (2006), 559-571.

[12] F. V. Jensen: *An Introduction to Bayesian Networks*. University College London Press, 1996.

[13] 狩野裕, 宮村理: 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, (2006).

[14] N. Nariai, Y. Tamada, S. Imoto and S. Miyano: Estimating gene regulatory networks and protein-protein interactions of Saccharomyces cerevisiae from multiple genome-wide data. *Bioinformatics*, 21 (2005), ii 206-ii 212.

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

[16] J. Suzuki: Learning Bayesian belief networks based on the MDL principle: An efficient algorithm using the branch and bound technique. *IEICE Trans. Information and Systems*, E 81-D (1998), 356-367.

[17] Y. Tamada, S. Kim, H. Bannai, S. Imoto, K. Tashiro, S. Kuhara and S. Miyano: Estimating gene networks from gene expression data by combining Bayesian network model with promoter element detection. *Bioinformatics*, 19 (2003), ii 227-ii 236.

[18] G. Schwarz: Estimating the dimension of a model. *Annals of Statistics*, 6 (1978), 461-464.