マイクロアレイデータ解析における統計的方法論の開発

井元 清哉

東京大学医科学研究所ヒトゲノム解析センター

要旨

マイクロアレイ技術の発展に伴い，数千から 1 万以上の遺伝子の発現状態を同時に観測することが可能となり，生命現象を理解するための網羅的な研究がバイオインフォマティクスと呼ばれる分野において集中的に行われている．本稿では，マイクロアレイデータ解析のための統計的方法論に注目し，データの正規化，遺伝子・疾患のクラスタリング，遺伝子間の制御関係を表す遺伝子ネットワークの推定に関して最新の研究成果を交え論じる．

1. 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，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

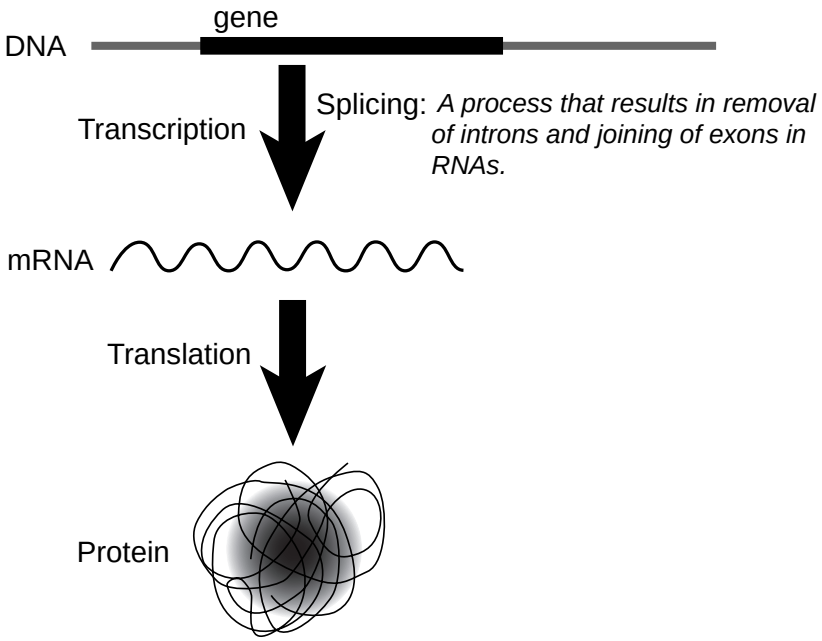


図 1: DNA からタンパク質への変換

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

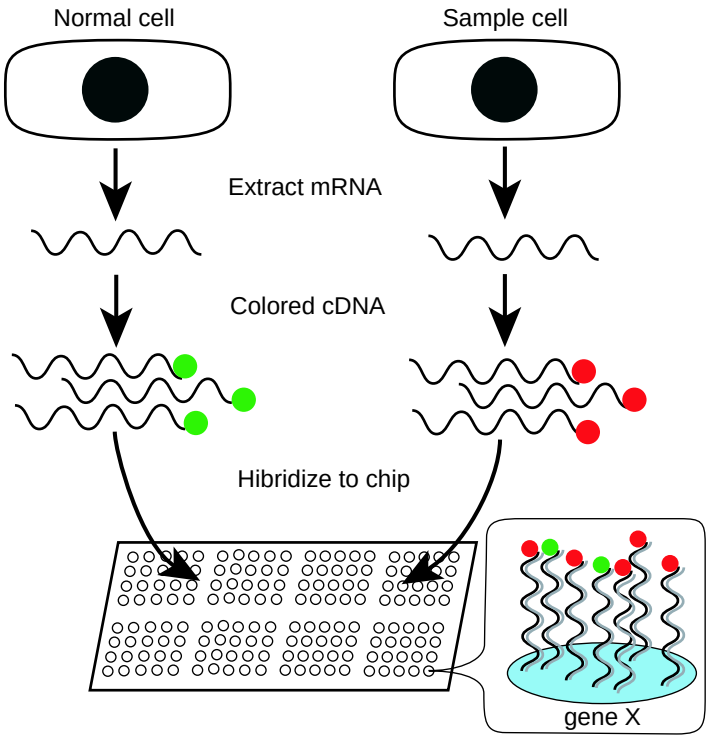
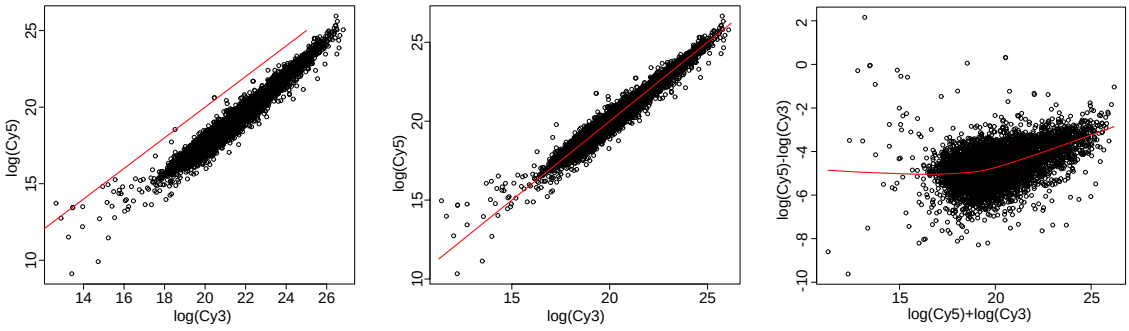


図 2: cDNA マイクロアレイ

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 R_i/G_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. 正規化

ほとんどの場合，観測されたマイクロアレイデータには正規化を施す必要がある．その理由として，正常細胞とサンプル細胞におけるもともとの mRNA の不均一性，Cy3, Cy5 の蛍光色素が本来持っている色の特性の違いなどが挙げられる．特に蛍光色素の特性の違いは顕著に現れ，例えば，Cy5 は Cy3 に比べて劣化が早い．つまり，マイクロアレイを作成してからスキャンするまでに時間がかかるとその分 Cy5 は劣化しインテンシティは小さくなってしまう．したがって，そのような場合，図 3(a) のように Cy3 と Cy5 のインテンシティには系統的な偏りが現れる．ここで，図 3(a) の赤線は $y = x$



(a) (b) (c)

図 3: cDNA マイクロアレイデータ．(a) 正規化前 (赤線: $y = x$)．(b) 総インテンシティ正規化後 (赤線: $y = x$)．(c) 正規化前の M-A plot と当てはめた lowess 曲線 (赤線)．

を表す．正常細胞とサンプル細胞の違いの程度にも依るが，ほとんどの遺伝子は両細胞間で同程度の発現レベルであり，一部の遺伝子が特異的に変化していると生物学的に考えられる．したがって，このような系統的な偏りには上に挙げたような原因が考えられる．

マイクロアレイデータの正規化は大きくは，大域的正規化と局所的正規化の 2 種に分けられる．大域的正規化に関しては総インテンシティ正規化や Lowess 正規化が代表的である．

総インテンシティ正規化: 今，2 つの細胞間でもともとの mRNA の総量はほとんど同じであり，遺伝子は偏り無く選択されマイクロアレイ上に割り付けられているとする．このとき，ハイブリダイゼーション後の Cy3, Cy5 それぞれの総インテンシティは同じであることが期待される．したがって，N をマイクロアレイ上に割り付けられた遺伝子数としたとき，$T_{total} = \sum_{i=1}^{N} R_i / \sum_{i=1}^{N} G_i$ を用いて $G'_i = T_{total} G_i$, $R'_i = R_i$ と補正する．図 3(a) で表されたマイクロアレイデータに対して総インテンシティ正規化を行った結果，図 3(b) が得られる．

Lowess 正規化: 総インテンシティ正規化は Cy3, Cy5 のインテンシティに一律に補正を施すというものであるが，インテンシティの大きさに依存した偏りの存在も報告されている．つまり，インテンシティの大きさによって偏りの程度が変わることがある．そのような場合に対して，ノンパラメトリック回帰の一手法である lowess を利用した正規化の方法が提案されている．図 3(a) を時計回りに 45 度回転させたものは M-A plot (または，R-I plot) と呼ばれ，M-A plot に対して lowess 曲線を当てはめ補正する．図 3(c) は図 3(a) の M-A plot に対して lowess 曲線を当てはめた例である．M-A plot では，横軸は Cy3, Cy5 の対数インテンシティの和，縦軸は対数インテンシティの差であり，$\log_2(\text{Cy5}) - \log_2(\text{Cy3}) = \log_2(\text{Cy5}/\text{Cy3})$ であるのでインテンシティの大きさに応じて対数比を補正していることに対応する．

局所的正規化: 蛍光されたcDNAは，384プレートから針先にごく少量つけられマイクロアレイ上へスポットされる．したがって，針の状態の差は，発現量の系統的な差を生じる．同じ針によってスポットされた遺伝子グループはペングループと呼ばれ，ペングループ間で系統的な差が無くなるように局所的正規化は多くの場合施される．このとき，各ペングループには十分な数の遺伝子が属し，マイクロアレイ上に遺伝子はランダムに割り付けられているという仮定が必要となる．

図3から分かるように，ほとんどのマイクロアレイデータに対して等分散性は成り立たない．したがって，正規化後の解析においては，この不等分散性を考慮に入れたモデリングが必要となる．マイクロアレイの正規化に関しては，Chen *et al.* (1997), Tseng *et al.* (2001), 大谷ら (2002), Quackenbush (2002), Chen *et al.* (2003) を参照されたい．

3. クラスタリング

マイクロアレイデータの解析において，現在最も使用されている統計的手法はクラスタリングであると思われる．パン酵母 (*Saccharomyces cerevisiae*) は約6000個の遺伝子から成り，今，我々はそれらの遺伝子に関してN枚のマイクロアレイを観測したとする．このとき，マイクロアレイデータは$6000 \times N$次の行列として与えられ，(i, j)成分はi番目の遺伝子のj番目のマイクロアレイによって観測された発現レベルに対応する．また，人 (*Human*) のマイクロアレイでは，約30000個の遺伝子に関してマイクロアレイデータは観測される．したがって，クラスタリングにおいては遺伝子をクラスタリングするのか，それともマイクロアレイをクラスタリングするのかによって問題の困難さが大きく異なる．例えば，パン酵母6000遺伝子を遺伝子破壊実験によって得られた120枚のマイクロアレイに基づき階層型クラスタリングを行った例を図4に挙げる．このとき，解析の目的としては，パン酵母の約6000遺伝子中2399個の遺伝子に関してはいまだその機能がわかっていない (2003年4月現在．MIPSデータベースより) ことから，それらの機能を予測することが考えられる．各遺伝子は120次元の特徴を表すベクトルを持っており，その特徴ベクトルに基づいてクラスタリングは実行される．その際，6000遺伝子，120次元特徴ベクトルという極めて次元の高いデータを取り扱うため，推定されたグループの安定性・信頼性などに疑問が残る．

次に，マイクロアレイの分類を考える．具体的には癌の分類問題について考察する．つまり，マイクロアレイは癌患者に対応し，その患者の遺伝子発現パターンから癌の種類を分類する問題である．従来，癌の分類は数人の専門家による癌細胞の組織に基く診断により行われていた．しかし，近年，組織からでは判断するのが困難な癌のサブクラスの同定が注目を集めている．その背景として，組織からは同じ種類に分類される癌であっても，ある人には抗癌剤が劇的に効き，また別の人は抗癌剤が効かなかったり副作用などにより亡くなってしまうというような事例が数多く報告されていることが挙げられる．つまり，癌のサブクラスの同定はこのような個人差を考慮に入れたオーダーメイド医療を実現させる可能性を持っている．現在，マイクロアレイデータ

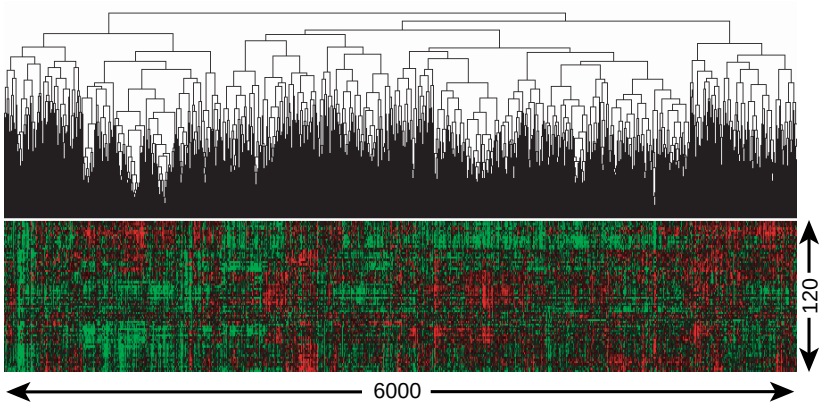


図 4: パン酵母遺伝子のクラスタリング例

は，このようなサブクラスの同定問題に利用されつつあり，大きな注目を集めている．

マイクロアレイデータに関するクラスタリングについては，階層型クラスタリング，k-means 法，自己組織化マップ，混合分布モデルなどの利用が研究されており，詳しくは，Eisen *et al.* (1998), Tamayo *et al.* (1999), Golub *et al.* (1999), McLachlan *et al.* (2002), De Hoon *et al.* (2002), Nagayama *et al.* (2002) を参照されたい．また，マイクロアレイデータに特化したクラスタリングソフトウェアも数多く開発されており，Eisen *et al.* による “Cluster and TreeView”, Tamayo *et al.* による “GENECLUSTER”, De Hoon *et al.* による “C clustering library” などが挙げられる．

4. 遺伝子ネットワーク

マイクロアレイデータに基づく遺伝子ネットワークの推定問題は，現在バイオインフォマティクスにおいて最も精力的に研究が進められている分野の 1 つである．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としては，ブーリアンネットワーク，ベイジアンネットワーク，常微分方程式モデル，グラフィカル・ガウシアンモデルなどが提案されている．図 5 は推定された遺伝子ネットワークの例である．遺伝子間の制御関係は有向グラフとして表される．

ブーリ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ためには，マイクロアレイデータを 2 値化する必要がある．2 値化は，ノイズの大きかった初期の cDNA マイクロアレイデータに対してはノイズのフィルタリングの役割も担い有効であったが，閾値の決定法や近年の精度の上がったマイクロアレイデータに対しては情報の過剰な損失が問題となる．また，ブーリアンネットワークによって遺伝子のネットワークを推定するためには多量のマイクロアレイが必要となることも問題として挙げられる．

ベイジアンネットワークは，現在，遺伝子ネットワーク推定問題に対して広く用いられている手法である．離散データに対するベイジアンネットワークモデルに加え，ノン

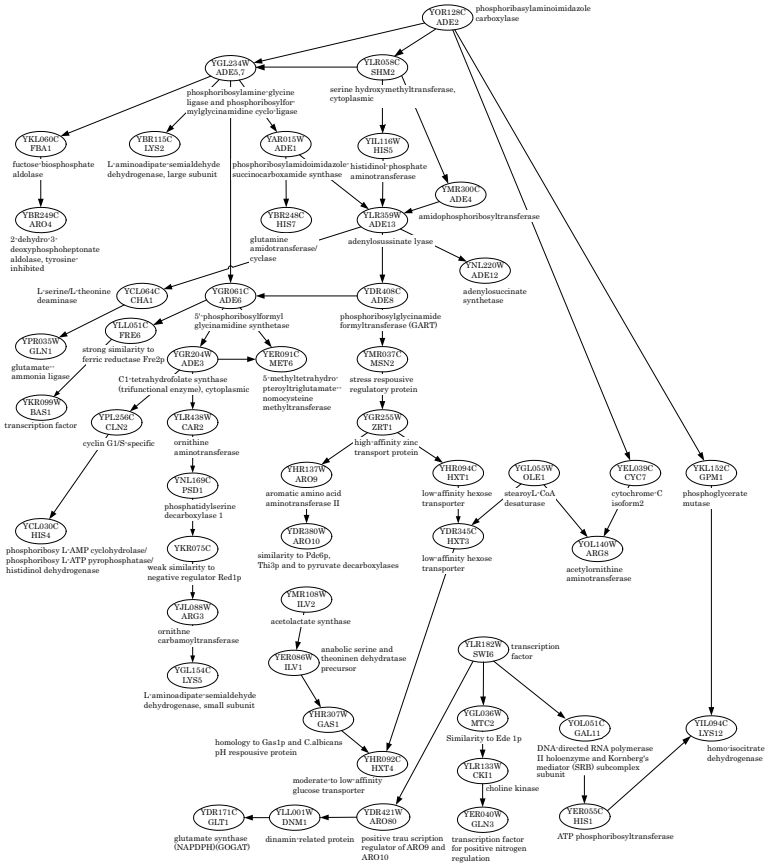


図 5: 遺伝子ネットワーク (Imoto *et al.* (2002b) より転載)

パラメトリック回帰モデルを組み合わせたベイジアンネットワークも提案され，いくつかの成果を挙げている．しかし，ベイジアンネットワークは非周期のネットワークを前提としており，実際の遺伝子ネットワークには多数の周期的制御関係があることからその点が実用上問題となる．その解決策として，時系列マイクロアレイデータが得られた際にはダイナミック・ベイジアンネットワークモデルの利用が提案されている．また，常微分方程式モデルも時系列マイクロアレイデータを前提とした手法である．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については，(ブーリアンネットワーク) Akutsu *et al.* (1998, 1999, 2000a, 2000b), Liang *et al.* (1998), Shmulevich *et al.* (2002). (ベイジアンネットワーク) Friedman & Goldszmidt (1996, 1998), Hartemink *et al.* (2001), Imoto *et al.* (2002a, 2002b, 2003a), Kim *et al.* (2003). (常微分方程式モデル) Chen *et al.* (1999), De Hoon *et al.* (2003). (グラフィカル・ガウシアンモデル) Toh & Horimoto (2002) などを参照されたい．また，推定された遺伝子ネットワークを利用した現実問題への応用として，薬剤ターゲット遺伝子の同定が挙げられる．詳しくは，宮野・Savoie (2002)，Savoie *et al.* (2003), Imoto *et al.* (2003b) を参照されたい．

マイクロアレイデータからの遺伝子ネットワーク推定に対しては，マイクロアレイデータの持つ情報量の不足が常に問題となる．つまり，パン酵母 6000 遺伝子のネットワークを数 100 枚のマイクロアレイデータから推定するには根本的に情報が不足していると言わざるを得ない．通常は，解析ターゲット遺伝子をある機能 (例えば細胞周期など) を持つものに限定し，ネットワークに含まれる遺伝子数を減らすなどの処理が事前に行われる．このような情報不足に対する解決策の 1 つとして，マイクロアレイデータに他の生物学的データを付加情報として加えた解析が大きな注目を集め，遺伝子ネットワークの推定に関しても研究されている．組み合わせる生物学的データとしては，タンパク質間相互作用 (protein-protein interaction)，タンパク質–DNA 相互作用 (protein-DNA interaction)，プロモーター領域に含まれる共通配列 (consensus motif)，文献情報などが用いられている．マイクロアレイデータと他の生物学的データとを合わせた解析としては，Bannai *et al.* (2002), Bussemaker *et al.* (2001), Ideker *et al.* (2002), Jenssen *et al.* (2001), Lee *et al.* (2002), Masys *et al.* (2001), Pilpel *et al.* (2001), Segal *et al.* (2002) などが挙げられ，特に遺伝子ネットワークの推定としては，Hartemink *et al.* (2002), Imoto *et al.* (2003c), Tamada *et al.* (2003) により研究されている．

参考文献

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

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

Akutsu, T., Miyano, S. & 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.

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

Bussemaker, H.J., Li, H. & Siggia, E.D. (2001). Regulatory element detection using correlation with expression. *Nature Genetics*, **27**, 167-171.

Chen, T., He, H.L. & Church, G.M. (1999). Modeling gene expression with differential equations. *Proc. Pac. Symposium on Biocomputing*, **4**, 29-40.

Chen, Y., Dougherty, E.R. & Bittner, M.L. (1997). Ratio-based decisions and the quantitative analysis of cDNA microarray images. *J. Biomed. Optics.*, **2**, 364-374.

Chen, Y.J., Kodell, R., Sistare, F., Thompson, K.L., Morris, S. & Chen, J.J. (2003). Normalization methods for analysis of microarray gene-expression data. *J. Biopharm. Stat.*, **13**(1), 57-74.

De Hoon, M.J.L., Imoto, S. & Miyano, S. (2002). Statistical analysis of a small set of time-ordered gene expression data using linear splines. *Bioinformatics*, **18**, 1477-1485.

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

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

Friedman, N. & Goldszmidt, M. (1996). Learning Bayesian networks with local structure. *Proc. 12th Conf. on Uncertainty in Artificial Intelligence*, 252-262.

Friedman, N. & Goldszmidt, M. (1998). Learning Bayesian networks with local structure. in M.I. Jordan *ed.*, Kluwer Academic Publisher.

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

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. & Lander, E.S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **285**, 531-537.

Hartemink, A.J., Gifford, D.K., Jaakkola, T.S. & Young, R.A. (2001). Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. *Proc. Pacific Symposium on Biocomputing*, **6**, 422-433.

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

Ideker, T., Ozier, O., Schwikowski, B. & Siegel, A.F. (2002). Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, **18**, Suppl.1 (ISMB 2002), 233-240.

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

Imoto, S., Kim, S., Goto, T., Aburatani, S., Tashiro, K., Kuhara, S. & Miyano, S. (2002b). Bayesian network and nonparametric heteroscedastic regression for nonlinear modeling of genetic network. *Proc. IEEE Computer Society Bioinformatics Conference*, 219-227.

Imoto, S., Kim, S., Goto, T., Aburatani, S., Tashiro, K., Kuhara, S. & Miyano, S. (2003a). Bayesian network and nonparametric heteroscedastic regression for nonlinear modeling of genetic network. *Journal of Bioinformatics and Computational Biology*, **1**, in press.

Imoto, S., Savoie, C.J., Aburatani, S., Kim, S., Tashiro, K., Kuhara, S. & Miyano, S. (2003b). Use of gene networks for identifying and validating drug targets. *Journal of Bioinformatics and Computational Biology*, **1**, in press.

Imoto, S., Higuchi, T., Goto, T., Tashiro, K., Kuhara, S. & Miyano, S. (2003c). Combining microarrays and biological knowledge for estimating gene networks via Bayesian networks. *Proc. IEEE Computer Society Bioinformatics Conference*, in press.

Jenssen, T.-K., Lægreid, A. Komorowski, J. & Hovig, E. (2001). A literature network of human genes for high-throughput analysis of gene expression. *Nature Genetics*, **28**, 21-28.

Kim, S., Imoto, S. & Miyano, S. (2003). Dynamic Bayesian network and nonparametric regression for nonlinear modeling of gene networks from time series gene expression data. *Proc. Computational Methods in Systems Biology*, Lecture Note in Computer Science, Springer-Verlag. 104-113.

Lee, T.I., Rinaldi, N.J., Robert, F., Odom, D.T., Bar-Joseph, Z., Gerber, G.K., Hannett, N.M., Harbison, C.T., Thompson, C.M., Simon, I., Zeitlinger, J., Jennings, E.G., Murray, H.L., Gordon, D.B., Ren, B., Wyrick, J.J., Tagne, J-B., Volkert, T.L., Fraenkel, E., Gifford, D.K. & Young, R.A. (2002). Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science*, **298**, 799-804.

Liang, S., Fuhrman, S. & Somogyi, R. (1998). REVEAL, a general reverse en-gineering algorithm for inference of genetic network architectures. *Proc. Pac. Symposium on Biocomputing*, **3**, 18-29.

Masys, D.R. (2001). Linking microarray data to the literature. *Nature Genetics*, **28**, 9-10.

McLachlan, G.J., Bean, R.W. & Peel, D. (2002). A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics*, **18**, 413-422.

宮野 悟・Savoie, C.J. (2002). バイオインフォマティクスの創薬応用. 実験医学. 12 月号. 2632-2637.

Nagayama, S., Katagiri, T., Tsunoda, T., Hosaka, T., Nakashima, Y., Araki, N., Kusuzaki, K., Nakayama, T., Tsuboyama, T., Nakamura, T., Imamura, M., Nakamura, Y. & Toguchida, J.

(2002). Genome-wide analysis of gene expression in synovial sarcomas using a cDNA microarray. *Cancer Res.*, **62**(20), 5859-5866.

大谷 敬子, 大瀧 慈, 佐藤 健一, 西山 正彦. (2002). cDNA マイクロアレイデータに基づく遺伝子発現状況の解析. 2002 年度統計関連学会連合大会講演報告集, 68-69.

Pe'er, D., Regev, A., Elidan, G. & Friedman, N. (2001). Inferring subnetworks from perturbed expression profiles. *Bioinformatics*, **17**, Suppl.1 (ISMB2001), 215-224.

Pilpel, Y., Sudarsanam, P. & Church, G.M. (2001). Identifying regulatory networks by combinatorial analysis of promoter elements. *Nature Genetics*, **29**, 153-159.

Quackenbush, J. (2002). Microarray data normalization and transformation. *Nature Genetics*, **32**, 496-501.

Savoie, C.J. Aburatani, S. Watanabe, S. Eguchi, Y. Muta, S. Miyano, S., Imoto, S., Kuhara, S. & Tashiro, K. (2003). Use of gene networks from full genome microarray libraries to identify functionally relevant drug-affected genes and gene regulation cascades. *DNA Research*, **10**, 19-25.

Segal, E., Barash, Y., Simon, I., Friedman, N. & Koller, D. (2002). From promoter sequence to expression: a probabilistic framework. *Proc. The Sixth Annual International Conference on Research in Computational Molecular Biology* (RECOMB 2002), 263-272.

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

Tamada, Y., Kim, S., Bannai, H., Imoto, S., Tashiro, K., Kuhara, S. & Miyano, S. (2003). Estimating gene networks from gene expression data by combining Bayesian network model with promoter element detection. *Bioinformatics*, **19**, (ECCB 2003), in press.

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander,E.S. & 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.

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

Tseng, G.C., Oh, M.K., Rohlin, L., Liao, J.C. & Wong, W.H. (2001). Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Res.*, **29**, 2549-2557.

ソフトウェア

Cluster and TreeView
`http://rana.lbl.gov/EisenSoftware.htm`

C clustering library
`http://bonsai.ims.u-tokyo.ac.jp/~mdehoon/software/cluster/index.html`

GENECLUSTER
`http://www-genome.wi.mit.edu/cancer/software/software.html`

著者連絡先

〒 108-8639 東京都港区白金台 4-6-1
東京大学 医科学研究所 ヒトゲノム解析センター DNA 情報解析分野
井元 清哉
Tel: 03-5449-5615　　Fax: 03-5449-5442
E-mail: imoto@ims.u-tokyo.ac.jp