日本統計学会誌
第38巻，第2号，2009年3月
237頁～250頁

遺伝子発現状態 'On/Off' 仮説に基づいたマイクロアレイデータ解析

大瀧　慈*，大谷敬子*，檜山桂子**，佐藤健一*，檜山英三***

Microarraydata analysis based on 'on/off' hypothesis

Megu Ohtaki*, Keiko Otani*, Keiko Hiyama**, Kenichi Satoh* and Eiso Hiyama***

数千から数万種といった規模の遺伝子発現を同時に観察することができるマイクロアレイは，遺伝子あるいは遺伝子間の相互作用と細胞レベルの機能を結びつける知見が得られるという点で生物医学領域に大きな期待を与えてきたが，研究成果の再現性に問題があることも次第に明らかになりつつある．マイクロアレイデータ解析が確実な研究基盤をもった研究分野に成熟するためには，マイクロアレイデータが得られるまでの複雑な背景を考慮したノイズの調整が必須である．本稿では，マイクロアレイデータ解析において遺伝子の発現・非発現（On/Off）を導入することにより，Affymetrix社GeneChipのプローブレベルのデータの数理構造を明らかにし，遺伝子発現の特徴量を'遺伝子発現強度の大きさ'を表す指標と'遺伝子がOnである確率'の2つの指標で表すことを提案した．また，それらの指標を用いて神経芽細胞腫の予後に関連する遺伝子の探索を行い提案した指標の有効性について検証した．

Microarray technology is now increasingly being used in versatile medical and biological fields as a high-throughput method for measuring the expression levels of thousand of genes simultaneously. It is often applied for disease diagnosis, identifying biomarkers, and studying function of genes. On the other hand, microarray analyses have a critical problem that they produce differing results with the same data. To establish a certain algorithm for adjustment various kinds of noise is a key point for microarray data analyses obtaining reliable results. We gave a mathematical description to Affymetrix GeneChip probe level data based on 'On/Off' hypothesis and proposed to express a feature of it with a pair of novel indicators; one expresses a signal intensity, the other expresses a probability of gene being 'On/Off'. The validity of the proposed method was shown by identifying 5 genes relating to the outcome of neuroblastoma.

Key Words and Phrases: Affymetrix GeneChip probe level data, Microarray, Neuroblastoma, Weibull-normal mixture

1. はじめに

1990年代後半に網羅的な遺伝子発現強度の測定技術として，マイクロアレイが登場し，生命科学の研究を進展させるものとして多大な期待を集めた．マイクロアレイには遺伝子から生成されるmRNAとは相補的な配列をした短いDNA断片（プローブ）が一枚のスライドグラス上に高密度にスポットされており，その作成方式によりスタンフォード型のcDNA（complementary DNA）マイクロアレイと，オリゴヌクレオチドアレイ（Affymetrix社のGeneChipやAgilent社のDNAマイクロアレイ等）がある．プローブの長さは，cDNAでは数百塩基であり，Affymetrix社のGeneChipでは25塩基，Agilent社のものでは60塩基である．マイクロアレイ

* 広島大学原爆放射線医科学研究所計量生物研究分野，** 同遺伝子診断治療開発研究分野，*** 同自然科学研究支援開発センター：〒734-8551 広島市南区霞1-2-3

には複数のプラットフォームが存在するため，異なるプラットフォームで解析したデータを比較研究可能か？あるいは，異なる施設間で解析されたデータの比較は可能か？等という疑問に答えるべく，米国食品医薬品局が先導して，米国で入手できるDNAチップとマイクロアレイを総て比較した遺伝子発現解析の品質管理プログラム（マイクロアレイ品質管理プログラム，MAQC）を発表した．MAQC Consortium（2006）の公式報告において，それらの再現性について一定の評価が与えられ，このことによりマイクロアレイがさまざまな病院や研究施設で安定して用いられる可能性がでてきたといえる．一方，マイクロアレイデータを解析する上で注意しなければならないいくつかの問題点がある．アレイ上にスポットされているプローブとしては，遺伝子に特異的に反応する塩基領域が選ばれているが，Cambon et al.（2007）の報告にもあるように，目的外のmRNAと結合するクロスハイブリダイゼイションの影響は無視することはできない．また，Naef et al.（2001）によれば，塩基対を作る際，A，TよりG，Cのほうが強い結合力を持つことからおこるハイブリダイゼイションの効率の違いがあり，このことに起因するバイアスも考慮する必要がある．また，マイクロアレイ実験は，多くの複雑な工程で構成されていることにより，様々な種類の実験環境による偏りや誤差が生じていることが，Schuchhardt et al.（2000），Dudoi et al.（2000）によって明らかにされた．そのため解析の前処理として，バックグランドの調整，ノーマリゼーションおよび規準化が必要とされている．これらについては Irizarry et al.（2003a），Ohtaki et al.（2005），倉橋 他（2007）などの論文を参照されたい．特に，クロスハイブリダイゼイションなどの非特異的結合は，偽陽性な遺伝子発現の原因の一つであり，その検出の可否は解析精度に大きく関わるものである．これら前処理に関しては数多くの報告がされているが，解析方法により，同一のデータから異なった結果が得られるというCope et al.（2004），Milenaar et al.（2006），Seo and Hoffman（2006）の報告もみられ，Allison et al.（2006）も述べているように，生物学者や医学者は一体どの方法を用いれば良いのか困惑しているのが現状である．特にAffymetrix社GeneChipのプローブレベルでのデータの要約方法については，Li and Wong（2001a, 2001b）によって開発されたdChip，Affymetrix（2002）により開発されたMAS5.0，Irrizarry et al.（2003b）により開発されたRMA法，Wu et al.（2004）によって開発されたGC-RMA，Affymetrix（2005）によって開発されたPLIER等の数多くのアルゴリズムが提案されているが，未だ定石法と呼べる信頼性の高いアルゴリズムが確立されていないのが現状である．Irrizaryは Eisenstein（2006）の報告において，マイクロアレイのデータはそれ自体よい精度をもっているが，生データの前処理の方法により得られた遺伝子発現強度に大きな違いがあり，その後の統計解析に大きな影響を与えると述べている．また，Shi et al.（2005）は，前処理を適切に行えば，異なるプラットフォームでの一致性を改善できると報告している．さらに，同じサンプルデータに対して，いろいろな解析方法が選択可能であるが，これらの方法の可能性と限界についての科学的な検証が十分行われていないと述べている．

多くの場合，マイクロアレイデータ解析における標本数（細胞の個数）は相対的に小さく，通常，高々数百個程度であるのに対して，解析の対象とすべき遺伝子の個数は数千～2万個以上であり，いわゆる非正則かつ超多次元構造を有している．このような特殊なデータ構造により，古典的な数理統計的諸手法の適用が困難である．マイクロアレイデータに代表されるハイスループットデータの統計解析におけるさまざまな問題点については松浦 他（2004）の論文にまとめられている．

Ohtaki et al.（2005）は，上記のマイクロアレイデータ解析に関する問題点に対して，生物学および統計学の知見に基づいた独自の数理モデルを構築し，有用な解析手法の開発を進めてきた．その基盤は，遺伝子発現状況に‘通常On’，‘異常On’，‘擬On’および‘Off’の概念導入

にある．遺伝子の発現状態が 'On' とは，遺伝子が mRNA を作っている状況を意味しており 'Off' とは，mRNA を作っていない状態を意味するものとした．そして，Off の場合の測定値はクロスハイブリダイゼイションなどによる非特異的結合由来の偽発現量や実験誤差および測定誤差を表しているものと解釈した．遺伝子の発現・非発現の概念の導入により，マイクロアレイデータの数理構造を詳細に定式化することが可能となり，統計解析の効率や精度を上げることができた．

本研究では，Affymetrix 社 GeneChip のプローブレベルのデータの数理構造を明らかにし，遺伝子発現の特徴量を '遺伝子発現強度の大きさ' を表す指標と '遺伝子が On である確率' の 2 つの指標で表すことを提案し，それらの指標を用いて神経芽細胞腫の予後に関連する遺伝子の探索を行い，提案した指標の有効性について検証する．

2. Affymetrix 社の GeneChip データについて

Affymetrix（2001）によれば GeneChip マイクロアレイ（商品名）は，目的以外の mRNA が結合するクロスハイブリダイゼイション等の非特異的な結合によるバックグランド値の大きさを評価するために独自に設計されたマイクロアレイである．各ターゲット遺伝子に対して，遺伝子特異的な信号を検出するための PM プローブ，非特異的信号を検出するための 25 塩基からなる PM の中央の塩基を適当な塩基で置き換えた MM プローブからなるプローブペアが，11〜20 対用意されている．MM 値をどう扱うかにより，バックグランドに関する補正結果がかなり異なったものとなる．MM 値を用いる代表的なものとして Affymetrix 社が提供している MAS5.0（Affymetrix, 2002）があり，MM 値を用いない代表的なものとして RMA 法

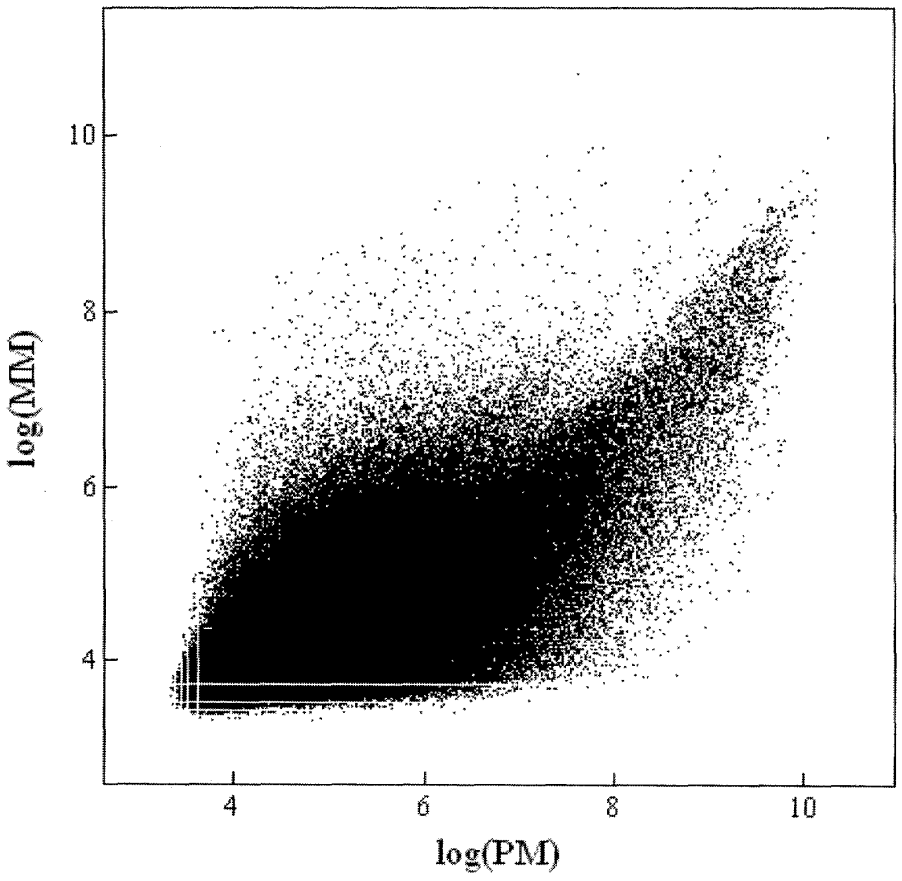

図 1　Affymetrix 社 GeneChip プローブレベルデータから得られた（*PM*, *MM*）の両対数散布図

(Irizarry et al. (2003)) が挙げられる．MAS5.0 は基本的には PM 値から MM 値を引いた値の対数値をもって補正後のシグナルとしているが，MM 値の代わりに新たに IMM 値という MM 値の調整値を定め，PM 値から MM 値を引いた値が負の値を取るのを避けている．また，RMA 法では，シグナル (S) に指数分布をノイズ (N) に正規分布を仮定し，測定値を $O=S+N$ として表し，O が与えられたもとでの事後平均を補正後のシグナルとしており，MM 値は用いていない．図1に現実のデータから得られた PM 値と MM 値の両対数散布図を示す．$\log PM$ と $\log MM$ の間に有意な相関が認められ (r=0.69, p 値<0.0001)，MM 値は非特異的な結合による偽発現に加え真のシグナル成分を含んでいることが示唆された．このことからも，PM 値から MM 値の単なる引き算によって非特異的な誤差を調整することは望ましくないことがわかる．我々は PM 値および MM 値に対して，非特異的結合やハイブリ効率を考慮に入れたな数理モデルを構築した．

本報告では，ターゲット遺伝子 $g(g=1,\cdots,G)$ の各プローブペア $j(j=1,\cdots,J)$ から得られた PM 値と MM 値のペアを $(PM_j^{(g)}, MM_j^{(g)})$ と表記する．また，ターゲット遺伝子に対応するプローブセット内において $PM>MM$ となるプローブペアの数 $(\#\{j|PM_j>MM_j, j=1,\cdots,J\})$ を X とする．

2.1 数理モデル

PM 値の対数値，MM 値の対数値に対して．

$$\begin{cases}\log PM_j^{(g)}=\log(1+\tau^{(g)}h_j^{(g)}s^{(g)}e^{u_{PM_j}^{(g)}}+\xi_{PM_j}^{(g)}e^{v_{PM_j}^{(g)}})+\varepsilon_{PM_j}^{(g)},\\ \log MM_j^{(g)}=\log(1+\tau^{(g)}\eta h_j^{(g)}s^{(g)}e^{u_{MM_j}^{(g)}}+\xi_{MM_j}^{(g)}e^{v_{MM_j}^{(g)}})+\varepsilon_{MM_j}^{(g)}\end{cases}$$

という定式化を行った．ここで，$\tau^{(g)}$ は遺伝子 g が On のときは1，Off のときは0をとる指示変数，$s^{(g)}$ は遺伝子 g の真の発現強度，$(h_j^{(g)}, \eta h_j^{(g)}), (0<\eta<1)$ は遺伝子 g のサイト j における PM プローブ，MM プローブへのハイブリ効率，$(\xi_{PM_j}^{(g)}, \xi_{MM_j}^{(g)})$ は非特異的な結合の大きさ，$(u_{PM_j}^{(g)}, u_{MM_j}^{(g)})$ および $(v_{PM_j}^{(g)}, v_{MM_j}^{(g)})$ は，それぞれ発現強度および非特異的結合の個体差，$(\varepsilon_{PM_j}^{(g)}, \varepsilon_{MM_j}^{(g)})$ は，測定誤差を表すものとする．遺伝子の発現状況が Off の場合は，

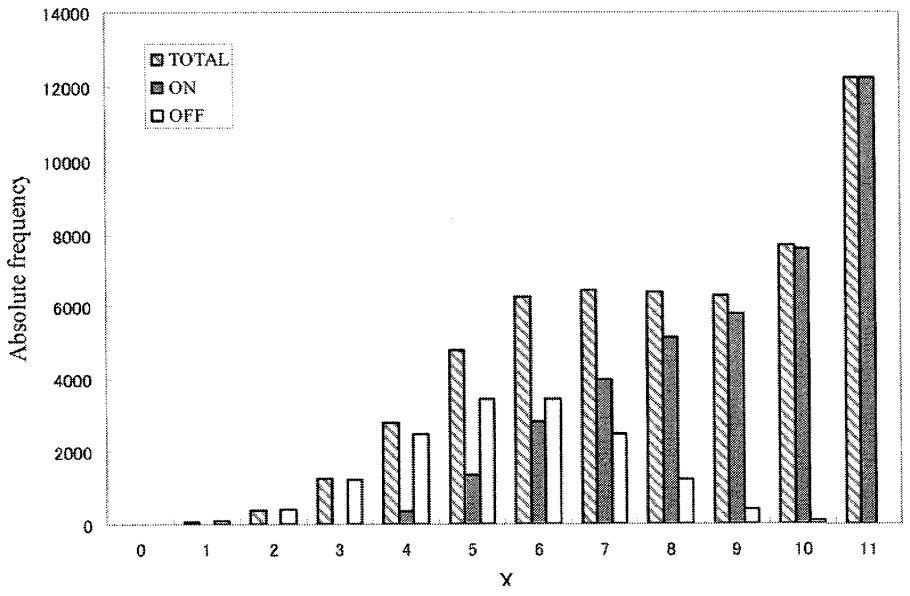

図2 Affymetrix 社 GeneChip プローブレベルデータの X 別頻度分布．ストライプの棒は全遺伝数，白抜きの棒は推定された Off 遺伝子数，黒棒は推定された On 遺伝子数を表す．

$$\begin{cases}\log PM_j^{(g)} = \log(1+\xi_{PM_j}^{(g)} e^{u_{PM_j}^{(g)}}) + \varepsilon_{PM_j}^{(g)}, \\ \log MM_j^{(g)} = \log(1+\xi_{MM_j}^{(g)} e^{v_{MM_j}^{(g)}}) + \varepsilon_{MM_j}^{(g)}\end{cases}$$

と表すことができ，PM，MM は同一な分布に従う，互いに独立な確率変数と見なすことができる．したがって X は，試行回数 J 回，成功確率 0.5 の二項分布に従うと考えられる．このことを利用して Off 遺伝子の数の推定をおこなった．図 2 に，X 値別にみた遺伝子数の度数分布を示す．全遺伝子数はストライプの棒，Off 遺伝子数は白抜きの棒，On 遺伝子数は黒抜きの棒で示した．全ゲノムの遺伝子発現強度の度数分布は，On 遺伝子と Off 遺伝子の混合分布として認識出来ることが示唆されている．詳細については Otani et al.（2007）の論文に述べてある．

3. 遺伝子特徴量としての 2 つの指標の導入

図 3 に，あるマイクロアレイデータから得られた遺伝子発現強度の，X 値別の箱型図表示を示す．発現強度が十分に大きければ，その遺伝子のプローブセット内のほとんどのプローブペアは $PM > MM$ を満たしている．また，プローブセット内に $MM \geq PM$ となるプローブペアが多数存在している遺伝子の発現強度は高くない．また，図 4 (a)，(b) に，発現強度が十分高い場合（a）と高くない場合（b）のプローブペア別にみた PM 値と MM 値をプロットしたものを示す．Cambon et al.（2007）の報告によれば，発現強度が十分に高いときは，クロスハイブリダイゼイションの大きさは，発現強度に比べ小さいということなので，遺伝子発現強度の大きさをプローブペアの要約量で表すことは可能であると考えられる．しかし，発現強度があまり高くない遺伝子については，プローブセット内に $MM \geq PM$ となるプローブペアが多数存在し，またクロスハイブリダイゼイションも真の遺伝子発現強度に比べて大きいと考えられるので，それらの要約量は遺伝子発現強度の大きさを表しているとは限らず，その場合，要約量の比較だけでは遺伝子の発現と非発現の分離も困難である．そこで我々は遺伝子発現の特徴量を，

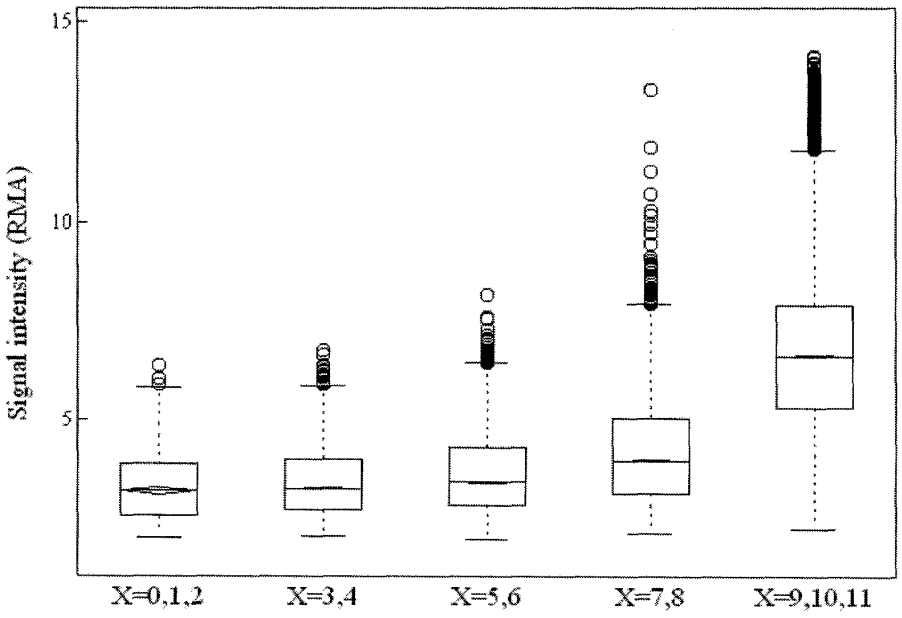

図 3　X 値別にみた遺伝子発現強度（RMA 法による）

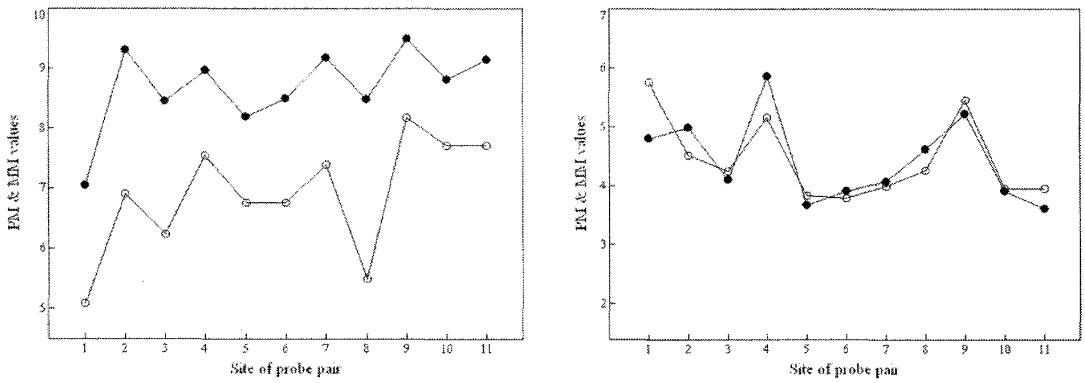

図4 プローブペアごとの測定値．○はPM値（対数値），●はMM値（対数値）を表す．(a) 発現強度が十分高い場合，(b) 高くない場合

1）プローブペアを一つの値に要約した遺伝子発現強度の大きさを表す指標，
2）PM値とMM値の大小関係に基づき推定した遺伝子の発現状態（On/Off）を表す指標，
により定められる2つの指標で表すことを提案した．

3.1 プローブペアの要約方法

MM値も遺伝子発現強度の情報を持っていることから，我々はプローブペアの要約方法としてPM値，MM値の両者を用いる方法を提案する．

アレイ i における遺伝子 g の発現強度を，

$$y_i^{(g)}=w\overline{PM}_i^{(g)}+(1-w)\overline{MM}_i^{(g)}$$

として定義する．ここで，$\overline{PM}_i^{(g)}=1/J\sum_j PM_{ji}^{(g)}$, $\overline{MM}_i^{(g)}=1/J\sum_j MM_{ji}^{(g)}$ である．重み w はN枚のマイクロアレイから得られた $2J$ 次元ベクトルの集合で構成されるプローブレベルデータ $\{(PM_{1i}^{(g)},\cdots, PM_{Ji}^{(g)}, MM_{1i}^{(g)},\cdots, MM_{Ji}^{(g)})|g=1,\cdots, G, i=1,\cdots, N\}$ に主成分分析を適応し，得られた第1主成分の固有ベクトル $(\alpha_1,\cdots,\alpha_J,\alpha_{J+1},\cdots,\alpha_{2J})$ を用いて，$w=\alpha_{PM}/\alpha_{MM}+\alpha_{PM}$ として与える．ここで，$\alpha_{PM}=1/J\sum_{j=1}^{J}\alpha_j$, $\alpha_{MM}=1/J\sum_{j=J+1}^{2J}\alpha_j$ である．図5に，あるマイクロアレイデータを用いて得られた第一主成分における固有ベクトル成分を示す．固有ベクトル成分のサイト依存性は特に認められず，PMの平均値とMMの平均値の重み付き平均値で遺伝子の発現強度の大きさを表すことができることが示唆された．PM値に対する重みとして0.6，MM値に対する重みとして0.4を得た．主成分分析を用いた要約方法の妥当性は，公開データ Human Genome U95 data set を用いて検証した．

3.2 Onになりやすさの指標 λ_g の導入

ターゲット遺伝子 g に対応するプローブセット j におけるPM値とMM値の大小関係を表す確率変数 $U_j^{(g)}$, $(j=1,\cdots, J)$ を $PM>MM$ ならば1，$PM\leq MM$ ならば0と定義する．PM値とMM値の大小関係に関して，$E\{\mathrm{logit}(1/N\sum_{i=1}^{N}U_{ji}^{(g)})\}=\lambda_g+\beta_j^{(g)}$ というモデル化を行う．

ここで，λ_g は遺伝子が発現する傾向の大きさを表す．$\beta_j^{(g)}$ は，ハイブリ効率などによるサイト特異的なブロック効果を表し，$\sum_{j=1}^{J}\beta_j^{(g)}=0$ を満たすものとする．よって λ_g の推定値として

$$\hat{\lambda}_g=\frac{1}{J}\sum_{j=1}^{J}\mathrm{logit}\left(\frac{1}{N}\sum_{i=1}^{N}U_{ji}^{(g)}\right)$$

が導かれる．

3.3 遺伝子の発現状況 'On/Off' の推定

真の発現強度 t の密度関数 f を

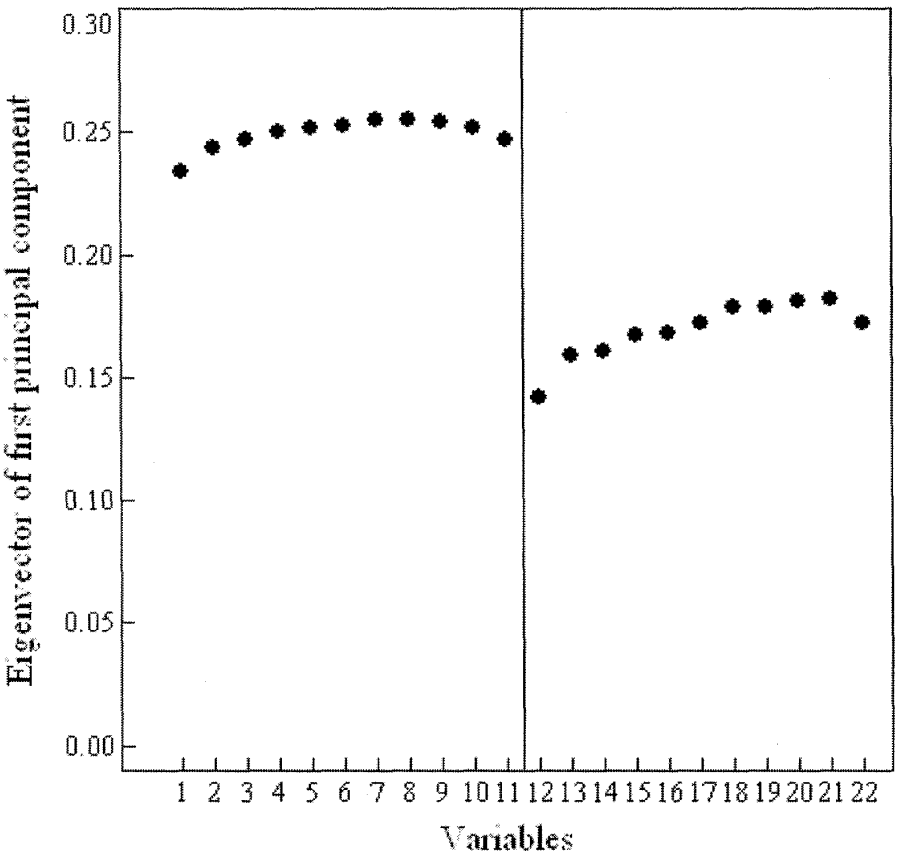

図5　PM 値と MM 値の対数変換値における第一主成分の固有ベクトル成分

$$f(t|\mu, \alpha, \xi)=\xi\delta(t)+(1-\xi)f_w(t|\mu, \alpha)$$

と表す．ここで，$\delta(t)$ はディラック関数，$f_w(t)$ は位置パラメータ μ，形状パラメータ α のワイブル密度関数，ξ は Off 遺伝子の混合割合を表すものとする．一方，測定誤差 w の密度関数は，正規密度関数 $\phi(w|\sigma^2)=1/\sqrt{2\pi\sigma^2}\exp(-w^2/2\sigma^2)$ に従うものとする．このとき $\hat{\lambda}_g$ の実現値をとと y すると，$y(=t+w)$ の密度関数 h は，

$$\begin{aligned}h(y|\xi, \mu, \alpha, \sigma^2)&=\int_0^{+\infty} f(t|\xi, \mu, \alpha)\phi(y-t|\sigma^2)dt\\&=\xi\phi(y|\sigma^2)dt+(1-\xi)\int_0^{+\infty} f_w(t|\mu, \alpha)\phi(y-t|\sigma^2)dt\end{aligned}$$

と表すことができる．データセット $\{\hat{\lambda}_1, \cdots, \hat{\lambda}_G\}$ が与えられているとき，対数尤度 $l(\theta)=\sum_{g=1}^{G}\log(h(\hat{\lambda}_g|\theta))$ の最大化により，パラメータ $\theta=(\xi, \mu, \alpha, \sigma^2)'$ の推定値 $\hat{\theta}$ を求める．このとき，遺伝子発現状態が Off あるいは On となる確率は，$\hat{\theta}$ および y が与えられたもとでの事後確率

$$\begin{cases}\Pr(g \text{ is Off}|y, \hat{\theta})=\dfrac{\hat{\xi}\phi(y|\hat{\sigma}^2)}{h(y|\hat{\theta})},\\ \Pr(g \text{ is On}|y, \hat{\theta})=1-\Pr(g \text{ is Off}|y, \hat{\theta})\end{cases}$$

によって与えられる．

神経芽細胞腫 40 例を用いた解析結果を図 6 (a)，(b) に示す．(a) において破線は実現値の頻度分布を示し，太い実践は推定されたワイブル・ノーマル混合分布，細い実践は On 遺伝子

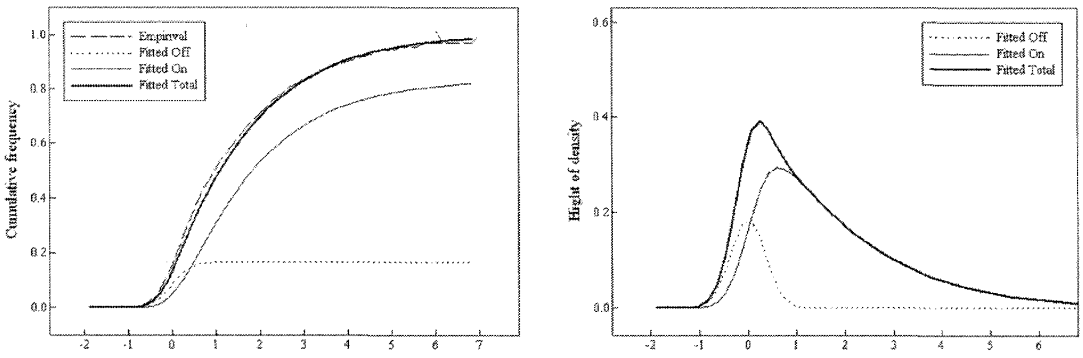

図6 (a) 推定されたワイブル・ノーマル混合分布関数と経験分布関数の適合度．太い実践は全遺伝子，細い実線はOn遺伝子，点線はOff遺伝子が対象．破線は経験分布関数を表す．(b) 推定されたワイブル・ノーマル混合密度関数．太い実践は全遺伝子，細い実線はOn遺伝子，点線はOff遺伝子が対象

についての分布関数，点線はOff遺伝子についての分布関数を表す．(b) は，それぞれの密度関数を表し，実線はワイブル・ノーマル混合密度関数，細い実線はOn遺伝子，点線はOff遺伝子についての密度関数を表す．なお，この場合のワイブル・ノーマル混合分布のパラメータ θ の推定値は，$(\hat{\xi}, \hat{\mu}, \hat{\alpha}, \hat{\sigma}^2) = (0.17, 1.96, 1.1, 0.13)$ で与えられた．

4. 表現型関連遺伝子としてのOR型遺伝子とAND型遺伝子

Tang et al. (2006)，Hiyama et al. (2008) によれば，小児の癌である神経芽細胞腫の予後は，死に至る予後不良なものと，腫瘍が完全に消失し治癒する予後良好なものに2極化しているのを特徴とし，*MYCN* 遺伝子は発現すれば予後が悪くなる予後関連遺伝子の一つである．図7に予後良好40症例，予後不良21症例 *MYCN* の発現強度と $X = \#\{j | PM_j > MM_j\}$, $(j = 1, \cdots, 11)$ の散布図を示す．白丸は予後良好症例，黒丸は予後不良症例である．また，便宜的に $X \geq 7$ ならば遺伝子発現状況はOn，$X \leq 6$ ならばOffとしたときの，

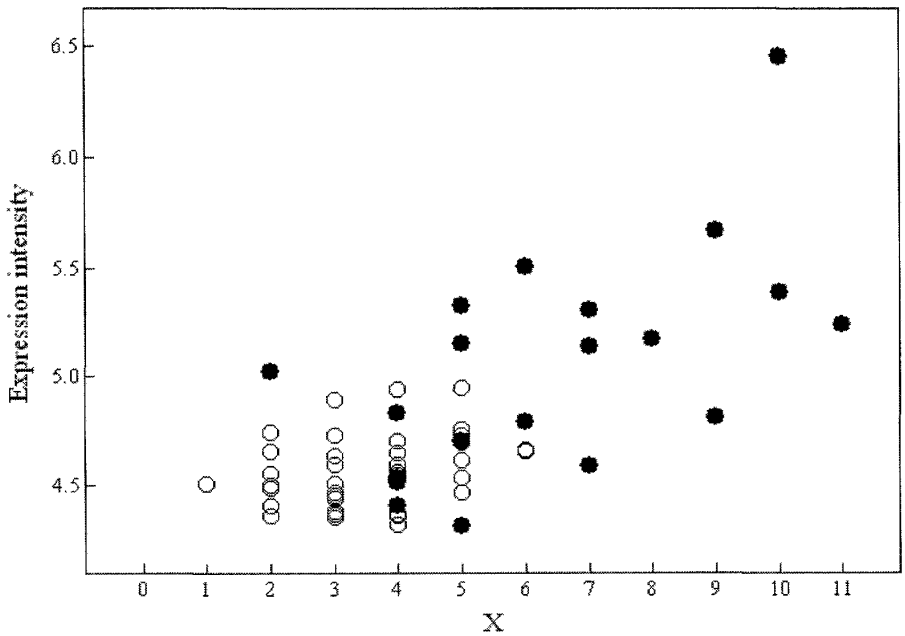

図7 神経芽細胞腫各症例におけるMYCN遺伝子のとシグナル強度．○は予後良好症例，●は予後不良症例を示す．

MYCN の発現・非発現と予後の関係を表 1 に示す．*MYCN* が On ならば予後不良である．その対偶をとれば，予後良好ならば全ての症例で *MYCN* は Off であることを示している．予後不良群においては On と Off が混在しており，遺伝子関連遺伝子の探索として通常用いられている発現強度の 2 群における平均値の比較によっては，このような遺伝子は候補遺伝子として選択されず偽陰性となる可能性が高い．また，*MYCN* 以外の予後に直接関連する遺伝子の存在が示唆される．このことに注目し，複数の遺伝子と表現型の関連性について 'AND' 型と 'OR' 型の単純なモデルを想定した．図 8 にその概念図を示す．R は表現型を G は遺伝子を表している．さらに，遺伝子の発現状況と表現型との関係を表すために，4 つのタイプの遺伝子を下記のように定義した．

定義 1　g^+ ならば R^+ のとき，遺伝子 g は OR（On）型といい，g^- ならば R^+ のとき，OR（Off）型という．また，R^+ ならば g^+ のとき，遺伝子 g は AND（On）型といい，R^+ ならば g^- のとき，遺伝子 g は AND（Off）型という．ここで，g^+ は On 遺伝子，g^- は Off 遺伝子，R^+, R^- は 2 値型の表現型を表す．

それぞれのタイプの遺伝子について，表現型と遺伝子の発現・非発現の関連性を示す 2×2 表を表 2 に示す．この表より，*MYCN* は予後不良に対し OR（On）型遺伝子であることが示唆されている．

5. 実データの解析

神経芽細胞腫で予後良好な 40 症例と予後不良な 21 症例から得られたマイクロアレイデータを用いて，OR（On）型遺伝子の探索を以下の手順で行った．

Step 1：予後良好群，予後良好群を用いて遺伝子が 'On/Off' となる事後確率を算出した．
Step 2：OR（On）型遺伝子は，予後良好群で全て Off なので，予後良好群で Off となる確率が

表 1　神経芽細胞腫の予後別 MYCN の発現・非発現の度数

	予後不良	予後良好
MYCN: On	9	0
Off	12	40

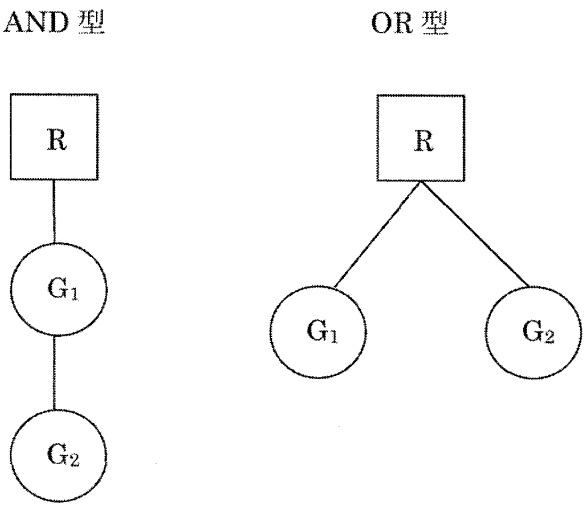

図 8　遺伝子発現の表現型に対する寄与：AND 型モデルと OR 型モデル

表2 OR型遺伝子とAND型遺伝子の発現状態と表現型の関係

OR(On) 型

	R^+	R^-
g^-	a	0
g^-	N_1-a	N_2
Total	N_1	N_2

OR(Off) 型

	R^+	R^-
g^+	N_1-b	N_2
g^-	b	0
Total	N_1	N_2

AND(On) 型

	R^+	R^-
g^+	N_1	c
g^-	0	N_2-c
Total	N_1	N_2

AND(Off) 型

	R^+	R^-
g^+	0	d
g^-	N_1	N_2-d
Total	N_1	N_2

0.5以上である遺伝子群を選択した．予後不良例ではOnである場合とOffである場合があるので，選択した遺伝子の中から，予後不良症例でOnとなる確率が予後良好群でOnとなる確率より0.2高い遺伝子を選択した．（この段階で候補遺伝子同定の対象となるプローブセットを54109個から744個に絞り込むことができた．）

Step 3：得られた744遺伝子について，予後良好群でOffとなる確率の高いもの上位100遺伝子を選択し，その中で，予後不良群でOnとなる確率が高いもの上位5遺伝子をOR（On）型の予後関連遺伝子とした．

選択された上位5個の遺伝子の予後良好・不良群における発現状況を表3に示す．また，得られたシグナル強度の2群間での平均値，標準偏差，t統計量およびt統計量を大きさ順に並び替えた場合の順位を表4に示す．選択された5遺伝子についてRT-PCR法（reverse transcription-polymerase chain reaction）でmRNAの定量化を行った結果，それらがいずれも神経芽細胞腫の候補遺伝子であることが確認された．また，選択された遺伝子が神経芽細胞腫に関連した機能を持っていることがAlaminos et al.（2003），Lastowska et al.（2007）の論文に示されているが，*MYCN*以外の遺伝子名については現段階では公表を差し控えさせていただきたい．

6. 考　　察

我々は，2001年に遺伝子多様性モデル開発事業に参入以来，マイクロアレイデータ解析の可能性と限界を明らかにすることをテーマにcDNAマイクロアレイやオリゴヌクレオチドマイクロアレイデータの数理構造について検討してきた．マイクロアレイ解析の基本的な目的である発現レベルの類似や相違について解析を行う前に，マイクロアレイデータが得られるまでの複雑な背景を考慮したノイズの調整が必須であると考えてきた．たとえば，前処理の第一段階としてのバックグラウンド調整は多くの場合，Eisen（1999）が示しているように，観測値（O）からバックグランド値（B）を減じた後，対数変換をする方法が用いられているが，この方法を用いると，観測値とバックグラウンドが近い値をとる場合，補整値は非常に不安定なものとなる．例えば，$O-B$が負値をとり対数変換できないものについては適当な正の数で置き換えるなどのことを行っているようである．図9にGE Healthcare社のヒト全ゲノムマイクロアレイデータを用いて，$\log B$と$\log O$の関係を示す．直線は$Y=X$のアイデンティティーラ

表3　選択された5つの遺伝子の予後良好・不良群における発現状況

遺伝子名	予後良好群で On である確率[a]	予後不良好群で On である確率[b]	確率の差 $b-a$	順位
G1	0.26	0.89	0.63	1
G2	0.38	0.94	0.56	2
G3	0.38	0.88	0.50	3
G4	0.16	0.58	0.42	4
MYCN	0.19	0.58	0.39	5

表4　選択された5つの遺伝子発現強度の予後良好・不良群における平均値とおよびそれらの差に関する t 値

	予後良好群		予後不良群		t 値	順位
	シグナル平均	標準偏差	シグナル平均	標準偏差		
G1	4.932	0.190	5.403	0.522	5.02	69
G2	4.591	0.187	5.211	0.617	5.77	14
G3	4.589	0.152	5.121	0.523	5.86	11
G4	4.169	0.105	4.628	0.670	4.16	674
MYCN	4.571	0.161	5.020	0.506	5.04	65

インを示す．観測値が低いとき（発現状態は 'Off' のとき）はバックグランドも観測値も同等であることが示されている．Ohtaki et al.（2005）は，$\log O - \log B$ によりバックグラウンド調整を行い，アレイ間のデータの規準化には，Off 遺伝子群における平均と分散を用いた調整の適用を提案した．上記の2つの方法によるバックグラウンド調整法を適用して得られた遺伝子発現強度分布を図10に示す．(a) は既存の方法による遺伝子発現強度分布を示している．負の値をとり対数変換できないものは欠落値扱いとし−9で表した．(b) は提案法による遺伝子発現強度分布である．適切なバックグランド調整により遺伝子発現強度分布が On 遺伝子と Off 遺伝子の混合分布となっていることが示唆されている．このように，マイクロアレイを用いた研究が再現性のある信頼性の高い結果を得るためには，非特異的結合によるバックグランド値の調整や，チップ間の比較を可能にするためのノーマリゼイション，複数のプローブペアから得られた測定値の一つのターゲット遺伝子の発現強度への要約処理などという前処理に対して細かな検討を重ねる必要があると考える．

一方，ハイスループットデータ解析の問題点として，説明変数の数（遺伝子数）がサンプル数（細胞数）に比べてきわめて大きいということがある．生体内における複数の遺伝子のネットワークの推定は，マイクロアレイデータ解析の最終的なゴールの一つであるが，井元（2007）も，たとえ全遺伝子でネットワークが計算上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サンプル数を考えると決して高くなく，解析に用いる遺伝子セットの適切な定義が必要であると述べている．この問題に対しても遺伝子発現・非発現の概念の導入は説明変数の絞り込みという点で有効に働くものと考えている．本報告においては，神経芽細胞腫の予後関連遺伝子の探索を例に，遺伝子の発現・非発現に基づく表現型関連遺伝子の絞り込みと，複数の遺伝子と表現型の関係を念頭においた表現型関連遺伝子の探索方法の提案を行った．新しい試みとして OR（On）型の関連遺伝子の探索を行った．その結果，生物学的な見地からも妥当な遺伝子を同定することができた．

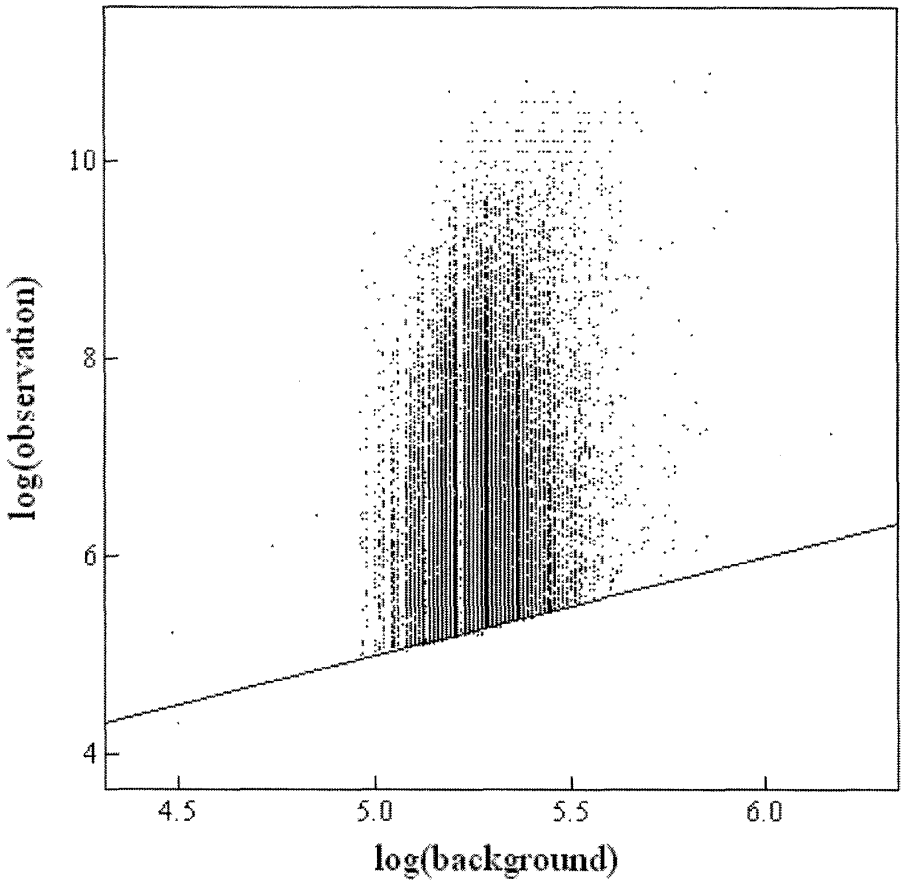

図9 (a) 従来法によるバックグランド調整後の遺伝子発現強度のヒストグラム．負値のため対数変換ができないものは欠落値（−9）で置き換えた．(b) 提案法によるバックグランド調整後の遺伝子発現強度．

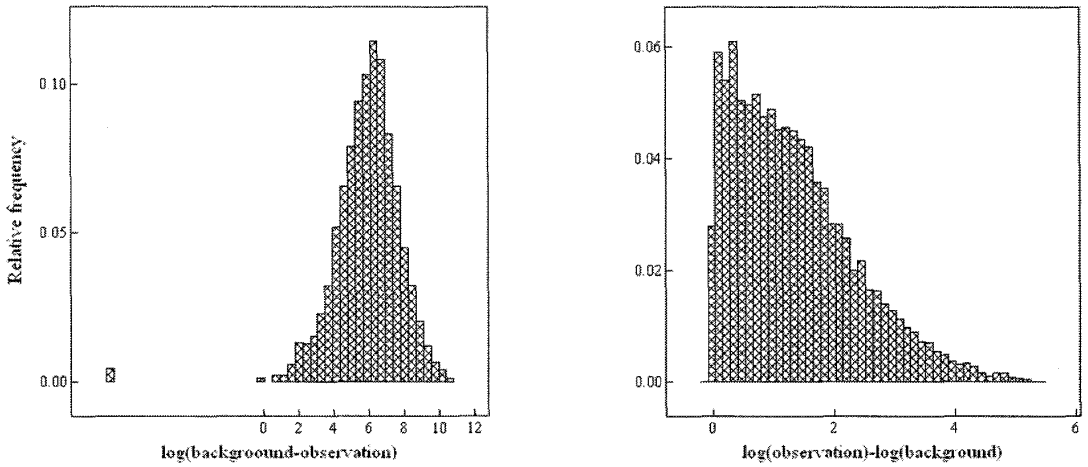

図10 バックグランド（B）の対数値と観測値（O）の対数値の散布図．直線は $O=B$ を表す．

謝　辞

本研究の一部はNEDOおよび，文部科学省科学研究費基盤研究（B），課題番号18300095の支援を受けて行ったものである．

参考文献

Affymetrix, Inc. (2001). GeneChip arrays provide optimal sensitivity and specificity for microarray expression

analysis, Technical Note.
Affymetrix, Inc. (2002). Statistical Algorithms Description Document.
Affymetrix, Inc. (2005). Guide to probe logarithmic intensity error (PLIER) estimation, Technical Note.
Alaminos, M., Mora, J., Cheung, NK. V., Smith, A., Qin, J., Chen, L. and Gerald, W. L. (2003). Genome-wide analysis of gene expression associated with MYCN in human neuloblastoma, *Cancer Research*, **63**, 4538–4546.
Allison, D. B., Cui, X., Page, G. P. and Sabripour, M. (2006). Microarray data analysis: from disarray to consolidation and consensus, *Nature Rev. Genet.*, 55–65.
Cambon, A. C. et al. (2007). Analysis of probe level patterns in Affymetrix microarray data, *BMC Bioinformatics*, **8**, 146.
Cope, L. M., Irizarry, R. A., Jaffee, H. A., Wu, Z. and Speed, T. P. (2004). A benchmark for Affymetrix GeneChip expression measures, *Bioinformatics*, **20**, 323–331.
Dudoit, S., Yang, YH., Callow, MJ. and Speed, TP. (2000). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments, Technical Report #578
Eisen, M. (1999). Scan Analyze User Manual, http//rana.lbl.gov/manuals/ScanAnalyzeDoc.pdf
Eisenstein, M. (2006). Quality control, *Nature*, **442**(31), 1067–1070.
Hiyama, E., et. al. (2008). Screening at 6 months of age reduced mortality of neroblastoma: A retrospective population-based cohort study including more than 13 million Japanese screened infants, *Lancet*, **371**, 1173–80.
井元清哉 (2007). 「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」, 『日本統計学会誌』, **37**(1), 55–69.
Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U. and Speed, T. P. (2003a). Exploration, normalization, and summaries of high density oligonucleotide array probe level data, *Biostatistics*, **4**, 249–264.
Irizarry, R. A., Bolstad, B. M., Collin, F., Cope, L. M., Hobbs, B. and Speed, T. P. (2003b). Summaries of Affymetrix GeneChip probe level data, *Nucleic. Acids. Res.*, **31**(4), e15.
Lastowska, M., Viprey, V., Santibanez-Koerf, M., Wapper, I., Peters, H., Cullinane, C., Roberts, P., Hall, A. G., Tweddle, D. A., Pearson A. D. J., Lewis, I., Burchill, S. A. and Jackson M. S. (2007). Identification of candidate genes involved in neuroblastoma progression by combining genomic and expression microarray with survival data, *Oncogene*, 1–13.
倉橋一也, 伊藤陽一, 松山裕, 大橋靖雄, 西尾和人 (2007). 「cDNA マイクロアレイ解析における正規化手法の性能比較」, 『日本統計学会誌』, **36**(2), 147–163.
Li, C. and Wong, W. H. (2001a). Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection, *Proc. Natl. Acad. Sci. USA*, **98**(1), 31–36.
Li, C. and Wong, W. H. (2001b). Model-based analysis of oligonucleotide arrays: model valiation, design issues and standard error application, *Genome Biol*, **2**, 1–11.
松浦正明, 牛嶋大, 宮田敏 (2004). 「メディカルインフォマティクスのためのゲノム関連データの解析法とその問題点」, 『計量生物学』, **25**(2), 117–34.
MAQC Consortium (2006). The Microarray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements, *Nature Biotechnology* **24**, 1151–1161.
Millenaar, F F., Okyere, J., May, S T., Zanten, M V., Voesenek, L ACJ. and Peeters, A JM. (2006). How to decide? Different methods of calculating gene expression from short oligonucleotide array data will give different results, *BMC Bioinformatics*, **7**, 137.
Naef, F., Lim, D. A., Pantil, N. and Magnasco M. A. (2001). From features to expression: High-density oligonucleotide array analysis revisited, *Technical Report* **1**, 1–9.
Ohtaki, M., Otani, K., Satoh, K., Kawamura, T., Huiyama, K. and Nishiyama M. (2005). Model-based analysis of microarray data: expression of differentially expressed genes between two cell types based on a two-dimensional mixed normal model, *Jpn. J. of Biometrics*, **2**(1), 31–48.
Otani, K, Hiyama, K., Satoh K., Shimamoto, T., Mohamad D., Andoh, M., Tonda, T., Kohda, M., Ohazaki, Y., Nishiyama, M. and Hiyama, E. (2007). A Mathematical Model for Affymetrix GeneChip Probe Level Data, *JP Journal of Biostatistics*, **1**(3), 283–306.
Seo, J. and Hoffman, E. P. (2006) Probe set algorithms: is there a rational best bet?, *BMC Bioinformatics*, **7**, 395.
Shi, L., Tong, W., Fang, H., Scerf, U., Han, J., Ouri, R. K., Frueh,F. W., Goodsaid, F. M., Gou, L., Su, Z., Han, T., Fuscoe, C. J., Xu, Z. A., Patterson, T. A., Hong, H., Xie, Q., Perkins, R. G., Chen, J. J. and Casciano, D. A. (2005). Cross-platform comparability of microarray technology: Intra-platform consistency and appropriate

data analysis procedures are essential, *BMC Bioinformatics*, **6**, S12.

Tang, X. X., Zhao, H., Kung, B., Kim, D. Y., Hicks, S. L., Dohn, S. L., Cheng, N. K., Seeger, R. C., Evans, A. E. and Ikegaki, N. (2006). The MYCN enigma: significance of MYCN expression in neuloblastoma, *Cancer Res*, **66** (5), 2826–2833.

Schuchhardt, J., Beule,D., Malik, A., et al. (2000). Normalization strategies for cDNA microarrays, *Nucleic Acids Research*, **28**, E47, 1–5.

Wu, Z., Irizarry, R. A., Gentleman, R., Martinez-Murillo, F. and Spencer, F. (2004). A model-based background adjustment for oligonucleotide expression arrays, *Journal of the American Statistical Association*, **99**(468), 909–917.