日本統計学会誌
第36巻，第2号，2007年3月
147頁～163頁

cDNAマクロアレイデータ解析における正規化手法の性能評価

倉橋一成*，伊藤陽一*，松山　裕*，大橋靖雄*，西尾和人**

Evaluation of Normalization Methods for cDNA Macroarray Data

Issei Kurahashi*, Youichi Ito*, Yutaka Matsuyama*, Yasuo Ohashi* and Kazuto Nishio**

近年DNAアレイ技術によって何千もの遺伝子の発現レベルを同時に観察し，生物学的に重要な情報を大量に得ることができるようになった．しかし測定された発現強度には様々な測定バイアスが含まれることや，発現強度に依存した不等分散性が存在することが指摘されている．これらの問題に対して，正規化と呼ばれる一連の手法が様々提案されている．これらの正規化手法を比較した研究は多くあるが，主にバイアス補正に注目しており，不等分散性はあまり重視していない．そこで本研究では不等分散性を改善するため，Huberによる分散安定化変換を単純化した簡便な正規化手法を提案し，既存の手法と比較，評価した．

In recent years, microarray technology allows the monitoring of expression levels for thousands of genes simultaneously and the acquiring large amount of biologically important information. However, many experimenters indicate that measured intensity involve various measurement biases and non-constant variance depended on intensity. To solve these problems, a variety of normalization methods are proposed. Many studies focused on comparison of normalization methods from the viewpoint of correcting biases, but not of stabilizing variation. Therefore, this study aims to propose convenient normalization method simplified Huber's variance stabilizing transformation and compare it to existent methods.

Key Words and Phrases: Microarray, Macroarray, Normalization method, Measurement bias, variance stabilization

1. はじめに

近年Folderら（1991）やSchenaら（1995）によって発明されたDNAアレイ技術の発達によって何千もの遺伝子の発現レベルを同時に観察することが可能となり，生物学や医学研究など，遺伝子発現の研究の中で幅広く使われるようになった．しかし，測定される発現強度には以下の2つの問題点が指摘されている．まず第1には，DNAアレイ実験を行う過程で混入する発現強度への測定バイアスが挙げられる．測定バイアスが入る可能性のある実験過程の例は，Schuchhardtら（2000）やFanら（2004）が挙げているように，アレイの作成，サンプルからのmRNAの抽出，mRNAからcDNAへの転写，cDNAへのラベリング，cDNAのアレイへのハイブリダイゼーション，アレイの洗浄，スキャナによる発現強度の測定などである．また第2には発現強度に依存した不等分散性が挙げられる．

* 東京大学大学院医学系研究科，〒113-0033 東京都文京区本郷7-3-1
** 近畿大学医学部ゲノム生物学教室，〒589-8511 大阪狭山市大野東377-2

これら2つの問題に対して，正規化と呼ばれる一連の手法が様々提案されてきた．第1の問題である測定バイアスの1つとして，二蛍光標識を用いたマイクロアレイにおけるDye biasの存在が指摘され，Global正規化が提案されている（Simonら（2003））．Dye biasは，遺伝子にラベリングされている二種類の蛍光色素（Dye）の発光強度が，サンプル間で系統的に異なっている場合に確認される．Global正規化は，Dye biasが存在しなければアレイ毎の発現強度の中央値はほぼ一定であるという仮定のもと，アレイ毎に得られる対数発現強度の中央値を遺伝子毎の対数発現強度から引くことで，このバイアスを取り除くことを試みている．これはアレイ内発現強度をアレイ毎の中央値で割ることに相当する．また，Bolstadら（2003）は単純にGlobal正規化を行うのではなく，中央値の中央値（アレイ毎に得られた発現強度の中央値を全体とした際の中央値）をかけることによって，他の正規化手法と比較が可能にできるように工夫を行っている．その後，Dye biasが発現強度に依存するというIntensity biasが指摘され，Global正規化では十分にバイアスを取り除けないため，非線形な正規化が提案されている．まずYangら（2001）がLOESS正規化を提案し，その後Workmanら（2002）によってスプライン関数による正規化が提案されている．

その後，Dye bias以外の様々な測定バイアスが指摘され，それらを取り除く正規化手法が提案されてきた．これらのバイアスは，アレイ内に存在するバイアスと，アレイ間に存在するバイアスの2つに分類できる．アレイ内バイアスの例を挙げると，アレイを作成する際のprint tipが原因で起こるprint-order biasやprint tipグループ毎のintensity bias，ハイブリダイゼーションの際にサンプルがアレイ上に均一に広がらない場合に起こるspatial biasなどがある．これらはそれぞれ，Uchidaら（2004），Bergerら（2004），Colantuoniら（2002）が指摘し，それらを取り除くための正規化手法を提案している．また，アレイ間に存在するバイアスは主にintensity biasや実験条件が異なることが原因で起こるバイアスであり，これを取り除くために，Bolstadら（2003）による線形なバイアスを取り除く正規化，Bolstadら（2003）やEdwerdsら（2003）による非線形なバイアスを取り除く正規化，Edwerdsら（2003）やYoonら（2003）によるbackground noiseを利用した正規化などが提案されている．さらに，cDNAマクロアレイにおいては，放射性同位体をラベルとして使用しているため，遺伝子発現量が少なくても検出できるという利点がある反面，放射線の曝露時間が長いと遺伝子発現量が大きいスポットが近接するスポットに影響を及ぼしてしまうというneighborhood biasが存在することが，Dugganら（1999）やHollowayら（2002）によって指摘されている．そのために，Schuchhardtら（2000）が遺伝子ごとに測定されたbackground noiseを利用してバイアスを取り除く正規化手法を提案しているが，データの性質によってはうまく機能しないことが経験上知られている．

さらに近年においては，第2の問題である発現強度に依存した不等分散性が問題視されている．これまではデータの分散を一定にするために一般的に対数変換が行われてきたが，Cheadleら（2003）によって対数変換にかわる正規化としてZスコアによる変換が，Huberら（2002）やInoueら（2004）によって逆双曲線正弦関数による変換が，またGellerら（2003）やKonishiら（2004）によって他の変数変換による変換が提案されてきた．しかし，これらの変換は一般に計算アルゴリズムが複雑なものが多く，データ依存性が高いと思われる．

これらの正規化に関する研究結果から，正規化を適用していく流れは図1のようになると思われる．1．まずアレイデータの確認を行い，2．測定バイアスを除去する．3．そして分散を安定化させ，4．最後に様々な遺伝子解析を行う．ここで重要なことは，測定バイアスは実験系に応じてそれぞれ違うものが存在していると考えられるので，実験毎に考察していくべきであるという事．また分散の安定化に関しては，全てのアレイ実験系に共通な問題であると考え

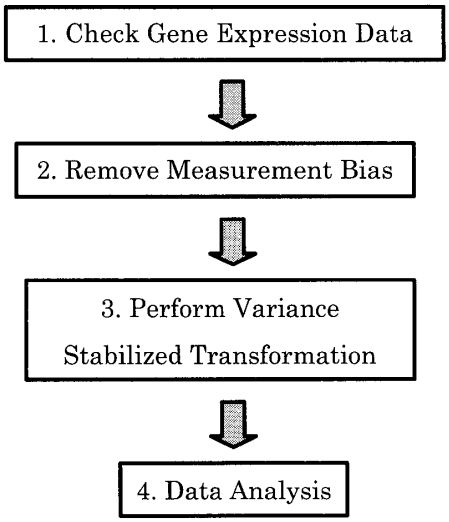

図1　正規化手法選択のプロセスを示したフローチャート

られるので，全てのアレイ実験系に共通な変数変換が必要であると考えられるという事である．

以上のことを踏まえて，本研究では cDNA マクロアレイ特有の neighborhood bias を，アレイ内繰り返し測定を利用して取り除く正規化を行う．また，不等分散性を改善するため，Huber の分散安定化変換を単純化した簡便な正規化手法を提案する．さらに既存の正規化手法と比較することによって，提案した正規化手法の性能を確認する．

2. cDNA マクロアレイデータ

2.1 実験データ

本研究で使用したデータは国立がんセンター薬効試験部において行われた，Clontech 社のカスタムアレイである Atlas Human Pharmacology array ver. 4.6（新フィルターアレイ）と Atlas Human Pharmacology array ver. 3（旧フィルターアレイ）の比較をするための実験データである．これらのフィルターアレイにはラベルとして放射性同位体が使われており，Clontech 社のカスタムサービスを利用して興味のある遺伝子のみをスポットしている．新フィルターアレイにおいては，1 枚のアレイに 1176 個のスポットがあり，positive control の遺伝子が 15 スポット，negative control の blank が 19 スポット，またいくつかの遺伝子に関して繰り返し測定が可能となるように設計されている．同一アレイ内での繰り返し測定は，バイアスの程度を調べる目的でスポットしており，2 回繰り返し測定されているものが 31 遺伝子（62 スポット），3 回繰り返し測定されているものが 62 遺伝子（186 スポット）である．そのため，測定された遺伝子の数は 984 遺伝子となる．今回解析の対象としたのは新フィルターアレイの 3 枚で，control や繰り返し測定も含めた 1176 個の全スポットを使用した．これら 3 枚の新フィルターアレイの散布図を，図 2 に示す．

2.2 発現強度の構造

アレイ実験系の正規化手法は，測定バイアスを取り除くための正規化手法と，分散安定化変換についての正規化手法の 2 種類に分けることができる．そのため本研究では，観測された発現強度は次の 2 つの部分から構成されていると考えた．1 つ目は遺伝子発現に直接関係の無い部分であり，2 つ目は遺伝子発現に直接関係している部分である．この 2 つ目の部分が真の遺伝子発現であると考えると 1 つ目の部分を測定バイアスと考えることができ，これを式で表すと次のようになる．

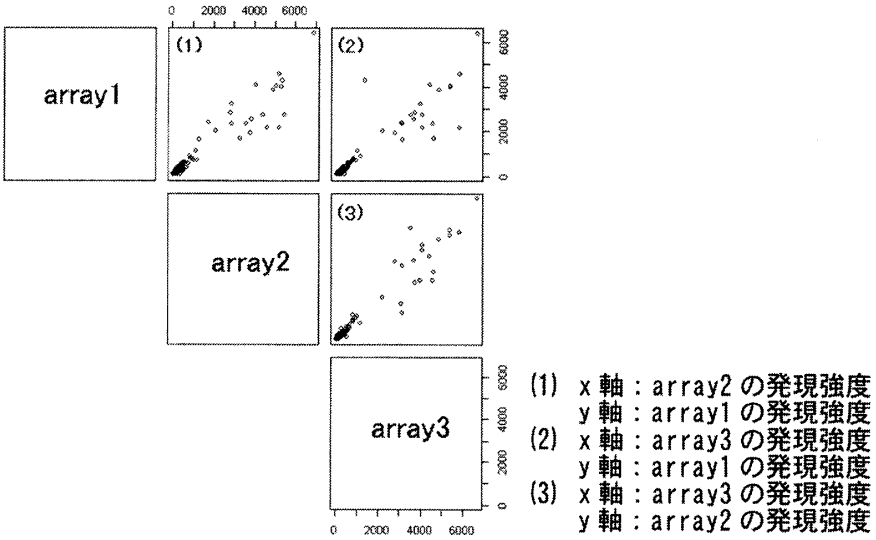

図2 新フィルターアレイの発現強度を示した散布図行列

$$Intensity = Measurement\ Bias + Gene\ Expression$$

今後は観測された発現強度を“発現強度”と表現し，真の遺伝子発現を“遺伝子発現”と表現する．さらに第1項はアレイ間測定バイアスとアレイ内測定バイアスに分けることができるので，発現強度は次のように表現できる．

$$Intensity = Inter\text{-}array\ Bias + Intra\text{-}array\ Bias + Gene\ Expression$$

さらにアレイ内測定バイアスは，遺伝子またはスポットに特異なバイアスの2種類に分けることができるであろう．

2.3 測定バイアスと不等分散性

本研究では発現強度の，遺伝子発現に関係していない部分は測定バイアスであると考えた．そのため，backgroundも測定バイアスの一部であると考えられる．しかしbackgroundは，background noiseと表現されることもあり混乱が生じる可能性があるので，本研究でbackgroundに対して次のように定義する．まずbackgroundはアレイ間測定バイアスであり，background値はアレイ毎の特有な値である．またbackground noiseはスポット毎に測定されている数値であり，background値にnoise（誤差）が加わったものと定義する．Affymetrix社が行っているbackground correctionは，background値を推定せず直接background noiseをスポット毎に引いている．またEdwerdsら（2003）やYoonら（2004）はbackground noiseを利用してbackground値を推定し，それを発現強度から引いている．本研究では後者の方法を用いた．またその他の測定バイアスは，実験データから推定されたものを発現強度から引くことで取り除く．測定バイアスを取り除く正規化手法のうち，本研究で評価を行ったものは3.1節で説明する．

また本論文では図1に示したように，測定バイアスを取り除いた後に分散安定化変換を行う．発現強度の不等分散性は様々な先行研究によって述べられており，分散安定化変換も大きな意味で正規化であると考えられる．

3. 正 規 化

3.1 測定バイアスを取り除く正規化手法

散布図を確認すると，本実験ではアレイ 1 の発現強度がアレイ 2 とアレイ 3 に比べて小さくなっていると思われる．また図 3 や図 4 によって，neighborhood bias の存在も確認できる．図 3 は新フィルターアレイ 1 の画像の一部であり，図 4 はアレイ内繰り返し測定されている遺伝子の発現強度をプロットしたものである．よって，これら 2 種のバイアスを取り除く必要がある．つまり，アレイ間測定バイアスとしてアレイ間の線形な測定バイアスが存在し，アレイ内測定バイアスとして neighborhood bias が存在していると考えた．そのため本研究では前者の測定バイアスを取り除くために，Bolstad ら（2003）による Global 正規化と Edwerds ら（2003）による background noise を利用した正規化を適用した．この際 background noise が測定されていなかったので，代わりに blank spot の発現強度を利用した．しかし blank spot にも neighborhood bias が存在していると思われたスポットがあったので，そのスポットを除いた平均値を background 値の推定値とした．また後者の測定バイアスは既存の正規化手法では充分に取り除くことができなかった．そのため同一遺伝子のアレイ内繰り返し測定を利用し，本研究独自の正規化を適用した．以下にこの正規化の説明を行う．

neighborhood bias を取り除く正規化

本研究では neighborhood bias は cDNA マクロアレイ実験系に特異的なバイアスであると考えた．またこの測定バイアスは発現強度のみに依存していると考え，次の式によって取り除くことを考えた．

$$I_g' = I_g^l - f(I_{g'}^l) \qquad (1)$$

ここで，I_g' は遺伝子 g における neighborhood bias を取り除いた後の発現強度，I_g^l は四方に隣接するいずれかのスポットに発現強度の大きいスポット（本研究では発現強度が 400 以上のスポットとした）がある遺伝子 g の発現強度，$I_{g'}^l$ は遺伝子 g のスポットに隣接する発現強度の大きい遺伝子 g' の発現強度である．発現強度の大きいスポットが四方に 2 個以上存在する場

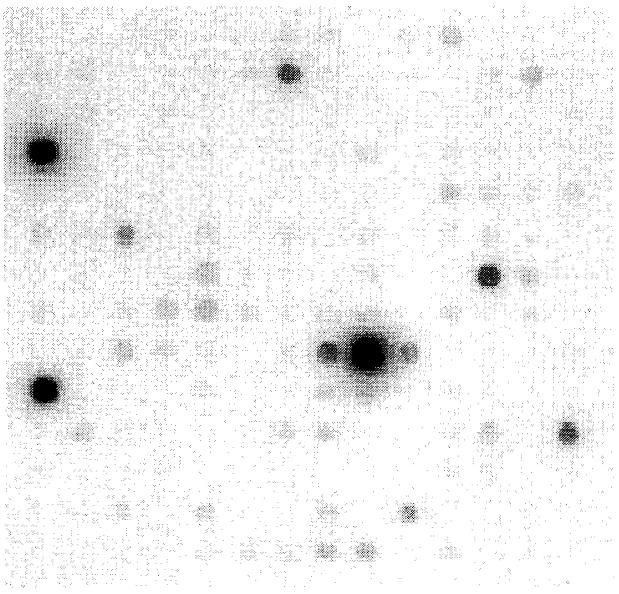

図 3 新フィルターアレイ 1 の画像の一部

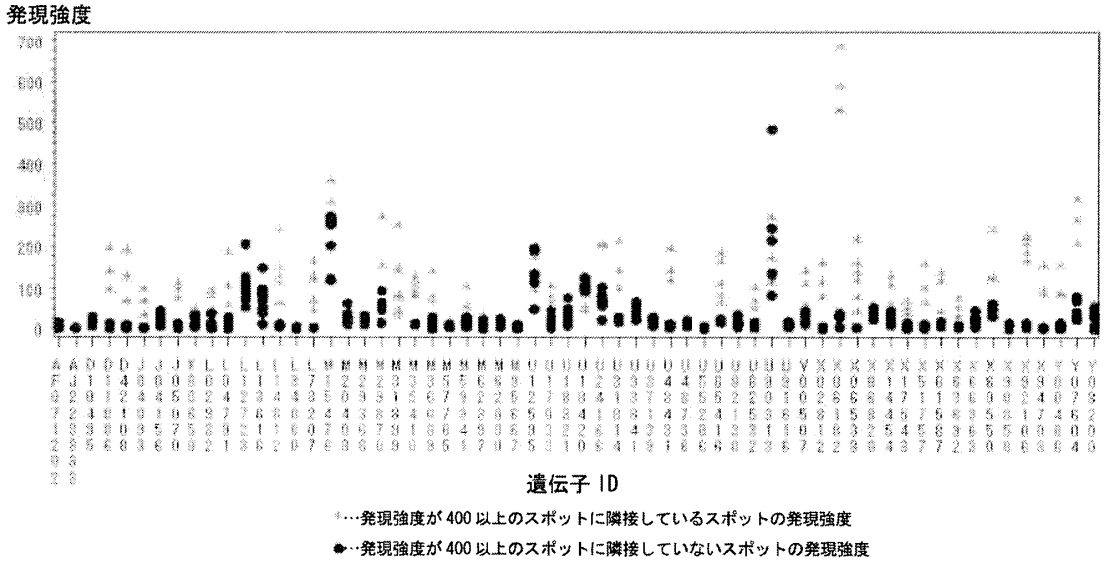

図4 アレイ内で繰り返し測定されている遺伝子の発現強度（3枚の結果をまとめたもの）

合は，その中で最も大きい発現強度のスポットを用いることにした．また関数$f(I_{g'}^{l})$によって得られる値がneighborhood biasの推定値である．本研究では同一遺伝子のアレイ内繰り返し測定を利用して，以下のようにこの関数を得た．まずアレイ内繰り返し測定されている遺伝子の発現強度に含まれているneighborhood biasを，次のように推定する．

$$\hat{E}_{g'}^{l}=I_{g}^{l}-mean(I_{g}^{s}) \qquad (2)$$

$\hat{E}_{g'}^{l}$は遺伝子g'が近接するスポットに及ぼすneighborhood bias，I_{g}^{l}，I_{g}^{s}はそれぞれ，隣接するスポットに発現強度の大きなスポットが存在する，もしくは存在しない遺伝子gの発現強度である．これは，隣接するスポットに影響を受けていないとみなされるスポットの平均値をもってその遺伝子の遺伝子発現とし，影響を受けている発現強度との差をとることでバイアスの推定を行っている．本研究では，x軸に$I_{g'}^{l}$，y軸に$\hat{E}_{g'}^{l}$をプロットし，関数型$f(I_{g'}^{l})$を線形回帰によって求めた．

3.2 分散安定化変換

cDNAマイクロアレイ，cDNAマクロアレイ実験系では，発現強度の値が大きいほどばらつきが大きくなる．これを解消するため，通常は発現強度に対して対数変換が行われることが多いが，対数変換を行うと今度は値の小さい範囲でばらつきが大きくなる．そのため，Huberら(2002)はこの不等分散性を解消するための変換を提案し，竹内（2003）の研究でその変換の良さは示されている．しかし，この分散安定化変換はアレイ効果の除去をパラメータ推定に組み込んでいる．つまり測定バイアスの除去と分散安定化を同時に行うようなモデルを想定しており，データの性質によってはうまく機能しない可能性が考えられる．そのため本研究ではHuberら（2002）の方法を単純化した，簡便な変数変換を提案する．以下では，3.2.1でHuberによる分散安定化変換を簡単に説明し，3.2.2でそれを単純化した変数変換を説明する．

3.2.1 Huberによる分散安定化変換

異なるアレイを用いても同じ試料の同じ遺伝子の遺伝子発現は等しいはずであるから，次の線形変換によってアレイ効果を取り除く．これはアレイ間測定バイアスを取り除くことに相当する．

$$I'_{gk}=o_k+s_kI_{gk} \tag{3}$$

ここで，o_k, s_k はアレイ k に対する線形変換のためのパラメータである．さらに変換後の発現強度に対して，Rocke ら（1995）のモデルを想定する．

$$I'_{gk}=\alpha_k+\mu_{gk}e^{\eta_{gk}}+\nu_{gk} \tag{4}$$

η_{gk} と ν_{gk} は独立に，平均は 0 の正規分布に分布すると仮定し，$e^{\eta_{gk}}$ と ν_{gk} の分散をそれぞれ，S_η^2, σ_ν^2 とする．すると，発現強度の分散が平均値に依存しなくなるような変数変換は次のように表せる．

$$h_k(I'_{gk})=\frac{1}{S_\eta}\mathrm{asinh}\left(-\alpha_k\frac{S_\eta}{\sigma_\nu}+\frac{S_\eta}{\sigma_\nu}I'_{gk}\right) \tag{5}$$

さらに，式 $I'_{gk}=o_k+s_kI_{gk}$ を代入して変数を整理すると次のようになる．

$$h(I_{gk})=\mathrm{asinh}(a_k+b_kI_{gk}) \tag{6}$$

ここで，$a_k=-\alpha_k S_\eta/\sigma_\nu+o_k S_\eta/\sigma_\nu$, $b_k=s_k S_\eta/\sigma_\nu$ と置き換えた．この $(a_1, b_1,..., a_k, b_k)$ を Huber ら（2003）は以下の対数プロファイル尤度関数を L-BFGS-B 法によって最大化することで推定している．

$$\frac{|G'|K}{2}\log\left(\sum_{g\in G'}\sum_k(h_k(I_{gk})-\hat{\mu}_k)^2\right)+\sum_{g\in G'}\sum_k\log h'_k(I_{gk}) \tag{7}$$

ここで，G', $|G'|$ はそれぞれ推定の際に用いる遺伝子の集合と個数，$\hat{\mu}_k$ は $h(I_{gk})$ によって変換した後に推定されるアレイ平均，h'_k は関数 h_k の一次導関数である．集合 G' の初期値を全遺伝子としてパラメータ推定し，推定したパラメータを固定したもとで $\hat{\mu}_k$ からの残差を計算する．残差が小さい遺伝子を集合 G' と再定義しパラメータを再推定することで，外れ値に対して頑健になるように考慮している．具体的には平均強度の大きい順に 10 分割し，それぞれのグループについて残差の大きさが小さいものから q%$(50<q\leq100)$ までの遺伝子を集合 G' とする．この操作をパラメータが収束するまで繰り返し行う．

この解析には統計解析ソフトウェアである R（R Development Core Team, 2005），及び Gentleman ら（2005）による Bioconductor project が提供している vsn 関数を使用した．

3.2.2 逆双曲線正弦変換

本研究では発現強度に対して以下のようなモデルを想定する．

$$I_{gk}=\alpha_k+\beta_{gk}+\mu_ge^{\eta_{gk}} \tag{8}$$

ここで α_k は発現強度に混入しているアレイ間測定バイアスを表し，β_{gk} はアレイ内測定バイアスを表す．また μ_g は遺伝子 g の本来の遺伝子発現であり，η_{gk} はその誤差を乗法的に説明する項である．つまり遺伝子発現が関係している項は $\mu_ge^{\eta_{gk}}$ だけであり，この項は常に正の値を取る．さらに測定バイアス（$\alpha_k+\beta_{gk}$）は常に正であると仮定し，これらを実験データによって推定する．これを式で表すと以下のようになる．

$$\alpha_k+\beta_{gk}=\hat{\alpha}_k+\hat{\beta}_{gk}+\nu_k+\gamma_{gk} \tag{9}$$

ここで，ν_k, γ_{gk}, η_{gk} は独立に平均 0，分散はそれぞれ，σ_ν^2, $\sigma_{\gamma_k}^2$, $\sigma_{\eta_k}^2$ の正規分布に従うと仮定し，測定バイアスの推定値を取り除いた後の発現強度を I'_{gk} とすると，I'_{gk} は（8）式と（9）式よ

り次のようになる.

$$I'_{gk}=I_{gk}-(\hat{\alpha}_k+\hat{\beta}_{gk})=\mu_g e^{\eta_{gk}}+\nu_k+\gamma_{gk} \tag{10}$$

I_{gk} は測定バイアスを含んでいるので常に正の値をとるが，I'_{gk} は測定バイアスの推定値を引いているので，遺伝子発現である μ_g が小さい場合は負の値もとり得る.

ここから本研究では，I'_{gk} に対して分散安定化変換を考える．まず，I'_{gk} の期待値と分散は次のように表すことができる.

$$E[I'_{gk}]=M_k\mu_g,\quad V[I'_{gk}]=D_k^2+\mu_g^2S_k^2 \tag{11}$$

ここで，$M_k=e^{\sigma_{\eta_k}^2/2}$, $D_k^2=\sigma_\nu^2+\sigma_{\gamma_k}^2$, $S_k^2=e^{\sigma_{\eta_k}^2}(e^{\sigma_{\eta_k}^2}-1)$ である．よって，この分散を期待値の関数で表すと次のようになる.

$$V[I'_{gk}]=D_k^2+\left(\frac{S_k}{M_k}\right)^2E[I'_{gk}]^2 \tag{12}$$

このように分散が期待値の関数になっている場合，分散が期待値に依存しなくなるような変換関数はデルタ法によって以下のように導かれる.

$$h(I'_{gk})=\int_{I'_{gk}}\frac{1}{\sqrt{V[E[I'_{gk}]]}}\cdot dE[I'_{gk}] \tag{13}$$

よって上の分散と期待値の関係式から，次の式が導ける.

$$h(I'_{gk})=\frac{M_k}{S_k}\log\left\{\frac{S_k}{D_kM_k}\cdot I'_{gk}+\sqrt{\left(\frac{S_k}{D_kM_k}\cdot I'_{gk}\right)+1}\right\}=\frac{M_k}{S_k}\mathrm{asinh}\left(\frac{S_k}{D_kM_k}\cdot I'_{gk}\right) \tag{14}$$

アレイ毎に分散安定化変換を行うとすると，M_k/S_k は全ての測定値に共通な値なので無視することができる．さらに $S_k/D_kM_k=c_k$ とし，正規化のための変換関数を次のように再定義する.

$$h(I'_{gk})=\mathrm{asinh}(c_kI'_{gk}) \tag{15}$$

正規化を行うためには，この c_k をアレイ毎に推定しなければならない．よって本研究では c_k を，測定バイアスの推定値を取り除いた後のアレイ毎の発現強度平均値 $I'_{\cdot k}$ の逆数とすることを提案する．この変数変換を改めて式で表すと次のようになる.

$$h(I'_{gk})=\mathrm{asinh}\left(\frac{I'_{gk}}{I'_k}\right)=\log\left\{\frac{I'_{gk}}{I'_k}+\sqrt{\left(\frac{I'_{gk}}{I'_k}\right)^2+1}\right\} \tag{16}$$

逆双曲線正弦関数 $\mathrm{asinh}(x)$ は引数の値が 0 から 1 の範囲では直線的な形状となり，その他の範囲では対数に似た形状をもつ．よって c_k を $I'_{\cdot k}$ の逆数とし，この変換を施すことで，測定バイアスの推定値を取り除いた後の発現強度がアレイ内平均値以下である遺伝子は変数変換前と同等な挙動を示し，アレイ内平均値より大きい遺伝子は対数変換を施したものと同等の挙動をとる（図5）．一般にアレイ実験系は極端に大きく発現している遺伝子がいくつかあり，平均値がそれらの影響で大きくなってしまう可能性があるが，敢えて平均値を使うことにより，遺伝子発現が密集している範囲を直線的に変換できると考えた．これによって変数変換を行わない発現強度と対数変換を行った発現強度の，それぞれの欠点を補うことができると考えられる．また測定バイアスの推定値を取り除いた後の発現強度が負の値である場合は対数変換を行うこ

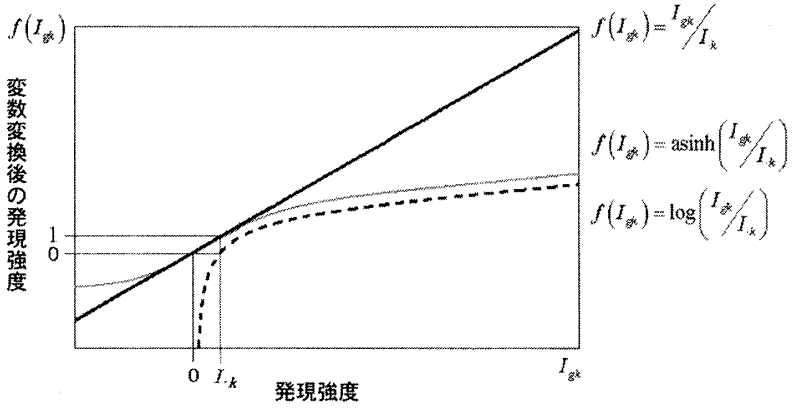

図 5　発現強度と各変数変換との関係のイメージ

とはできないが，逆双曲線正弦変換では負の値であっても計算できる．

4. 正規化手法の評価

4.1 測定バイアスを取り除く正規化

図 6 に neighborhood bias と発現強度との関係を示す．図 6 は，（2）式によって求めた neighborhood bias と発現強度との関係を，3 枚のアレイデータに関してまとめて示したものである．この図から発現強度が 6000 を超える 3 点は外れ値だと考えられ，かつ neighborhood bias は発現強度が大きくなると線形的に増加するのではないかと考えられる．そこで，発現強度が 6000 を超える値を除いて線形回帰を行ったものが図 7 であり，この回帰式に従って neighborhood bias を取り除いた．また外れ値と考えた 3 点に隣接するスポットは，neighborhood bias の推定値（$\hat{E}_g^l$）の 3 つの平均値である 580 を，発現強度から引いた．

また，本研究では測定バイアスを取り除く正規化手法を評価するために，アレイ間 MSE とアレイ内 MSE を利用した．Park ら（2002）がアレイ間 MSE を評価基準として用いることを提案しており，本研究ではこれに加えて，アレイ内繰り返し測定を利用して計算したアレイ内 MSE も評価基準とした．MSE は真の分散とバイアスの 2 乗の和であるので，正規化によって

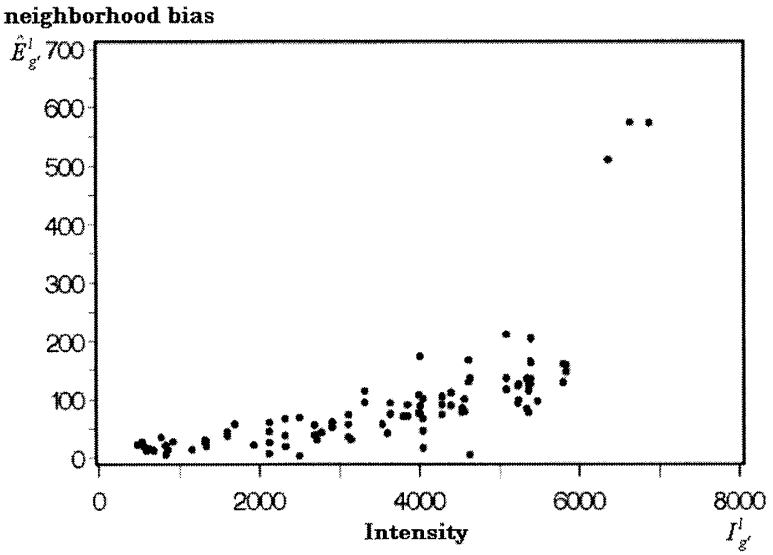

図 6　Neighborhood bias と発現強度との関係

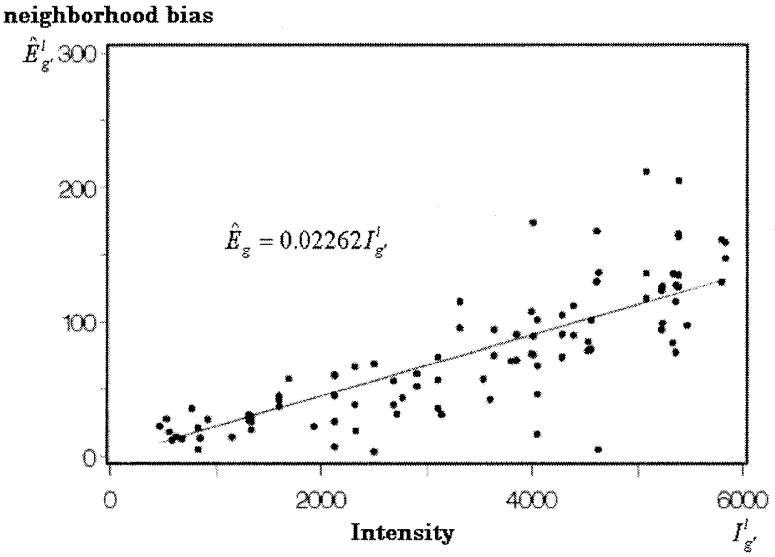

図7 Neighborhood bias と発現強度との関係に回帰直線を当てはめたもの
（外れ値と考えた3点を取り除いた）

測定バイアスを取り除くことができれば，各 MSE は小さくなる．まず遺伝子毎の各 MSE を分散の推定値と考え，次のように表現する．

$$\begin{cases} \text{遺伝子毎のアレイ間 MSE}\cdots\hat{\sigma}_g^2=\dfrac{1}{K-1}\sum_k(I_{gk\cdot}-I_{g\cdot\cdot})^2 \\ \text{遺伝子毎のアレイ内 MSE}\cdots\hat{\sigma}_{gk}^2=\dfrac{1}{R-1}\sum_r(I_{gkr}-I_{gk\cdot})^2,\ r\neq 1 \end{cases} \tag{17}$$

$$where \quad I_{gk\cdot}=\frac{1}{R}\sum_r I_{gkr}, \quad I_{g\cdot\cdot}=\frac{1}{K}\sum_k I_{gk\cdot},$$
$$g=1,\ldots,G, \quad k=1,\ldots,K, \quad r=1,\ldots,R$$

r はアレイ内繰り返し番号を表す．そして本研究ではアレイ内繰り返し数によって分散の推定精度が異なると考えた．そのため遺伝子をアレイ内繰り返し回数によって分類し，アレイ間 MSE，アレイ内 MSE 共にそれらの中で MSE の中央値を求めた．これらの値は以下のようになり，これを評価基準とした．

$$\begin{cases} \text{アレイ間 MSE}\cdots\hat{\sigma}_r^2=\underset{g\in Gr}{median}(\hat{\sigma}_g^2) \\ \text{アレイ内 MSE}\cdots\hat{\sigma}_{kr}^2=\underset{g\in Gr}{median}(\hat{\sigma}_{gk}^2),\ r\neq 1 \end{cases} \tag{18}$$

ただし G_r はアレイ内繰り返し数が r 回である遺伝子の集合である．またアレイ内繰り返し測定が無い場合はアレイ間 MSE しか計算できず，この際の評価基準は以下のようになる．

$$\text{アレイ間 MSE}\cdots\hat{\sigma}^2=\underset{g}{median}(\hat{\sigma}_g^2), \quad \hat{\sigma}_g^2=\frac{1}{K-1}\sum_k(I_{gk}-I_{g\cdot}) \tag{19}$$

Park ら（2002）は中央値ではなく平均値を用いているが，本研究では外れ値を考慮して中央値を用いた．表1と表2に，アレイ間 MSE とアレイ内 MSE の結果を示す．まずアレイ間 MSE は，background noise を利用した正規化ではどの MSE の値も減少したが，Global 正規化や neighborhood bias を取り除く正規化では必ずしもそうではなかった．特に Global 正規化で

表1　アレイ間 MSE

正規化手法	アレイ内繰り返し測定回数		
	1回	2回	3回
生データ	36.80	48.56	74.00
background noiseを利用した正規化	10.75	9.50	68.68
Global正規化	10.61	21.64	88.45
neighborhood biasを取り除く正規化	36.66	52.23	88.45

表2　アレイ内 MSE

正規化手法	アレイ内繰り返し測定回数					
	アレイ1		アレイ2		アレイ3	
	2回	3回	2回	3回	2回	3回
生データ	81.18	574.64	56.70	1488.99	21.58	1113.66
background noiseを利用した正規化	88.18	574.64	48.10	1488.99	21.58	1082.99
Global正規化	88.18	574.64	48.50	1272.93	24.34	1255.72
neighborhood biasを取り除く正規化	58.10	93.52	35.10	144.74	21.58	190.38

は，散布図行列によってアレイ 2 とアレイ 3 がアレイ 1 に比べて発現強度が大きいと予想されたにもかかわらず，補正の際に用いた中央値がそれぞれ 73.46，79.78，69.18 とアレイ 1 よりアレイ 3 の方が小さくなっていたため，アレイ間測定バイアスを適切に取り除くことができなかったと考えられる．次にアレイ内 MSE は，neighborhood bias を取り除く正規化で値が著しく減少した．また background noise を利用した正規化や Global 正規化ではあまり変化しないか，増加しているものもあった．さらに各正規化手法を組み合わせると，アレイ間 MSE とアレイ内 MSE 共に，組み合わせている正規化手法の性質を併せ持った結果となった．以上の結果から，本研究のデータでは background noise を利用した正規化と neighborhood bias を取り除く正規化を適用することで測定バイアスを除去できると考え，この順番で適用した．

4.2　分散安定化変換

本研究では，対数変換と Huber による分散安定化変換と逆双曲線正弦変換の比較，評価を行った．これらの変数変換は測定バイアスを取り除いた後に行った．評価基準としては，%Mean-SD plot，divided %Mean-SD plot，第 1 主成分の固有値割合を用いた．以下にこれらの評価基準の説明を行う．%Mean-SD plot は x 軸に発現強度のアレイ間平均値を，y 軸に発現強度のアレイ間標準偏差を遺伝子毎にプロットした Mean-SD plot を改良したものである．y 発現強度の大小に関わらず分散が均一になっていれば，標準偏差も均一になっており Mean-SD plot は平坦なものになる．そこで本研究では異なる変数変換を行った結果の比較を可能とするために，Mean-SD plot の各軸を，遺伝子毎に計算した発現強度平均値の最大値を基準とした割合で表す．これを%Mean-SD plot と呼び，各軸は以下のように表せる．

$$\begin{aligned} x\text{ 軸：}\%Mean &= \frac{I_{g\cdot}}{\max_g(I_{g\cdot})}\times 100, \\ y\text{ 軸：}\%SD &= \frac{\sqrt{\dfrac{1}{K-1}\sum_k (I_{gk}-I_{g\cdot})^2}}{\max_g(I_{g\cdot})}\times 100 \end{aligned} \tag{20}$$

正確に言えば%SD は割合ではないが，簡便のためこのように表現する．この%SD は変動係数（CV）に近い指標になっているが，分母にどの遺伝子にも共通な値を用いているので CV ではない．y 軸に CV を用いるとプロットは右上がりの直線となるため，分散が均一になっているかどうかを視覚的に確認することが困難であると思われる．またアレイ内繰り返し測定がされている遺伝子は，アレイ内平均値をそのアレイでの発現強度とみなして，アレイ間平均値

とアレイ間標準偏差を計算した．

しかし%Mean-SD plotでは，プロットがある値付近に集中している場合やプロットにあまり変化が見られない場合が多い．そのような場合は分散の均一性を確認することは難しく，以下のdivided %Mean-SD plotを評価基準とすることを提案する．まず%Mean-SD plotをx軸に関して10分割し，それぞれの範囲での%Meanと%SDの平均値をその区画での代表値とする．そしてその10個の値（分割した区画に値が存在しない場合は点の数は10個より少なくなる）をプロットする．本研究ではこの図をdivided %Mean-SD plotと呼ぶ．この図によって発現強度と分散の平均的な挙動を確認することができる．

さらに，分散安定化変換を定量的に比較する指標として，積和行列の主成分分析から計算できる第1主成分の固有値割合を用いる．このとき，遺伝子をオブザベーション，アレイを変数としたG×K行列をデータ行列として積和行列を計算する．Kohase et al.（2004）によるとどんな実験系においても，ほとんどの遺伝子の発現量は変わらないという．そのため第1主成分の固有値割合の大きい正規化手法ほど，アレイ間のばらつきが小さく分散も均一になっており，より良い正規化手法と解釈することができると思われる．

図8と図9に%Mean-SD plotとdivided %Mean-SD plotの結果を，表3に第1主成分の固有値割合の結果を示す．また，補助的に散布図も図10に示す．散布図のプロットが直線$y=x$上に乗り，かつばらつきが小さく均一であるほど良い．%Mean-SD plotとdivided %Mean-SD plotを見ると，変数変換前の発現強度は値の大きい範囲でばらつきが大きく，対数変換を施した発現強度では逆に値の小さい範囲でばらつきが大きくなっていることがわかる．またHuberによる分散安定化変換は，%Mean-SD plotやdivided %Mean-SD plotでは分散は均一になっていると思われるが，散布図行列を見ると外れ値がいくつか出現している．これらの値は分散安定化変換を施す前は負の値をとっており，それが原因で外れ値となったと考えられる．さらに，Huberによる分散安定化変換では%Meanが0〜30付近の値を示しているものが全く無いことから，発現強度がほとんど0であった遺伝子もある程度の値をとるような変換になっているこ

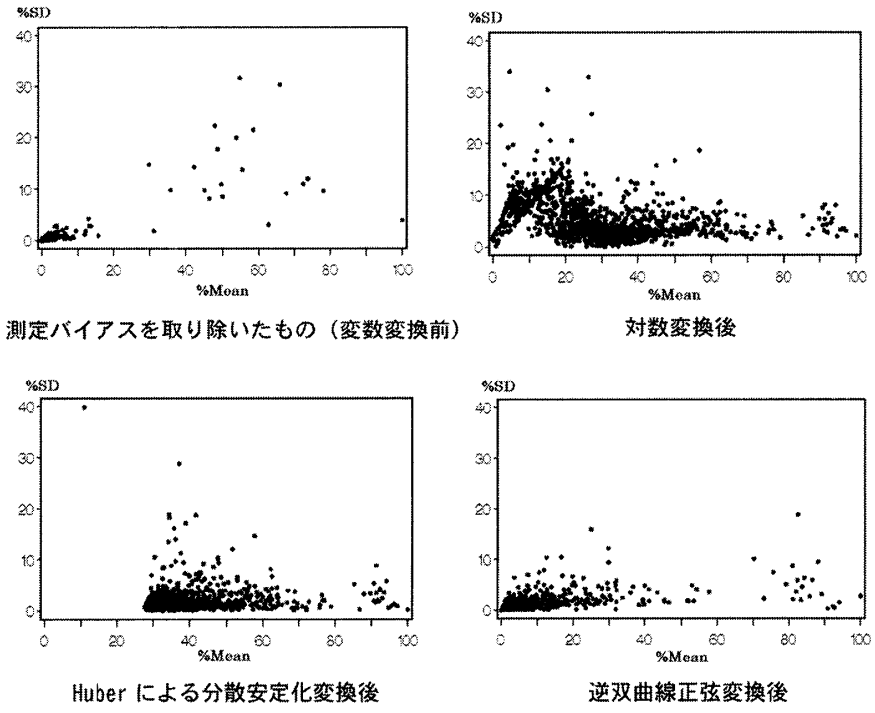

図8 %Mean-SD plot

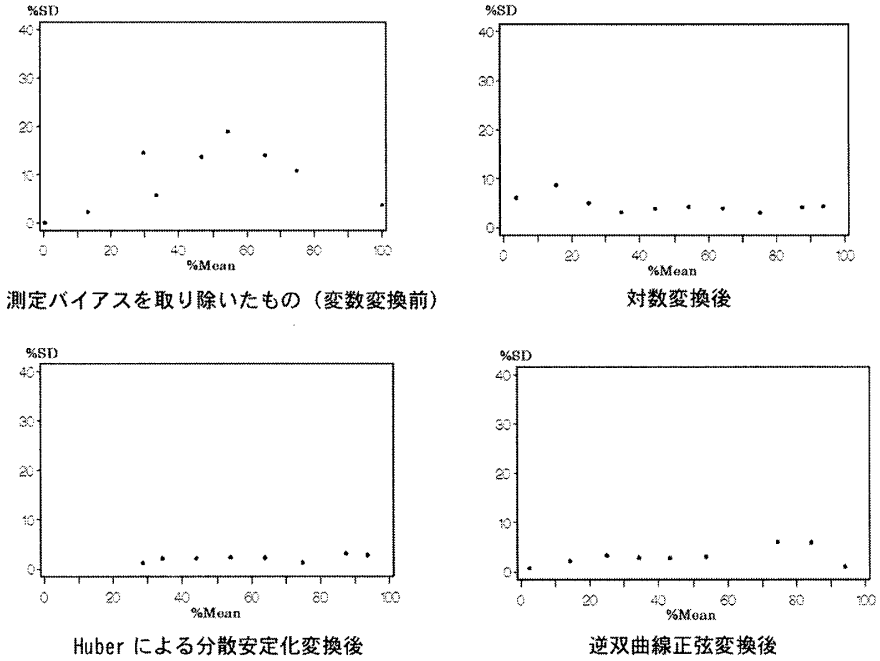

図 9　divided %Mean-SD plot

表 3　第 1 主成分の固有値割合

正規化手法	第1主成分の固有値割合
測定バイアスを取り除く前のデータ	0.9716
測定バイアスを取り除いた後のデータ	0.9702
対数変換	0.9838
Huberによる分散安定化変換	0.9851
逆双曲線正弦関数による変換	0.9900

とが分かる．この結果は，本研究に特異的なものである可能性はあるが，このような変数変換が正規化に適しているかどうかには疑問が残る．その点，逆双曲線正弦変換は分散が均一になっており，変換を施した結果も妥当な値であると思われる．さらに主成分分析における第 1 主成分の固有値割合で比較しても，わずかではあるが逆双曲線正弦変換が大きくなっている．以上の結果を総合して，本研究で用いたデータでは，逆双曲線正弦変換によって分散を均一にできると判断した．

5. 考　　察

本研究では，散布図行列，アレイ間 MSE，アレイ内 MSE，%Mean-SD plot，divided %Mean-SD plot，第 1 主成分の固有値割合を総合的に検討することにより，提案した正規化手法の評価を行った．この評価基準について，測定バイアスの評価としてアレイ間 MSE を用い，不等分散性の評価として%Mean-SD plot や第 1 主成分の固有値割合のみを用いるという単純な評価方法も考えられる．しかしアレイ間 MSE とアレイ内 MSE を共に評価することによって，より詳細にアレイ内測定バイアスの程度を知ることができることが本研究により示唆された．また，Huber による分散安定化変換は%Mean-SD plot や第 1 主成分の固有値割合だけで評価すると良い変換であると思われたが，散布図行列を確認すると非常に歪んだ結果となっており，とても正規化の役割を果たしているとは考えられなかった．したがって単純な評価基準だけではなく，散布図行列と共に評価を行うことが非常に重要であると思われる．

本研究で用いた cDNA マクロアレイでは，発現強度の最も大きいスポットであっても近接

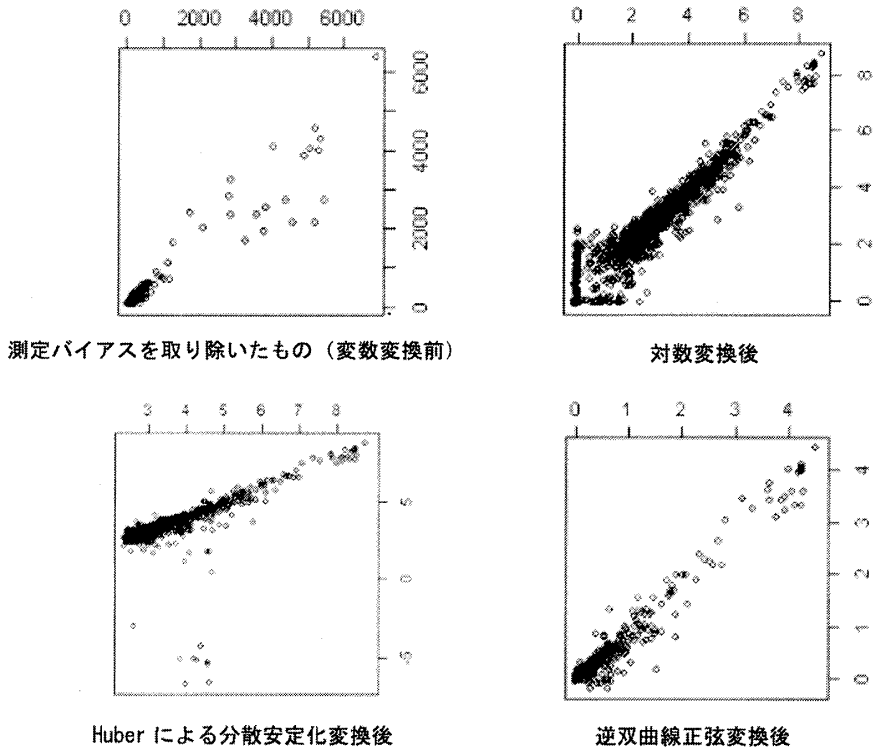

図 10　散布図行列の一部（x 軸：array 2 の発現強度，y 軸：array 1 の発現強度）

するスポットへの影響はそれほど大きいものではなく，少し霞がかっている程度であった．そのため neighborhood bias を推定し，取り除くことが可能であった．しかし発現強度があまりにも大きすぎるために，隣のスポットを完全に覆い尽くしてしまうといった場合もあり得る．このような場合は neighborhood bias を推定することはできず，本研究で用いた正規化を行うことはできない．そのため，あらかじめ Querec ら（2004）の方法などによってそのような事態にならないように工夫を行うことが必要であると思われる．Querec の方法はフィルムへの放射線の曝露時間とスポットの発現強度の関係を事前に測定しておき，最も有効であると思われる曝露時間を決定しておくというものである．また，今回使用したデータには background noise が測定されていなかったため，アレイ内繰り返し測定を利用して neighborhood bias を推定した．しかし，background noise が測定されている場合は，Schuchhardt ら（2000）の方法で neighborhood bias を推定できる可能性がある．

また，本研究の実験データでは遺伝子の同一アレイ内繰り返し測定がされていたので，これを利用することでアレイ内 MSE を計算した．しかし本研究では，アレイ内 MSE とアレイ間 MSE はアレイ内の繰り返し測定数が多い遺伝子のものほど大きいという結果となった．このことは発現強度に入っているスポット効果が大きいことを示唆している．また繰り返し測定を利用して neighborhood bias を取り除く正規化を行うこともでき，Fan ら（2004）も，アレイ内繰り返し測定遺伝子を利用して測定バイアスを推定することを提案している．そのため測定バイアスを定量的に評価する場合や，cDNA マクロアレイのようにどのような測定バイアスが入るか予想できるような場合は，アレイ作成の段階であらかじめアレイ内繰り返し測定ができるように計画を行うべきであろう．

Huber による分散安定化変換は，変換前の値が負であると変換後の値が負に大きな値となってしまうことが示唆された．このような現象が起こる原因は定かではないが，パラメータ推定を行う際の反復計算の段階で，推定値が望ましくない結果となっている可能性がある．その点，

逆双曲線正弦変換は単純な変数変換であるため，このような問題はほとんど無いと思われる．さらに Huber の分散安定化変換と逆双曲線正弦変換の最大の違いは，前者は測定バイアスの除去と分散安定化を同時に行っているのに対し，後者は測定バイアスを取り除いた後に分散安定化変換を行っている点である．この観点からみると，遺伝子発現解析における Kerr ら（2001），Dobbin ら（2002），Dobbin ら（2003），Wolfinger（2001）などの発現解析の方法にも異なった見解があると考えることができるかもしれない．Kerr や Dobbin はアレイ効果や遺伝子発現，Dye の効果，さらにそれらの交互作用を同時にモデル化しているのに対し，Wolfinger は発現強度に対してまずアレイ効果や治療効果をモデル化し，次にそれらの効果からの残差に対して遺伝子発現や交互作用をモデル化するといった 2 段階の方法となっている．どちらの方法が優れているかを一概に判断することはできないが，同時に全てを推定するよりもそれぞれ別々に推定した方が安定した推定値を得ることができ，計算負荷も小さい．

今後，cDNA マクロアレイ以外の DNA アレイを用いて正規化手法の評価を行う場合でも，本研究で用いた評価基準を応用することは容易に可能である．さらに，新たな正規化手法が提案され既存の正規化手法と比較する際や，本研究で比較していない正規化手法を比較する際にもこれらの評価基準を用いることで，正規化手法を適切に比較することができると思われる．特に発現強度の不等分散性は cDNA マクロアレイと cDNA マイクロアレイに共通な問題点なので，これを簡便に解消できる逆双曲線正弦変換が他の実験系においても有効であることが期待される．

最後に，根本的な問題として mRNA の発現量が最終的に生成される蛋白質量を反映していないのではないかという議論がなされおり，Stoughton ら（2005）などが DNA アレイの意義について疑問視する見方をしていることを指摘しておく．しかし Hughes ら（2000）が述べているように，遺伝子発現量はその遺伝子によって作られる蛋白質量の代替指標にはならないものの，個々の蛋白質の機能を修飾しているのではないかという意見もある．したがって，DNA アレイという実験系は今後も有用であり解析方法等のさらなる研究が必要であると思われる．

謝辞

本稿は大変有用な査読者のコメントによって大きく前進することが出来た．これについて心からお礼を申し上げたい．本研究は科学研究費補助金基盤研究（A）No. 16200022 の援助を受けた．

参 考 文 献

Argyropoulos, C., Chatziioannou, A. A., Nikifordis, G., et al. (2006). Operational criteria for selecting a cDNA microarray data normalization algorithm, *Oncology Reports*, **15**, 983–996

Berger, J. A., Hautaniemi, S., Jarvinen, A. K., et al. (2004). Optimized LOWESS normalization parameter selection for DNA microarray data, *BMC Bioinformatics*, **5**, 194. 1–13.

Bolstad, B. M., Irizarry, R. A., Åstrand, M., et al. (2003). A comparison of normalization methods for high density oligonucleotide array based on variance and bias, *Bioinformatics*, **19**, 185–193.

Cheadle, C., Vawter, M. P., Freed, W. J., et al. (2003). Analysis of microarray data using Z sore transformation, *Journal of Molecular Diagnostics*, **5**, 73–81.

Colantuoni, C., Henry, G., Zeger, S., et al. (2002). Local mean normalization of microarray element signal intensities across an array surface: quality control and correction of spatially systematic artifacts, *Biotechniques*, **32**, 1316–1320.

Dobbin, K. and Simon, R. (2002). Comparison of microarray designs for class comparison and class discovery, *Bioinformatics*, **18**, 1438–1445.

Dobbin, K., Shih, J. and Simon, R. (2003). Statistical design of reverse dye microarrays, *Bioinformatics*, **19**, 803–810.

Duggan, D. J., Bittner, M., Chen, Y., et al. (1999). Expression profiling using cDNA microarrays, *Nature Genetics*, **21**, 10-14.

Edwards, D. (2003). Non-linear normalization and background correction in one-channel cDNA microarray studies, *Bioinformatics*, **19**, 825-833.

Fan, J., Tam, P., Woude, G. V., et al. (2004). Normalization and analysis of cDNA microarrays using within-array replications applied to neuroblastoma cell response to a cytokine, *Proceedings of the National Academy of Sciences*, **101**, 1135-1140.

Folder, S. P. A., Read, J. L., Pirrung, M. C., et al. (1991). Light-directed, spatially addressable parallel chemical synthesis, *Science*, **251**, 767-773.

Futschik, M. and Crompton, T. (2004). Model selection and efficiency testing for normalization of cDNA microarray data, *Genome Biology*, **5**, R0060. 1-20.

Geller, S. C., Gregg, J. P., Hagerman, P., et al. (2003). Transformation and normalization of oligonucleotide microarray data, *Bioinformatics*, **19**, 1817-1823.

Gentleman, R., Carey, V., Huber. W., et al. (2005). *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, New York: Springer.

Hoffmann, R., Seidl, T. and Dugas, M. et al. (2002). Profound effect of normalization on detection of differentially expressed genes in oligonucleotide microarray data analysis, *Genome Biology*, **3**, R0033. 1-11

Holloway, A. J., Laar, R. K., Tothill, R. W., et al. (2002). Options available—from start to finish—for obtaining expression data by microarray, *Nature Genetics Supplement*, **32**, 481-489.

Huber, W., Heydebreck, A., Sultmann, H., et al. (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression, *Bioinformatics*, **18**, 96-104.

Huber, W., Heydebreck, A., Sultmann, H., et al. (2003). Parameter estimation for the calibration and variance stabilization of microarray data, *Statistical Applications in Genetics and Molocular Biology*, **2**, Article3. 1-22.

Hughes, T. R., Marton, M. J. Jones, A. R., et al. (2000). Functional discovery via a compendium of expression profiles, *Cell*, **102**, 109-126.

Inoue, M., Nishimura, S., Hori, G., et al. (2004). Improved parameter estimation for variance-stabilizing transformation of gene-expression microarray data, *Journal of Bioinformatics and Computational Biology*, **2**, 669-679.

Kerr, M. K. and Churchill, G. A. (2001). Statistical design and the analysis of gene expression microarray data, *Genetic Research*, **77**, 123-128.

Kohane, I. S., Kho, A. T. and Butte, A. J. 著, 黒田有人 訳. (2004), 統合ゲノミクスのためのマイクロアレイデータアナリシス, シュプリンガー・フェアラーク 東京.

Konishi, T. (2004). Three-parameter lognormal distribution ubiquitously found in cDNA microarray data and its application to parametric data treatment, *BMC Bioinformatics*, **5**, 5.

Park, T., Yi, S., Kang, S., et al. (2003). Evaluation of normalization methods for microarray data, *BMC Bioinformatics*, **4**, 33.1-33.13.

Querec, T. D., Stoyanova, R., Ross, E., et al. (2004). A novel approach for increasing sensitivity and correcting saturation artifacts of radioactively labeled cDNA arrays, *Bioinformatics*, **20**, 1955-1961.

R Development Core Team. (2005). *R: A Language and Environment for Statistical Computing*, Vienna: R Foundation for Statistical Computing.

Rocke, D. M. and Lorenzato, S. (1995). A tow-component model for measurement error in analytical chemistry, *Technometrics*, **37**, 176-184.

SAS Institute Inc. (2004). *SAS/STAT* 9.1 *User's Guide, Cary*, NC: SAS Institute Inc.

Schena, M., Shalon, D., Davis, R. W., et al. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray, *Science*, **270**, 467-470.

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

Simon, R. M., Korn, E. L., McShane, L. M., et al. (2003). *Design and analysis of DNA microarray investigations*, New York: Springer.

Stoughton, R. G. (2005). Applications of DNA microarrays in biology, *Annual Review of Biochemistry*, **74**, 53-82.

竹内正弘. (2003). フィルターアレイの信頼性および抗癌剤感受性遺伝子に関する研究, 厚生労働科学研究費補助金分担研究報告書.

Uchida, S., Nishida, Y., Satou, K., et al. (2005). Detection and Normalization of biases present in spotted cDNA mcroarray data: a composite method addressing dye, intensity-dependent, spatially-dependent, and print-order

biases, *DNA Research*, **12**, 1–7.

Wolfinger, R. D., Gibson, G., Wolfinger, E. D., et al. (2001). Assessing gene significance from cDNA microarray expression data via mixed models, *Journal of Computational Biology*, **8**, 625–638

Workman, C., Jensen, L. J., Jarmer, H., et al. (2002). A new non-linear normalization method for reducing variability in DNA microarray experiments, *Genome Biology*, **3**, R0048. 1–16.

Yang, Y. H., Dudoit, S., Luu, P., et al. (2001). Normalization for cDNA microarray data, SPIE BiOS, Available at: http://www.stat.berkeley.edu/users/terry/zarray/Html/papersindex.html. Accessed January 18, 2006.

Yoon, D., Yi, S. G., Kim, J. H., et al. (2004). Two-stage normalization using background intensities in cDNA microarray data, *BMC Bioinformatics*, **5**, 97. 1–12.