目次

抄録

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光強度には様々な測定バイアスが含まれるため、正規化によってこれを取り除かなくてはならない。また、測定された発光強度が本質的に持つ不等分散性も正規化によって解消する。正規化には様々な方法が提案されており、これら正規化手法を比較する論文はいくつかあるが、発光強度の不等分散性を解消するための正規化手法を比較する論文はない。このような現状であると、実際に得られたデータを正規化する際に、どの正規化を採用すればよいか決定できない。

そのため本研究では、正規化手法を比較し選択していくときに踏むべきプロセスを明示し、実際の実験データに当てはめた。また、新たな正規化手法を提案し、既存の正規化手法との比較を行った。比較基準としては、散布図行列、アレイ間誤差分散とアレイ内誤差分散、Mean-SD プロット、第 1 主成分の固有値割合を用いた。その結果この実験系では、cDNA マクロアレイにおいて発光強度の大きいスポットが近接するスポットに及ぼすバイアスを取り除く正規化、及び発光強度の分散を安定化するための、逆双曲線正弦関数による正規化の性能が良いことが示された。

本研究では、遺伝子のアレイ内繰り返し測定を利用することで正規化手法の比較と測定バイアスの推定を行うことができた。そのため、今後の実験においても遺伝子のアレイ内繰り返しは必要である。また今後新たな手法が提案された場合でも、本研究と同様のプロセスを踏み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出来ると思われる。

Key words ： DNA アレイ　cDNA マクロアレイ　正規化　測定バイアス
分散の安定化

1. 緒言

近年、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[1,2]。DNA アレイ技術には様々な種類があり、一枚のアレイの中にスポットされている遺伝子数が比較的多いものを cDNA（complementary DNA）マイクロアレイ、比較的少ないものを cDNA マクロアレイという[3,4,5]。これらの DNA アレイ技術に共通な実験手順は 1)アレイの作成、2)試料の準備とアレイへのハイブリダイゼーション、3)スキャニングとデータ解析の 3 つに分けることができる[3,6]。これらの手順の概要は以下の通りである。ガラスまたはナイロンメンブレンなどのスライド上に、オリゴヌクレオチドまたは cDNA をスポットする。そしてそこへ染料もしくは放射性同位元素によってラベル付けした遺伝子をハイブリダイズさせ、ラベルの発光強度を測定するというものである[1,2,7]。

ここで、遺伝子発現量とラベルの発光強度が比例関係にあるという暗黙の仮定の上で、測定されたラベルの発光強度を遺伝子発現量の代替指標としている。しかし、測定される発光強度には DNA アレイ実験を行う過程で様々な測定バイアスが入ってしまうという問題がある。測定バイアスが入る可能性のある実験仮定の例を挙げると、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、スキャナによる発光強度の測定などである[8,9]。そこで、解析を行う前段階の準備として正規化という操作によってこれらの測定バイアスを取り除く必要がある[4]。

まず考えられる測定バイアスは cDNA マイクロアレイに入る染料によるバイアスであり、このバイアスを取り除くために中央値を用いた総正規化が行われてきた[7]。しかし、染料バイアスが発光強度に依存するという発光強度バイアスが発見され、総正規化では十分にこのバイアスを取り除けないため非線形な正規化が考案された。まず考案された正規化は LOWESS 正規化であり[10]、その後スプライン関数による正規化も考案された[11]。また、Bolstad らは中央値を用いた総正規化の代わりに、中央値の中央値を基準値として用いる正規化手法を用いている[12]。そして DNA アレイについての研究が進むにつれ、染料バイアス以外の様々な測定バイアスが指摘され、それを取り除く正規化手法が考案されてきた。これまでに提案されてきたもののうち、アレイ内のバイアスを除くものとしては、空間的なバイアスを取り除く正規化[13]、アレイにプリントしたチップ群でのプリント順によって起こるバイアスを取り除く正規化[14]、プリントチップ群ごとの発光強度バイアスを取り除く正規化[15]などがある。またアレイ間のバイアスを除くものとして、線形なバイアスを取り除くための正規化や[12]、非線形なバイアスを取り除く正規化[12,16]、バックグラウンドノイズを取り除く正規化[16,17]などがある。さらに、cDNA マクロアレイにおいては、ラベルとして使われている放射性同位元素が少量でも検出できるといった利点がある反面、遺伝子の発現量が大きいと発光強度が非常に強くなるという欠点もある。そのため、近接するスポットに影

響を及ぼしてしまうというバイアスも存在し[18,19]、このバイアスを取り除くための正規化手法が提案されている[8]。しかし、この正規化手法は遺伝子ごとに測定されたバックグラウンド値を使っているので、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、測定された発光強度は分散が均一でないことが問題視されている。それまではデータの分散を安定化するため一般的に対数変換が使われてきたが、これに代わる正規化として、他の変数変換による正規化手法が考案されてきた[20,21]。逆双曲線正弦関数による変換や[22,23]、Z スコアによる変換[24]などである。

このように様々な正規化手法が提案されていると、どの正規化手法を選択すればよいかという問題が上がってくる。この問題解決のために、測定バイアスを除去するための正規化手法を比較、選択している論文はいくつか存在する[4,12,25]。しかし、分散を安定化するための正規化手法を含めて比較している論文はまだ無い。正規化手法の選択はその後の解析に多大な影響を及ぼすため[26]、この問題は非常に重要であり、何らかの選択基準を確立しなくてはならない。

2. 目的

本研究では、ある特定の実験系を対象として、近接するスポットによるバイアスを取り除く正規化手法と発光強度の分散を安定化するための正規化手法を提案する。また同時に、他の実験系にも適用できることを目指し、正規化手法を測定バイアスと分散の均一性の観点から比較をするための評価基準を提案する。

3. 方法

3-1 比較する正規化手法

3-1-1 測定バイアスを取り除くための正規化手法

まず、測定されたデータにどのようなバイアスが入り得るかを確認する。その後、それらの測定バイアスを正規化によって順次取り除いていく。これまでは一般的に対数変換を施した後に測定バイアスを取り除いていた[27,28,29]。しかし、本研究では分散の安定化を図るため対数以外の関数による変数変換も行うので、対数変換を行う前に測定バイアスを取り除く。

今回比較する手法は下記の 3 つの手法で、**1** と **2** はこれまでに提案されている手法、**3**

は本研究で提案する手法である。この 3 つの正規化手法のうちどの正規化手法が良いか、またはどの正規化手法の組み合わせが良いかを検討する。

１．バックグラウンド正規化

スキャナで遺伝子の発光強度を測定する際に入るバックグラウンドノイズを取り除くために、いくつかの正規化手法が提案されている。例を挙げると広い範囲からバックグラウンドノイズを推定する方法や[30]、ベイズ流の方法などがある[31]。しかし、本研究では簡便で十分に有用であると思われる以下の方法によってバックグラウンドノイズを取り除くことを考える[16]。

$$N_{gk}=\begin{cases} I_{gk}-I_k^b & if \quad I_{gk}-I_k^b>1 \\ \exp[1-(I_k^b+1)/I_{gk}] & otherwise. \end{cases} \tag{1}$$

ここで、ここで、g は遺伝子 $\mathrm{ID}(g=1,...,G)$、k はアレイ $\mathrm{ID}(k=1,...,K)$、I は測定された発光強度、N_{gk} は正規化した後の発光強度、I_k^b はバックグラウンドノイズである。発光強度にはバックグラウンドノイズが含まれており、バックグラウンドノイズは通常スポットごとやアレイごとに測定されている。そのため、測定された発光強度からバックグラウンドノイズを引くことで、そのスポットの本来の発光強度を推定できる。また、$I_{gk}-I_k^b$ の値が小さくなる場合でも、上記の式によって補正すれば負の値とならないため、後の変数変換の際に欠損値とならない。なお、$\exp[1-(I_k^b+1)/I_{gk}]$ は I_{gk} に関して単調増加関数である[16]。

２．総正規化

まず､染料によるバイアスやアレイ間の測定バイアスを M-A プロットによって確認する。M はあるアレイとあるアレイの同一スポットの測定値の差（もしくは染料間の測定値の差）であり、A は測定値の平均値である。測定バイアスが無い場合は、値は直線 $y=0$ の周りに平坦にばらつく。測定バイアスが有る場合は、値は測定バイアスに応じて傾いてばらつく。この図によって測定バイアスが確認され、かつ線形な場合は総正規化によって取り除く。本論文では、Bolstad らの総正規化手法を用いた[12]。式で表すと以下のようになる。

$$N_{gk}=\frac{I_{gk}}{\alpha_k}$$
$$where \quad \alpha_k=\frac{\tilde{I}_k}{\tilde{\tilde{I}}}, \quad \tilde{I}_k=\underset{g}{median}(I_{gk}), \quad \tilde{\tilde{I}}=\underset{k}{median}\left(\tilde{I}_k\right) \tag{2}$$

ここで頑健性を考慮し、$\tilde{I}_k = \underset{g}{median}(I_{gk})$ を計算する際に発光強度の大きいものと小さいものを 2%ずつ除いた。

３．近接するスポットによるバイアスを取り除く正規化

放射性同位元素をラベルとして使っている cDNA マクロアレイでは、あるスポットの発光強度が非常に強いため、隣のスポットに影響を及してしまうというバイアスが存在する[8,18,19]。この測定バイアスを次のような式によって取り除くことを提案する（このバイアスはどのアレイにも共通であると考えたので添字 k は除いた）。

$$N_g = I_g^l - f(I^l) \tag{3}$$

ここで、I_g^l は四方のいずれかに発光強度の大きいスポット（本研究では発光強度が 400 以上のスポット）があるスポットの発光強度であり、I^l は隣にある発光強度の大きいスポットの発光強度である。発光強度が大きいスポットがあるスポットの四方に 2 個以上存在する場合は、その中で最も大きい発光強度のスポットを用いた。この関数型 $f(I^l)$ を得るためには、アレイ内で同一遺伝子の繰り返し測定が必要である。アレイ内で同一遺伝子の繰り返し測定を利用して、発光強度の大きいスポットが隣のスポットに及ぼす効果を次のように推定する。

$$\hat{E}_g^l = I_{gr}^l - \underset{r}{mean}(I_{gr}^s) \tag{4}$$

ここで、r はアレイ内の遺伝子繰り返し測定であり $(r = 1,...,R)$、I_{gr}^l は遺伝子 g の繰り返し測定のうち、隣に発光強度の大きいスポットがあるスポットの発光強度、I_{gr}^s は遺伝子 g の繰り返し測定のうち、隣に発光強度の大きいスポットが無いスポットの発光強度であり、右辺第 2 項はこの平均をとることを意味している。よって、x 軸に発光強度が I_{gr}^l であるスポットに近接するスポットの発光強度 $I_{g'}^l$（遺伝子 g によって遺伝子自体は異なるので添字を g' とした）を、y 軸に上の式によって求めた $\hat{E}_g^l$ をプロットし、線形回帰によって $f(I^l)$ を求めることができる。

3-1-2 分散を安定化するための正規化手法

測定バイアスを取り除く正規化によって測定バイアスを取り除いた後のデータに対して、下記の分散を均一にするための正規化を行い、それぞれ比較する。**1** は Huber が提案した手法 [22]、**2** はこれまで経験的に使われてきた手法 [7] を少し改良したもの、**3** は本研究で Huber の方法をより単純に改良した方法である。

1．Huber による分散安定化変換 [22]

異なるアレイを用いても同じ試料の同じ遺伝子の発現量は等しいはずであるから、線形変換によって異なるアレイでの測定値を等しくすることを目標とする。式で書くと次のようになる。

$$\hat{I}_{gk} = o_k + s_k I_{gk} \tag{5}$$

o、s はそれぞれ線形変換のためのパラメータである。さらに不等分散性を解消するために、発光強度の背後に潜在変数 η と誤差 ν を想定する[32]。

$$I = \alpha + \beta\, e^{\eta} + \nu \tag{6}$$

η と ν は独立に、平均は 0、分散はそれぞれ σ_η^2、σ_ν^2 の正規分布に分布すると仮定する。すると、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。

$$h_k(\hat{I}_{gk}) = \frac{1}{\sigma_\eta}\operatorname{arsinh}\left(-\alpha\frac{\sigma_\eta}{\sigma_\nu} + \frac{\sigma_\eta}{\sigma_\nu}\hat{I}_{gk}\right) \tag{7}$$

さらに、式 $\hat{I}_{gk} = o_k + s_k I_{gk}$ を代入して変数を整理すると次のようになる。

$$N_{gk} = h(I_{gk}) = \operatorname{arsinh}(a_k + b_k I_{gk}) \tag{8}$$

ここで $a_k = -\alpha\frac{\sigma_\eta}{\sigma_\nu} + o_k\frac{\sigma_\eta}{\sigma_\nu}$、$b_k = s_k\frac{\sigma_\eta}{\sigma_\nu}$ である。この $(a_1, b_1, ..., a_K, b_K)$ は以下のプロファイル尤度を L-BFGS-B 法によって最大化することで推定する。

$$-\frac{G'K}{2}\log\left(\sum_{g\in G'}\sum_k (h_k(I_{gk}) - \hat{\mu}_k)^2\right) + \sum_{g\in G'}\sum_k \log h_k'(I_{gk}) \tag{9}$$

ここで、$\hat{\mu}_k$ 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 の微分である。1 回目にパラメータを推定する際は G' として全遺伝子を使用するが、2 回目以降は $\hat{\mu}_k$ からの残差が小さい遺伝子を用いる。この際に、全遺伝子を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 %（$50 < q \le 100$）まで

の遺伝子を使用して再計算を行う。この操作をパラメータが収束するまで繰り返し行う[33]。この様に、一部の遺伝子を使うことで実験の効果を受けていると思われる遺伝子の影響を受けることなく、正規化が行えるよう配慮している。

この解析には統計解析ソフトウェアである R[34]、及び Bioconductor project 内で入手可能な遺伝子データ解析パッケージである vsn を使用した[35]。

２．対数変換

発光強度をプロットすると、発光強度の小さい範囲では分散が小さく、発光強度が大きい範囲になるほど分散が大きくなるといった性質がある。そのため、これまで一般的には発光強度に対して対数変換を施すことで、分散の安定化を図ってきた[7]。式で表すと次のようになる。

$$N_{gk} = \log(I_{gk}) \tag{10}$$

この変換は、遺伝子の発光強度に対して変動係数が一定であるというモデルを仮定しているものとみなせる[32]。このモデルに、遺伝子 g とアレイ k の効果を組み込むと次のように表せる。

$$I_{gk} = \mu_g e^{\eta_k} \tag{11}$$

ここで μ_g は遺伝子 g の本来の発光強度、η_k はアレイ k における誤差であり平均 0、分散 $\sigma^2_{\eta_k}$ の正規分布に従うと仮定する。正規化によってアレイ間の測定バイアスを取り除けている場合には添字 k は不必要である。しかし、正規化を行ってもバイアスが依然として存在していることもあり得るので、分散に対応するアレイの効果を取り除くために、本研究では次の式を用いて対数変換を行うことを提案する。

$$N_{gk} = \log\left(\frac{I_{gk}}{I_{\cdot k}} \cdot I_{\cdot\cdot}\right) \qquad where \quad I_{\cdot k} = \frac{1}{G}\sum_g I_{gk}, I_{\cdot\cdot} = \frac{1}{GK}\sum_g\sum_k I_{gk} \tag{12}$$

このように遺伝子平均値を用いて発光強度を補正することで、アレイ間の測定値を等しくでき、3-1-1 の１で述べた総正規化と同じような効果を期待することが出来る。

３．逆双曲線正弦関数による変換

同じ遺伝子について発光強度に対数変換を施したもの同士をプロットすると、発光強度そのもののプロット図とは反対に、発光強度の大きい範囲では分散が小さくなり、発光強度の小さい範囲ほど分散が大きくなるといった性質がある。そのため測定される発光強度

に対して以下のモデルを仮定する。

$$I_{gk} = \mu_g e^{\eta_k} + \nu_k \tag{13}$$

ここで、μ_g は(6)式と同じく遺伝子 g の本来の発光強度、η_k と ν_k は誤差をそれぞれ乗法的、加法的に説明する項である。η_k と ν_k は独立に、平均は 0、分散はそれぞれ $\sigma^2_{\eta_k}$ 、$\sigma^2_{\eta_k}$ の正規分布に分布すると仮定する。すると I_{gk} の平均と分散は次のように表すことができる。

$$E(I_{gk}) = \mu_g \text{、} \quad V(I_{gk}) = \sigma^2_{\eta_k}\mu_g^2 + \sigma^2_{\nu_k} \tag{14}$$

よって、この分散を平均で表すと次の式が得られる。

$$V(I_{gk}) = \sigma^2_{\eta_k} E(I_{gk})^2 + \sigma^2_{\nu_k} \tag{15}$$

このように、分散が平均の関数になっている場合に、分散が平均に依存しなくなるような変換関数（分散安定化変換）はデルタ法によって以下のように導かれる[36]。

$$N_{gk} = h(I_{gk}) = \int^{I_{gk}} 1/\sqrt{V(E(I_{gk}))}dE(I_{gk}) \tag{16}$$

上の分散と平均の関係式から、次の式が導ける。

$$h(I_{gk}) = \frac{1}{\sigma_{\eta_k}} \mathrm{arsinh}(\frac{\sigma_{\eta_k}}{\sigma_{\nu_k}} I_{gk}) \tag{17}$$

アレイごとに分散安定化のための正規化を行うとすると、$\frac{1}{\sigma_{\eta_k}}$ 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 $\frac{\sigma_{\eta_k}}{\sigma_{\nu_k}} = c_k$ とし、正規化のための変換を次のように定義し直す。

$$h(I_{gk}) = \mathrm{arsinh}(c_k I_{gk}) \tag{18}$$

この c_k をアレイごとに推定しなければならない。本研究では c_k をアレイごとの平均値 $I_{\cdot k}$ の逆数とすることを提案する。以上の正規化手法を式であらわすと次のようになる。

$$N_{gk} = \mathrm{arsinh}\left(\frac{I_{gk}}{I_{\cdot k}}\right) = \log\left(\frac{I_{gk}}{I_{\cdot k}} + \sqrt{\left(\frac{I_{gk}}{I_{\cdot k}}\right)^2 + 1}\right) \tag{19}$$

逆双曲線正弦関数は引数の値が 0 から 1 の範囲では直線的な形状となり、その外の範囲では対数に似た形状をもつ。よって、この変換を行うことで、発光強度がアレイの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、発光強度がアレイの平均値より大きい遺伝子は対数変換を施した際と同等の挙動をとる。これによって、変数変換を行う前と対数変換を行った後のそれぞれの欠点を補うことができる。

3-2 正規化手法の比較方法と比較のプロセス

正規化手法を比較するために、本研究では以下の評価基準を用いる。**1**はこれまで経験的に用いられてきたもの、**2**は Park らが提案したもの[4]を改良したもの、**3**はこれまで経験的に用いられていたもの[37]を改良したもの、**4**は本研究で提案するものである。

１．散布図行列

各正規化手法を最も簡単に比較するために、各正規化を施した際の発光強度をアレイごとに（例えば第 1 アレイを x 軸に、第 2 アレイを y 軸に）プロットしてグラフを作成する。サンプルに異なる操作を加えることで異なった発現を示している遺伝子を決定することが目的である実験においても、ほとんどの遺伝子の発現量は変わらないと思われる[38]。そのため、同一遺伝子は異なったアレイにおいてもほとんど同じ値をとるはずである。よってこのプロットが直線 $y = x$ 上に乗り、かつばらつきが小さく均一であるほど良い正規化手法であるといえる。

散布図行列に準ずる方法として、アレイ間の測定バイアスを明確にするための M-A プロットがある。しかし、前述のように M-A プロットは散布図を 45° 回転させただけの図なので、本質的には散布図を用いて評価していることと同じである。さらに、アレイの枚数が増えるとアレイ組み合わせの数も急激に増え、M や A を計算する手間も大きくなる。散布図行列にも、アレイの枚数が増えるとアレイ組み合わせの数も増え、示すべき図も多くなってしまうという問題がある。しかし、散布図行列の場合は、特に計算をする必要は無く、また視覚的直感によって測定バイアスの存在を充分確認することができる。そのため、本研究では散布図行列を評価基準とした。散布図行列を見て測定バイアスが存在しそうだと思われた場合に、M-A プロットを用いてさらに確認するとよいと思われる。

２．誤差分散

測定バイアスを取り除くための正規化手法を誤差分散によって比較する。これは Park らが提案した基準であり、繰り返し測定間の誤差分散が小さくなっている正規化手法ほど良いと言える[4]。しかし、Park らはアレイ間の誤差分散しか考えていない。アレイ間の誤差分散は、アレイ間の測定バイアスを取り除く正規化を行うことにより減少するが、アレイ

内の測定バイアスを取り除く正規化を行っても減少しない。そのため、アレイ内の誤差分散も用いて正規化手法を評価することを本研究で提案する。アレイ内の誤差分散を計算するためには、遺伝子の同一アレイ内での繰り返し測定が必要である。アレイ内の繰り返し測定が r 回 $(r=1,...,R)$ である遺伝子の誤差分散は以下のようになる。

$$\begin{cases} \text{遺伝子ごとのアレイ間誤差分散}\cdots \hat{\sigma}_g^2 = \dfrac{1}{K-1}\sum_k (I_{gk\cdot} - I_{g\cdot\cdot})^2 \quad where \quad I_{g\cdot\cdot} = \dfrac{1}{K}\sum_k I_{gk\cdot} \\ \text{遺伝子ごとのアレイ内誤差分散}\cdots \hat{\sigma}_{gk}^2 = \dfrac{1}{R-1}\sum_r (I_{gkr} - I_{gk\cdot})^2 \quad where \quad I_{gk\cdot} = \dfrac{1}{R}\sum_r I_{gkr} \end{cases} \tag{20}$$

そして、アレイ間の誤差分散、アレイ内の誤差分散ともに、遺伝子ごとの誤差分散の中央値を用いて正規化手法の比較を行う。式で表すと次のようになる。

$$\begin{cases} \text{アレイ間誤差分散}\cdots \hat{\sigma}^2 = \underset{g}{median}(\hat{\sigma}_g^2) \\ \text{アレイ内誤差分散}\cdots \hat{\sigma}_k^2 = \underset{g}{median}(\hat{\sigma}_{gk}^2) \end{cases} \tag{21}$$

３．分散の均一性

測定バイアスを取り除いた後、分散を均一にするための正規化手法を Mean-SD プロットによって比較する。Mean-SD プロットは x 軸に発光強度の平均値を、y 軸に発光強度の標準偏差を遺伝子ごとにプロットしたものである。正規化によって分散が均一になっている場合は標準偏差も均一になっているため、このプロットは平坦なものになる[37]。そして、異なる変数変換を行った結果間の比較を可能とするために、各軸を発光強度平均値の最大値を分母とした割合で表すことを提案する。式で表すと次のようになる。

$$x\text{軸}：\%Mean = \frac{I_{g\cdot}}{\max_g(I_{g\cdot})} \times 100 \qquad y\text{軸}：adjustedSD = \frac{\sqrt{\dfrac{1}{K-1}\sum_k (I_{gk} - I_{g\cdot})^2}}{\max_g(I_{g\cdot})} \times 100 \tag{22}$$

しかし、値が集中している場合や値にあまり変化が見られない場合は、この図によって確認することは難しい。よって、このプロットを発光強度の平均値に関して 10 分割し、その中で%Mean と adjusted SD の平均をとる。そしてその 10 個の点をプロットした図を使い、正規化手法間の比較を行うことを提案する。

４．第１主成分の固有値割合

測定バイアスを取り除き、分散を安定化した結果を総合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とを提案する。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 $G \times K$ 行列をデータ行列として積

和行列を計算する。大部分の遺伝子は異なったアレイにおいても同じ値をとるはずなので、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。

以上の解析は一部を除き SAS9.1[39]を用いて行った。

4. 対象とするアレイ実験系

本研究では実データとして国立がんセンターにおいて行われた、クロンテック社の新フィルターアレイ(Version4_6)と旧フィルターアレイ(Version3)の比較をするための実験データを用いた。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元素が使われている。実験の概要は、肺がん細胞株 PC-14 を PCR 法で増幅した後に分割し、旧フィルターアレイで 3 回、新フィルターアレイで 3 回、それぞれ一日中に繰り返し測定した。1 枚のアレイには 1176 個のスポットがあり、これには positive control が 15 スポット、negative control としての blank が 19 スポット、同一遺伝子の繰り返し測定が含まれている。この繰り返し測定がなされている遺伝子は、2 回繰り返し測定されているものが 31 個（62 スポット）、3 回繰り返し測定されているものが 62 個（186 スポット）である。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。

5. 結果

5-1 測定バイアスを取り除くための正規化手法の適用結果

本研究で使用した新フィルターアレイの発光強度をプロットしたものを図 1 に示す。

【図 1】

今後はこのデータを用いて、順次以下の正規化の評価を行った。

１．バックグラウンドノイズを取り除く正規化

まず、各アレイ内でバックグラウンドノイズを取り除く正規化を行った。バックグラウンドの値は、各アレイにおける blank スポットの平均値を使用した。blank スポットは各アレイに 19 スポットあるが、近接するスポットによるバイアスを受けていると思われるス

ポットもあるので、それらを除いた 14 スポットを使用した。それらのスポットの平均値はそれぞれアレイ 1 で 60.24、アレイ 2 で 66.11、アレイ 3 で 57.70 であった。

２．総正規化

次にあるアレイと、そのアレイを除いたアレイの平均値との M-A プロットを確認した。それぞれのアレイについての M-A プロットを図 2 に示す。

【図 2】

この M-A プロットによって、アレイ 1 の発光強度が他のアレイに比べて若干小さいことがわかる。よって、アレイ間に測定バイアスがあると思われ、また、この測定バイアスの非線形性は弱いと思われた。そのため、この測定バイアスを総正規化によって取り除いた。

３．近接するスポットによるバイアスを取り除く正規化

本研究で使用した新フィルターアレイ 1 をスキャンした画像を図 3 に示す。

【図 3】

この図では、黒い点が各遺伝子の発光強度を示しており、色の濃いものほど発光強度が強く測定されている。この図を見ると、発光強度の大きいスポットが、近接するスポットに影響を及ぼしていることがわかる。さらに、次の図 4 によって近接するスポットによって生じるバイアスを確認した。

【図 4】

図 4 は、同一アレイ内で繰り返し測定されている遺伝子について、観測された発光強度をプロットしたものである。✻は、隣に発光強度が 400 以上と測定されたスポットが存在するスポットの発光強度を表しており、●は隣に発光強度が 400 以上と測定されたスポットが存在しないスポットの発光強度を表している。この図を見ると、全体的に✻の発光強度の方が●の発光強度より大きいことが確認され、発光強度の大きいスポットは近接するスポットに影響を及ぼしていることがわかる。

さらに、近接するスポットに及ぼす効果を推定するため、次の図 5、6 によって発光強度と近接するスポットに及ぼす効果の関係を確認した。

【図 5】

【図 6】

図 5 は発光強度と近接するスポットに及ぼす効果の関係を、(4)式の $\hat{E}_g$ を使ってアレイ 1～3 のデータをまとめて表したものである。この図において、近接するスポットに及ぼす効果の非常に大きいものが 3 点あり、この 3 点は外れ値と思われる。図 6 は、この 3 つの外れ値を取り除き、発光強度から近接するスポットに及ぼす効果を推定するために回帰直線を引いたものである。

そして、図 6 から得られた回帰直線によって近接するスポットに及ぼす効果を推定し、隣に発光強度が 400 以上と測定されたスポットが存在するスポットの発光強度の調整を次の補正式に基づいて行った。

$$N_g = I_g^l - 0.02262 I^l \qquad where \quad I^l \geq 400 \tag{23}$$

また、推定する際に外れ値とみなし取り除いた 3 点が隣に存在するスポットについては、観測された発光強度から 3 つのスポットが及ぼす効果の平均値 580 を取り除いた。

5-2 測定バイアスを取り除くための正規化手法の比較結果

測定バイアスを取り除く正規化手法を単独あるいは組み合わせて用いたときのアレイ間誤差分散、アレイ内誤差分散によって比較した。まずアレイ間誤差分散を表 1 に示す。

【表 1】

バックグラウンド正規化ではどの遺伝子でも値が減少したが、総正規化や近接するスポットによるバイアスを取り除く正規化では、遺伝子によって必ずしもそうはならなかった。特に総正規化では、M-A プロットによってアレイ 2 とアレイ 3 がアレイ 1 に比べて全体的に発光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。そのためアレイ間の測定バイアスを取り除くことができなかったと考えられる。また正規化手法を組み合わせると、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。

次に、アレイ内誤差分散の結果を表 2 に示す。

【表 2】

近接するスポットによるバイアスを取り除く正規化でアレイ内誤差分散が著しく減少し、バックグラウンド正規化や総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。

また正規化手法を組み合わせると、アレイ間誤差分散の際と同じように、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、アレイ間やアレイ内の測定バイアスを取り除く正規化手法として、それぞれバックグラウンドノイズを取り除く正規化と近接するスポットによる測定バイアスを取り除く正規化を用いることに決定した。この 2 つの正規化手法を組み合わせ、測定バイアスを取り除いた後の散布図行列を以下に示す。

【図 7】

以下の変数変換では近接するスポットによるバイアスを取り除く正規化、バックグラウンドノイズを取り除く正規化の順に正規化を行った後のデータを使用した。

5-3 分散を安定化するための正規化手法の適用結果

１．対数変換

測定バイアスを取り除いたデータに対数変換を施した結果を図 8 に示す。

【図 8】

発光強度がバックグラウンドとほぼ同じ遺伝子は、バックグラウンド正規化によって値が 1 付近に補正されている。そのためそれらの遺伝子は、対数変換を施すことによって 0 付近に集中する結果となった。

２．逆双曲線正弦関数による変換

測定バイアスを取り除いたデータに逆双曲線正弦関数による変換を施した結果を図 9 に示す。

【図 9】

対数変換の際にみられた 0 付近の集中が解消され、全体的に均一にばらつく結果となった。

以上の対数変換、逆双曲線関数による変換の際に用いたアレイ 1～アレイ 3 の平均値は 83、108、108 であった。アレイ 1 の平均値が他のアレイのものと異なることから、総正規化で取り除くことのできなかったアレイ間の測定バイアスを、これらの変数変換の際に取り除くことができた可能性がある。

３．Huber による分散安定化変換

Huber による分散安定化変換を、測定バイアスを取り除く前のデータと、正規化によって測定バイアスを取り除いた後のデータに対して行った。

まず測定バイアスを取り除く前のデータに対して Huber による分散安定化変換を行った結果を図 10 に示す。

【図 10】

測定バイアスを取り除く前のデータに対して Huber による分散安定化変換を行うとばらつきがほぼ均一になり、分布も良い結果となった。

次に、異なる手順で測定バイアスを取り除いた後のデータに対して Huber による分散安定化変換を行った結果を図 11、図 12 に示す。

【図 11】

【図 12】

データ 1 は近接するスポットによるバイアスを除去した後にバックグラウンドノイズを取り除いたデータであり、データ 2 はバックグラウンドノイズを取り除いた後に近接するスポットによるバイアスを除去した。データ 1 とデータ 2 の Huber による分散安定化変換を行う前の段階での大きな違いは、前者ではデータに負値が無いのに対して後者では負値が存在している点である。Huber による分散安定化変換を行った結果、前者ではばらつきがほぼ均一になり分布も良いが、後者ではばらつきもまばらなものがあるうえに分布も歪んだ結果となった。

5-4 分散を安定化するための正規化手法の比較結果

１．Mean-SD プロットによる比較

変数変換前のデータ、対数変換、逆双曲線正弦関数による変換、測定バイアスを取り除く前のデータに対する Huber による分散安定化変換、測定バイアスを取り除いた後のデータ 1､2 に対する Huber による分散安定化変換に対するそれぞれの Mean-SD プロットを図 13 に示す。

【図 13】

また、発光強度と標準偏差との関係をよりわかりやすくするために、それぞれの Mean-SD

プロットを分割したものを図 14 に示す。

【図 14】

これらの図から、変数変換前のデータでは発光強度が大きくなるに従って標準偏差が大きくなっていることがわかる。また、対数変換では逆に発光強度が小さい範囲で標準偏差が大きく、発光強度が大きい範囲では標準偏差が小さくなっていることがわかる。さらに逆双曲線正弦関数による変換や、測定バイアスを取り除く前のデータに対して Huber による分散安定化変換を行ったものは標準偏差が安定していることがわかる。しかし、測定バイアスを取り除いた後のデータに対して Huber による分散安定化変換を行ったものは、発光強度が小さい範囲で標準偏差が大きくなる傾向になった。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。%Mean が 70～100 の範囲は、もとの尺度に戻すと約 2～約 9 の範囲である。つまり、Huber による分散安定化変換は変換前に発光強度が 0 であった遺伝子もなにかしらの値をもつような変換になっていることが分かる。

２．第１主成分の固有値割合による比較

それぞれの正規化手法を行った際の第 1 主成分の固有値割合を表 3 に示す。

【表 3】

測定バイアスを取り除く前のデータに対して Huber による分散安定化変換を行ったものが 0.9986 となり最も大きく、測定バイアスを取り除いた後のデータ 1 に対して Huber による分散安定化変換を行ったものが 0.9973 となり次いで大きい結果となった。また、対数変換と逆双曲線正弦関数による変換では、後者の方が 0.9900 と大きくなった。

以上、１と２の結果を統合して、本データについては逆双曲線正弦による変換が適切と判断した。

6. 考察

本研究では、散布図行列、誤差分散、Mean-SD プロット、第 1 主成分の固有値割合を総合的に検討することで、得られたデータに最も適していると思われる正規化手法を最終的に選択した。測定バイアスが取り除けているかどうかを確認するための評価基準は誤差分散を用い、分散が均一になっているかどうかを確認するための評価基準は Mean-SD プロットや第 1 主成分の固有値割合を用いるという単純な評価方法もありえる。しかし、Huber

による分散安定化変換はMean-SDプロットや第1主成分の固有値割合だけで評価すると最も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思えなかった。したがって評価基準の値だけではなく、散布図行列を確認することは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、発光強度の最も大きいスポットであっても近接するスポットへの影響はさ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、近接するスポットへの影響を推定することができ、それによって近接するスポットによるバイアスを取り除くことが可能であった。しかし、発光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。このような場合は近接するスポットへの影響を推定することはできず、本研究で用いた正規化は行うことができない。そのため、あらかじめ Querec らの方法[40]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光強度の関係を実験によってあらかじめ測定しておき、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スポットごとにバックグラウンド値が測定されていなかったため、近接するスポットのバイアスを取り除く正規化手法を考案した。しかし、スポットごとにバックグラウンド値が測定されている場合は、Schuchhardt らの方法[8]で近接するスポットの効果を取り除くことができるかもしれない。

また、本研究で使用した実験では遺伝子の同一アレイ内繰り返し測定がなされていた。そのため、アレイ内繰り返し測定を用いることでアレイ内誤差分散を計算することができ、また近接するスポットによるバイアスを取り除く正規化を行うこともできた。アレイ内繰り返し測定が存在しないとアレイ内誤差分散が計算できず、アレイ内の測定バイアスを評価することができない。よって測定バイアスを取り除く正規化手法の比較が不十分なものとなってしまう。さらに、アレイ内繰り返し測定を利用することで、必要な場合は測定バイアスを推定することも出来る[9]。したがって、遺伝子の同一アレイ内繰り返し測定は重要であり、今後の実験においても必要である。

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比較を行う場合でも、本研究で用いたプロセスと評価基準を応用することは容易である。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プロセスと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。

最後に、根本的な問題として遺伝子発現量が何を示しているかという問題について様々な議論がなされおり、DNA アレイの意義について疑問視する見方もあることを指摘しておく[6]。しかし最近では、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言われている[41]。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。

7. 結論

放射性同位元素をラベルとする cDNA マクロアレイを対象として、適切な正規化手法を選択するためのプロセスと評価基準を提案した。バックグラウンド正規化と近接するスポットによるバイアスを取り除く正規化を組み合わせ、逆双曲線正弦関数による変換を行うことによって、分散をほぼ均一にすることができた。

8. 謝辞

本研究を行うにあたり御指導頂きました東京大学大学院医学系研究科健康科学・看護学専攻、生物統計学・疫学予防保健学教室の大橋靖雄教授、松山裕助教授、伊藤陽一助手、山口拓洋助手に深く感謝申し上げます。また、貴重なデータを提供して下さった国立がんセンター薬効試験部西尾和人先生に深く感謝申し上げると同時に、研究過程において親切に御指導下さいました生物統計学・疫学予防保健学の大学院生の皆様に心から感謝申し上げます。

9. 参考文献

[1] Schena M, et al. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*. 1995; **270**: 467-70.

[2] Folder SPA, et al. Light-directed, spatially addressable parallel chemical synthesis. *Science.* 1991; **251**: 767-73.

[3] Mocellin S, et al. DNA array-based gene profiling from surgical specimen to the molecular portrait of cancer. *Annals of Surgery*. 2005; **241(1)**: 16-26.

[4] Park T, et al. Evaluation of normalization methods for microarray data. *BMC Bioinfomatics*. 2003; **4**: 33.

[5] Murphy D, et al. Gene expression studies using microarrays: principles, problems, and prospects. *Advances in Physiology Education*. 2002; **26**: 256-70.

[6] Stoughton RG, et al. Applications of DNA microarrays in biology. *Annual Review of Biochemistry*. 2005; **74**: 53-82.

[7] Simon RM, et al. *Design and analysis of DNA microarray investigations*. New York:Springer. 2003.

[8] Schuchhardt J, et al. Normalization strategies for cDNA microarrays. *Nucleic Acids Research*. 2000; **28**: E47.

[9] Fan J, et al. 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.* 2004; **101(5)**: 1135-40.

[10] Yang YH, et al. Normalization for cDNA microarray data. *SPIE BiOS*. 2001. Available at: http://www.stat.berkeley.edu/users/terry/zarray/Html/papersindex.html. Accessed January 18, 2006.

[11] Workman C, et al. A new non-linear normalization method for reducing variability in DNA microarray experiments. *Genome Biology*. 2002; **3(9)**: R48.

[12] Bolstad BM, et al. A comparicon of normalization methods for high density oligonucleotide array based on variance and bias. *Bioinformatics*. 2003; **19(2)**: 185-93.

[13] Colantuoni C, et al. Local mean normalization of microarray element signal intensities across an array surface: quality control and correction of spatially systematic artifacts. *Biotechniques*. 2002; **32**: 1316-20.

[14] Uchida S, et al. 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 Reserch*. 2005; **12**: 1-7.

[15] Berger JA, et al. Optimized LOWESS normalization parameter selection for DNA microarray data. *BMC Bioinformatics*. 2004; **5**: 194.

[16] Edwards D, et al. Non-linear normalization and background correction in one-channel cDNA microarray studies. *Bioinformatics*. 2003; **19(7)**: 825-33.

[17] Yoon D, et al. Two-stage normalization using background intensities in cDNA microarray data. *BMC Bioinformatics*. 2004; **5**: 97.

[18] Bowtell DD. Options available—from start to finish—for obtaining expression data by microarray. *Nature Genetics*. 1999; **21**: 15-9.

[19] Duggan DJ, et al. Expression profiling using cDNA microarrays. *Nature Genetics*. 1999; **21**: 10-4.

[20] Konishi T. Three-parameter lognormal distribution ubiquitously found in cDNA microarray data and its application to parametric data treatment. *BMC Bioinformatics*. 2004; **5**: 5.

[21] Geller SC, et al. Transformation and normalization of oligonucleotide microarray data. *Bioinformatics*. 2003; **19(14)**: 1817-23.

[22] Huber W, et al. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*. 2002; **18**: S96-S104.

[23] Inoue M, et al. Improved parameter estimation for variance-stabilizing transformation of gene-expression microarray data. *Journal of Bioinformatics and Computational Biology*. 2004; **2(4)**: 669-79.

[24] Cheadle C, et al. Analysis of microarray data using Z sore transformation. *Journal of Molecular Diagnostics.* 2003; **5(2)**: 73-81.

[25] Futschik M, et al. Model selection and efficiency testing for normalization of cDNA microarray data. *Genome Biology*. 2004; **5(8)**: R60

[26] Hoffmann R, et al. Profound effect of normalization on detection of differentially expressed genes in oligonucleotide microarray data analysis. *Genome Biology*. 2002; **3(7)**: R33.

[27] Chen Y, et al. Ratio statistics of gene expression levels and applications to microarray data analysis. *Bioinformatics*. 2002; **18**: 1207-15.

[28] Yang MC, et al. A statistical method for flagging weak spots improves normalization and ratio estimates in microarrays. *Physiological Genomics*. 2001; **7(1)**: 45-53.

[29] Sasik R, et al. Statistical analysis of high-density oligonucleotide arrays: a multiplicative noise model. *Bioinformatics*. 2002; **18(12)**: 1633-40.

[30] Goryachev AB, et al. Unfolding of microarray data. *Journal of Computational Biology*. 2001; **8**: 443-61.

[31] Kooperberg C, et al. Improved background correction for spotted DNA microarrays. *Journal of Computational Biology*. 2002; **9**: 55-66.

[32] Rocke DM, et al. A tow-component model for measurement error in analytical chemistry. *Technomertics*. 1995; **37(2)**: 176-84.

[33] Huber W, et al. Parameter estimation or the calibration and variance atabilization of miaroarray data. *Statistical Applications in Genetics and Molocular Biology*. 2003; **2(1)**: Article3.

[34] R Development Core Team. *R: A language and environment for statistical computing*. Vienna:R Foundation for Statistical Computing. 2005.

[35] Gentleman R, et al. *Bioinformatics and computational biology solutions using R and bioconductor*. New York: Springer. 2005.

[36] Tibshirani R. Estimating transformations for regression via addivity and variance stabilization. *Journal of the American Statistical Association*. 2002; **83**: 394-405.

[37]竹内 正弘. 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. 2003.

[38] Kohane IS, et al. 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. 2004.

[39] SAS Institute Inc. *SAS/STAT 9.1 User's Guide.* Cary, NC: SAS Institute Inc.; 2004

[40] Querec TD, et al. A novel approach for increasing sensitivity and correcting saturation artifacts of radioactively labeled cDNA arrays. *Bioimformatics*. 2004; **20(12)**: 1955-61.

[41] Hughes TR, et al. Functional discovery via acompendium of expression profiles. *Cell*. 2000; **102**: 109-26.