2008－BIO－13 (9)
2008／6／26

検定多重性とサンプル個性を利用した臨床ラベル関連遺伝子探索

大羽成征† 石井信†

† 京都大学情報学研究科. 京都府宇治市五ヶ庄
E-mail: †{oba,ishii}@i.kyoto-u.ac.jp

あらまし 異なる臨床ラベルのついた標本群間で平均発現量に差異があるような遺伝子をマイクロアレイデータに基づいて検出する問題において，遺伝子発現パターン間に臨床ラベルに直接対応しない相関がある状況を考慮に入れた新手法を提案する．提案手法では，残差行列 (発現行列を標本群間差異で表現した残り) に特異値分解を施すことによって発現パターン間の相関を表現する因子統計量を定義し，これと通常の統計量とを組み合わせて多次元の検定統計量を用意し，経験ベイズ法に基づく検定を適用する．提案手法は，保守的な FDR 推定値によって第一種過誤率を制御することができるのみでなく，既存の検定スコアを超える検出力を示した．

キーワード FDR，遺伝子選択

Differential gene discovery with considering characteristic patterns

Shigeyuki OBA† and Shin ISHII†

† Graduate School of Informatics, Kyoto University. Gokasho, Uji, Kyoto, Japan
E-mail: †{oba,ishii}@i.kyoto-u.ac.jp

Abstract For better detection of significant genes with differential expression (DE) between different clinical groups based on microarray measurement, we consider, in this study, a situation where expression patterns of significant genes may show correlation which does not directly correspond to the clinical labels. In order to extract the correlation, we defined factor statistics based on a singular value decomposition of a residual matrix, that is, a gene expression profile matrix subtracted by mean differential expressions. We presented a multi-dimensional test statistic consisting of the conventional and the novel statistics, and applied the framework of empirical Bayesian statistical test. This new test had a conservative estimation of false discovery rate (FDR) and exhibited higher power than conventional statistical tests.

Key words FDR, gene selection

1. 導 入

マイクロアレイで得られるような数千から数万の遺伝子の発現量プロファイルデータの解析において，発現差解析は基礎的な課題である．例えば癌原因遺伝子を探索する研究では，癌と非癌のそれぞれ複数の細胞組織片からのデータに基づいて癌と非癌の間で異なる発現量を示す遺伝子（differentially expressed (DE) 遺伝子）を検出する統計的検定が行われ，以降の解析対象を選び出す第一段階スクリーニングとしてしばしば重要な意味を持っている．

DE 遺伝子検出問題は統計的検定を全遺伝子に対して同時に行う問題になっており，一般に同時多重仮説検定問題（Multiple Simultaneous Hypothesis Testing; MSHT）と呼ばれる．MSHT 状況で通常の検定方法をそのまま使用すると偽陽性率制御に問題が生じることが古くから知られていたが，近年では False Discovery Rate (FDR) 基準 [1] を用いる方法が広く受け入れられつつある．一方で，MSHT 状況を積極的に利用した検定統計量の工夫による検出力向上が試みられている．その最初の例は，非対称な両側検定によって検出力の向上を図る Significance Analysis of Microarrays (SAM) 法 [2] であった．Efron の経験ベイズ法 [3] では，検定を階層型モデルのベイズ推定として扱うことによって SAM 法とほぼ同様の結果を得，これにより SAM 法の理論的基盤を与えた．Storey の Optimal Discovery Procedure (ODP) 理論 [4] は，MSHT を構成する帰無仮説と対立仮説がそれぞれ厳密に尤度関数の形で書き下せるときに検出力最大の検定統計量を構成する方法を示し，これに基づく DE 遺伝子検出法 [5] は SAM 法や経験ベイズ法を越える検出力を示した．また経験ベイズ法を，一次元の t 統計量ではなく標準誤差統計量と t 統計量とをペアにした二次元統計量に対して適用する方法 [6] が開発され，ODP と同様の高い

DE 遺伝子検出力が示されている [7].

MSHT における DE 遺伝子検出力の最適化に関して，これまで行われてきた工夫の多くは，単純なグループ間混合モデル，すなわちラベルごとに異なる平均発現量がありそこに独立ノイズが載っているというモデルに依ってきた．しかし，実際の DE 遺伝子検出タスクにおいてこのモデルは不十分であり，改善の余地は大きい．例えば癌関連遺伝子は必ずしも癌と非癌の標本のあいだで平均的に異なる挙動を示しているわけではなく，癌サンプルのうちのごく一部においてのみ高い発現を示していと考えられている．このような遺伝子を癌外れ値遺伝子と呼び，これを検出するために特殊な統計量が工夫されている [8], [9]. また悪性度の強い癌とそうでない癌のグループの二群間で比較を行うときにも，しばしばグループ内に症例のサブクラスタの存在が疑われる [10].

そこで本研究では，一般に DE 遺伝子の発現パターン間に症例ラベルにより直接には決まらないような相関がある場合を考える．行列因子化に基づく因子統計量を定義することによってパターン間の相関を捕らえ，通常の統計量と因子統計量とを組み合わせた高次元統計量のもとで経験ベイズ法を用いるあたらしい検定スコアを定義し，これに基づく検定の性能を人工データと実データによって評価する．

2. 経験ベイズ法に基づく同時多重仮説検定

2.1 遺伝子発現差検定

第 i 遺伝子が第 j 症例において示した遺伝子発現量の観測値を $y_{ij} \in \mathbb{R}$ とし，これを全遺伝子，全症例についてまとめた行列を $\boldsymbol{Y} \in \mathbb{R}^{M \times N}$ とする．M は考慮する全遺伝子の個数，N は考慮する全症例の個数である．また第 i 遺伝子に対応する発現ベクトルを $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{ij}, \ldots, y_{iN})$ と書く．第 j 症例は，症例群 $1, 2$ のどちらかに所属するものとし，ベクトル $\boldsymbol{c} = (c(1), \ldots, c(N)), c(j) \in \{1, 2\}$ でその所属を表すものとする．

遺伝子発現差検定では，各遺伝子 i が症例群の違いに応じて平均的に異なる発現量を示しているか否かを検定するために以下のような確率的データ生成過程を想定する．

$$y_{ij} = \mu_{ic(j)} + \epsilon_{ij} \tag{1}$$

ここで $\mu_{ic(j)}$ は $c(j)$ の値によって μ_{i1}, μ_{i2} のどちらかを示す．ϵ_{ij} は $\sum_j \epsilon_{ij} = 0$ であるようなノイズ項であるが，必ずしも正規分布などの特定の分布形状を仮定しない．以上の前提のもとで以下のような帰無仮説と対立仮説とを比較する．

$$H_0 : \mu_i = 0, \qquad H_1 : \mu_i \neq 0 \tag{2}$$

ここで $\mu_i \stackrel{\text{def}}{=} \mu_{i1} - \mu_{i2}$ は遺伝子の真の群間平均発現差である．

2.2 統計量と棄却域

検定統計量 z_i を発現ベクトルの何らかの関数として $z_i = z(\boldsymbol{y}_i) \in Z$ と書く．適当な棄却域 $R \subset Z$ と，「統計量 z_i が棄却域に含まれる ($z_i \in R$) とき帰無仮説を棄却する」という判定基準とによって検定が定義される．

DE 遺伝子検出では以下の正則化 t 統計量がしばしば用いられる．

$$z(\boldsymbol{y}_i) = t_i = \frac{\bar{y}_{i1} - \bar{y}_{i2}}{s_i + s_0} \tag{3}$$

ここで $\bar{y}_{i1}, \bar{y}_{i2}$ はそれぞれ標本グループ 1，2 に関する発現量標本平均値，s_i はグループ間で共通のグループ内標準偏差の推定値 (pooled standard error) である．

$$s_i = \sqrt{\frac{n_1 + n_2}{(n_1 + n_2 - 2) n_1 n_2} \sum_{g=1,2} \sum_{j=1}^{n} I(c(j) = g)(y_{ij} - \bar{y}_{ig})^2}$$

$I(.)$ は括弧の中の論理式が真であるときに 1，偽であるときに 0 を出力する指標関数，$n_1 = \sum_j I(c(j) = 1)$ と $n_2 = \sum_j I(c(j) = 2)$, は各症例群に含まれる症例数である．s_0 は適当な正の数であり，標本標準偏差 s_i が偶然に小さい値をとったときに統計量 z が大きな値を示すことを防ぐ正則化項である．

一般に統計量 z_i の空間 Z は必ずしも一次元 $Z = \mathbb{R}$ でなくてもよい．2 次元統計量 $z(\boldsymbol{y}_i) = (t_i, \log s_i)$ に基づく経験ベイズ検定法が高い検出性能を示すことが報告されている [6]．また我々は，これが式 (1) のもとでの ODP [5] と漸近的に等価という意味で最適な統計量であることを示した [11]．本研究では，式 (1) の残余項が遺伝子間で一定の相関を持っている場合を考えるが，このとき上記の 2 次元統計量はもはや最適ではない．本研究では，3 節で以下で示すような発現行列の因子化モデルに基づいて残余項の相関構造を多次元統計量の形で抽出し，得られた新しい多次元統計量に基づく経験ベイズ検定法を提案する．

2.3 Efron の経験ベイズ検定

確率変数 h を導入し，帰無仮説もしくは対立仮説の成立を $h = 0$ もしくは $h = 1$ で表現することにする．帰無仮説が成立しているときの統計量 z の分布を帰無分布 $p_0(z) \stackrel{\text{def}}{=} p(z|h = 0)$, 対立仮説が成立しているときの統計量 z の分布を対立分布 $p_1(z) \stackrel{\text{def}}{=} p(z|h = 1)$ と呼ぶ．

検定対象全て（例えば全遺伝子）にわたる統計量 z の分布 $p(z)$（全分布と呼ぶ）が，以下の混合分布で得られるものと仮定する．

$$p(z) = \pi_0 p_0(z) + (1 - \pi_0) p_1(z) \tag{4}$$

ここで π_0 は帰無仮説が成り立つ事前確率である．このとき，統計量が観測された条件下で帰無仮説が成立する事後確率

$$p(h = 0 \mid z) = \frac{\pi_0 p_0(z)}{p(z)} \tag{5}$$

に対して適当なしきい値 λ を設けることで棄却域

$$R = \{z \in Z \mid \widehat{\text{lfdr}}(z) < \lambda\} \tag{6}$$

を設定して行う仮説検定が検出力の意味で最適である．このことは Neyman-Pearson の補題として知られている．この事後確率を，統計量 z を持つ対象に関する帰無仮説を棄却したときの FDR(false discovery rate) という意味で局所 fdr と呼び，

lfdr(z) と書く．

経験ベイズ検定[3]では，$p_0(z), p(z)$ を，サンプルに基づく経験分布で近似することによって局所 fdr を推定する．

$$\hat{\text{lfdr}}(z) \stackrel{\text{def}}{=} \frac{\hat{\pi}_0 \hat{q}_0(z)}{\hat{q}(z)} \tag{7}$$

ここで $\hat{\pi}_0$, $\hat{q}_0(z)$, $\hat{q}(z)$ はそれぞれ帰無仮説成立の推定事前確率，推定帰無分布，推定全分布である．

局所 fdr が統計量 z の関数であるの対して，棄却域 R の関数として定義される FDR は，特に大域的 FDR と呼ばれる．

$$\text{FDR}(R) = \frac{P(h=0)P(z \in R | h=0)}{P(z \in R)} \tag{8}$$

FDR に基づく多重検定では，FDR が予め定めた値（たとえば 0.1）を下回るように棄却域 R を決定する必要がある．そのためには大域的 FDR の推定値が，真値に対して負のバイアスを持たない（正のバイアスを持ってもよい）保証が必要であり，これを保守性と呼ぶ．適切に推定された局所 fdr に基づく式 (6) のような棄却域のもとで実際のサンプル $z_1, \ldots, z_M$ を評価したときに

$$\hat{\text{FDR}}(R) = \frac{\sum_{i=1}^{M} \hat{\text{lfdr}}(z_i) I(z_i \in R)}{\sum_{i=1}^{M} I(z_i \in R)} \tag{9}$$

が FDR の漸近的かつ保守的な推定値になる．

推定全分布 $\hat{q}(z)$ は，観測されたサンプル（全サンプル）$\{z_1, \ldots, z_M\}$ から求める．このとき例えば適当な窓関数に基づくノンパラメトリック密度推定によって $M \to \infty$ で $p(z)$ と一致する漸近不偏推定 $\hat{q}(z)$ を得ることが可能である．

推定帰無分布 $\hat{q}_0(z)$ の求め方には，工夫が必要である．Efron はクラス所属ベクトル $\boldsymbol{c}$ のランダムな並べ替えに基づいて帰無分布からのサンプル（帰無サンプル）$\{z_i^{*b} \in Z \mid i = 1, \ldots, M,\ b = 1, \ldots, B\}$ をシミュレートし，これに基づく分布推定をおこなった．

帰無分布 $p_0(z)$ と全分布 $p(z)$ の比を $r(z) \stackrel{\text{def}}{=} p_0(z)/p(z)$ と書く．$r(z)$ の推定値 $\hat{r}(z)$ は $\hat{r}(z) = \hat{q}_0(z)/\hat{q}(z)$ で得られるが，ロジスティック回帰の要領で $\hat{q}_0(z), \hat{q}(z)$ を経由せずに帰無サンプルと全サンプルから $\hat{r}(z)$ を直接に推定する方法（[3], [6], 詳細は省略する）が提案されており，本研究ではそれを採用した．

推定事前確率 $\hat{\pi}_0$ は，

$$\hat{\pi}_0 = \frac{\sum_{i=1}^{M} I(\hat{h}_i = 0)}{(1/B)\sum_{i=1}^{M}\sum_{b=1}^{B} I(\hat{h}_i^{*b} = 0)} \tag{10}$$

によって求めることができる．ここで $\hat{h}_i, \hat{h}_i^{*b}$ はそれぞれ統計量 z_i, z_i^{*b} に基づく暫定的な検定結果であり，たとえば適当な閾値 $\Delta > 0$ のもとで $\hat{h} = I(|z| > \Delta)$ で決定することで，保守性 $\lim_{M\to\infty} \hat{\pi}_0 \geq \pi_0$ が得られる．

3. 行列因子化統計量に基づく検定

3.1 行列因子化モデル

まず，式 (1) で定義された単純モデルのパラメタ表現を以下のように変更する．

$$y_{ij} = \mu_i b_{0j} + \mu_{0i} + \epsilon_{ij} \tag{11}$$

ここで，条件「$\mu_i = 0$」と条件「$\mu_{1i} = \mu_{2i}$」とが等価となるように，$b_{0j} = -(N/n_1)I(c(j)=1) + (N/n_2)I(c(j)=2), \mu_{0i} = (n_1/N)\mu_{1i} + (n_2/N)\mu_{2i}$ とおいた．

次に残余項 ϵ_{ij} に関して以下の因子化モデルを適用する．

$$\epsilon_{ij} = \sum_{k=1}^{K} u_{ik} v_{jk} + e_{ij}^{*} \tag{12}$$

u_{ik}, v_{ik} は因子化表現のパラメタであり，インデックス $k = 1, \ldots, K$ は因子を表す．因子の個数 K は M 未満かつ N 未満の適当な非負整数とする．e_{ij}^{*} をとくに因子化残余と呼び単純残余 ϵ_{ij} と区別する．相関を持つ因子を因子化表現 u_{ik}, v_{jk} の形で抜き出すことによって，単純残余 ϵ_{ij} が各遺伝子 i に関して独立であることを想定できない場合にも因子化残余 e_{ij}^{*} については各遺伝子 i と標本 j に関して独立と仮定できるようになる．

因子化残余 e_{ij}^{*} は j によらず，平均 0，分散 s_i^{*2} を持つ何らかの分布にしたがうものとするが，必ずしも正規分布を仮定しない．

3.2 因子統計量の算出

前節のモデルを構成する未知量は $\epsilon_{ij}, e_{ij}^{*}$ の二種類の残余が最小となるように求める．

(a) まず μ_i, μ_{0i} を $d_i = \sum_{j=1}^{N} \epsilon_{ij}^2$ が最小になるように求め，得た結果を $\hat{\mu}_i, \hat{\mu}_{0i}$ とする．

(b) 次に残余 $\hat{\epsilon}_{ij} = y_{ij} - \hat{\mu}_i b_{0j} + \hat{\mu}_{0i}$ に関する以下の行列因子化表現を特異値分解を用いて求める．

$$\hat{\epsilon}_{ij} = \sum_{k=1}^{K} \hat{u}_{ik} \hat{v}_{jk} + e_{ij}^{*} \tag{13}$$

ここで $d_0 = \sum_{i=1}^{M}\sum_{j=1}^{N} e_{ij}^{*2}$ を最小化し，なおかつ直交性（$k \neq k'$ のとき $\sum_{i=1}^{M} \hat{u}_{ik}\hat{u}_{ik'} = 0, \sum_{j=1}^{M} \hat{v}_{jk}\hat{u}_{jk'} = 0$）と，$v$ のスケール一定性（$\sum_{j=1}^{N} \hat{v}_{jk}^2 = N$）が満たされるようにする．

以上の計算に基づいて得られた推定値について各遺伝子に対応する部分を抜き出すと，これを遺伝子有意性検定の対象となる検定統計量として用いることができる．具体的には $\mu_i, u_{ik}, i = 1, \ldots, K$ の最小二乗推定量 $\hat{\mu}_i, \hat{u}_{ik}$ およびその残差スケール $\sigma_i = \sqrt{\sum_j \hat{\epsilon}_{ij}^2}$, $\sigma_i^{*} = \sqrt{\sum_j e_{ij}^{*2}}$ が i 番目遺伝子の性質を表す統計量となっている．そこでこれらをさらに整理し，$K+2$ 次元統計量 $z_i = (t_i, \log\sigma_i, u1_i, u2_i, \ldots, uK_i)$ を検定統計量とする．ここで第一次元は，t 統計量 $t_i \stackrel{\text{def}}{=} \hat{\mu}_i/\sigma_i$ であり仮説を直接定義しているパラメタの推定値であるため，これを主要統計量と呼ぶことにする．第二次元目以降を補助統計量と呼ぶが，第三次元目以降は $u1_i \stackrel{\text{def}}{=} \hat{u}_{i1}/\sigma_i^{*}, \ldots uK_i \stackrel{\text{def}}{=} \hat{u}_{iK}/\sigma_i^{*}$ で定義され，これらをとくに因子統計量と呼ぶ．バイアス項 $\hat{\mu}_{0i}$ は検定において重要でないので無視した．

3.3 因子統計量を用いた経験ベイズ法

上記の $K+2$ 次元統計量を用いて局所 fdr を計算するさい，Z の空間の次元が高いことで精度の高いノンパラメトリック分布推定が難しくなる「次元の呪い問題」と，因子統計量の帰無

分布のノンパラメトリックな定義が難しいという「帰無分布問題」の二つの困難がある．

「次元の呪い問題」に対しては，因子統計量間の条件付き独立性を利用し，低次元（2次元）分布推定結果を集めることによって対処した．「帰無分布問題」については，主要統計量 t に関する並べ替えシミュレーションと，因子統計量 uk に関する保守的な拾い出しとの組み合わせによるサンプリング法を適用することによって対処した．

具体的な手順について以下で示す．その手順の意味と得られた局所 fdr 推定値の正当性については次節で補足議論をする．

(a) まずラベル情報 $c(j)$ と遺伝子発現行列 y_{ij} の観測値と，行列因子化モデルに基づいて各遺伝子に対応する $K+2$ 次元統計量 $z_i = (t_i, \log\sigma_i, u1, u2, \ldots, uK)$ を算出する．これは統計量 z に関する真の全分布 $p(z)$ からの M 点の独立ランダムサンプリングと解釈される．次にラベル情報 $c(j)$ のランダムな並べ替えに基づいて，二次元統計量 $(t, \log\sigma)$ に関するサンプル集合 $\{(t_i^{*b}, \log\sigma_i^{*b}) \mid i = 1, \ldots, M, b = 1, \ldots, B\}$ を用意する．B は並べ替えを行った回数である．これを真の帰無分布 $p_0(t, \log\sigma)$ からのランダムサンプリングと同一視する．これと全分布からのサンプルに基づいて，$\hat{\pi}_0$，$\widehat{\mathrm{lfdr}}(t_i, \log\sigma_i)$，$\widehat{\mathrm{lfdr}}(t_i)$ を 2 章の方法によって求める．

(b) 次に，適当なしきい値（例えば $\lambda = 0.8$）のもとで事前検定を行い，その結果を $\hat{h}_i \stackrel{\mathrm{def}}{=} I(\widehat{\mathrm{lfdr}}(t_i) < \lambda)$ とする．これを用いて，二次元統計量 $(t, u1)$ の帰無分布からの標本を以下のように生成する．

$$\{(t_i^{*b}, u1_i) \in \mathbb{R}^2 \mid i \text{ s.t. } \hat{h}_i = 0, b = 1, \ldots, B\}$$

これと観測された全分布からの標本

$$\{(t_i, u1_i) \in \mathbb{R}^2 \mid i = 1, \ldots, M\}$$

とを用いて，二次元統計量 $(t, u1)$ に基づく局所 fdr, $\widehat{\mathrm{lfdr}}(t, u1)$ を求める．同様にして全ての因子統計量に関して主要統計量とのペアによる二次元統計量に関する局所 fdr, $\widehat{\mathrm{lfdr}}(t, uk)$ を求める．

(c) 以上の結果を以下のようにまとめることで，$K+2$ 次元統計量 z_i に関する局所 fdr の推定値とする．

$$\widehat{\mathrm{lfdr}}(z_i) = \widehat{\mathrm{lfdr}}(t_i, \log\sigma_i) \prod_{k=1}^{K} \frac{\widehat{\mathrm{lfdr}}(t_i, uk_i)}{\widehat{\mathrm{lfdr}}(t_i)}$$

3.4 局所 fdr 推定の保守性

$K+2$ 次元統計量に関する局所 fdr を 2 次元ずつペアにして計算してよいのは，因子統計量は直交行列分解によって得られたものであって，$p(t, u1, u2, \ldots, uK) = p(u1 \mid t)p(u2 \mid t) \cdots p(uK \mid t)p(t)$ の条件付き独立性を仮定することができるからである．(注1) 個々の 2 次元局所 fdr は 2 次元統計量空間に適当な大きさのグリッドを切り，帰無標本・全標本に関する 2 次元度数分布を求め，空間的に平滑化をかけるという手順（詳細は省略）によって，少ない計算コストで精度よく求めることができる．

上記の (b) での帰無標本サンプリングを用いた理由は以下のとおり．まず帰無分布において主要統計量と因子統計量が独立 $p_0(t, u1) = p_0(t)p_0(u1)$ であることから，両者を別個にサンプリングした．t のサンプリングは (a) での並べ替えサンプリングの使い回しであるが，$u1$ のサンプリングは帰無分布 $p_0(u1) = p(u1 \mid h = 0)$ からのサンプリングの代わりに主要統計量 t に基づく事前検定結果が $\hat{h} = 0$ であったことを条件とした分布 $q_0(u1) \stackrel{\mathrm{def}}{=} p(u1 \mid \hat{h} = 0)$ からのサンプリングになっている．その結果として定まるサンプリング分布は以下のように，真の帰無分布に対立分布が多少混ざったものとなる．

$$q_0(u1) = \eta_0 p_0(u1) + (1 - \eta_0)p_1(u1)$$

ここで $\eta_0 = P(h = 0 \mid \hat{h} = 0)$ であり，しきい値 λ 次第で η_0 はいくらでも 1.0 に近づけることができるが，分布推定に使用できるサンプル数を減らさないために $\eta_0 < 1.0$ が必要となる．しかし $\eta_0 < 1.0$ であっても，統計量 $u1$ 上の棄却域 R について，上記のような近似サンプリングに基づく推定 FDR に関する以下のような保守性が言える．(注2)

$$\lim_{M \to \infty} \widehat{\mathrm{FDR}}(R) > \mathrm{FDR}(R)$$

4. 実　　験

マイクロアレイデータに基づく発現差解析による有意遺伝子検出のデモンストレーションを，人工データと実データに対して行った．

人工データは 2560 遺伝子× 32 症例の遺伝子発現量行列であり，32 症例のうち 16 症例に癌，残りの 16 症例に非癌のラベルがついている．全 2560 遺伝子のうち 1600 遺伝子は非 DE 遺伝子であり，その発現量はラベルに関係なく独立な正規分布 $N(0, 1)$ にしたがう．残り 960 遺伝子のうち 640 遺伝子を DE-A と呼び，非癌症例のうち共通の 8 例で平均発現量が 1.0 上昇しているものとし，320 遺伝子を DE-B と呼び，癌症例のうち共通の 4 例のみで平均発現量が 2.0 上昇しているものとした．DE-A は正常細胞の組織特異的発現を示す遺伝子であり，非癌判定の指標となるものを想定しており，DE-B はいわゆる癌外れ値遺伝子を想定している．どちらも，癌・非癌症例で平均した場合に発現差 0.5 を持つという意味で同程度に DE らしい遺伝子であるが，症例ラベルで表現されない症例特異的発現パターンに特徴があるため，このパターンに着目することに

(注1)：$\log\sigma$ を含める場合に関して，$p(t, \log\sigma, u1, u2, \ldots, uK) = p(t, \log\sigma)p(u1 \mid t)p(u2 \mid t) \cdots p(uK \mid t)p(t)$ を仮定する．これの妥当性については検討の余地が残っている．

(注2)：略証：サンプリング分布と真の帰無分布との間の差は $q_0(u1) - p_0(u1) = (1 - \eta_0)(p_1(u1) - p_0(u1))$ となるが，検定の棄却域 R において $p_1(u1) > p_0(u1)$ であるためこれは正値．したがって $z \in R$ において $\lim_{M \to \infty} \hat{q}_0(u1) > p_0(u1)$．全ての $k = 1, \ldots, K$ について同様．また $\lim_{M \to \infty} \hat{\pi}_0 > \pi_0$，$\lim_{M \to \infty} \hat{q}_0(t) = p_0(t)$．したがって $z \in R$ において

$$\lim_{M \to \infty} \widehat{\mathrm{lfdr}}(z) = \lim_{M \to \infty} \hat{\pi}_0 \frac{\hat{q}_0(t)}{\hat{q}(t)} \prod_{k=1}^{K} \frac{\hat{q}_0(t, uk)\hat{q}(t)}{\hat{q}(t, uk)\hat{q}_0(t)} > \mathrm{lfdr}(z)$$

(証終)

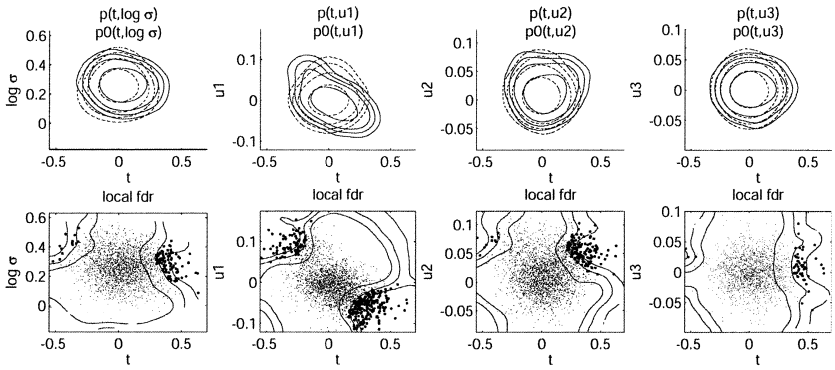

図 1　人工データに対して 4 種類の 2 次元統計量 a1,a2,a3,a4 に関する二次元局所 fdr を計算するデモンストレーション．上段の 4 パネルは帰無分布 (破線)，全分布 (実線) の推定結果を等値線で示している．下段の 4 パネルでは，局所 fdr の 0.1,0.2,0.5 等値線を示し，統計量の分布を黒点であらわし，とくに局所 fdr が 0.2 未満の遺伝子を濃い点で示している．

よって検出性能向上の可能性がある．

実データとしては乳癌の二変種間比較を行ったマイクロアレイデータ（BRCA）[12] を用いた．BRCA は 3226 遺伝子の発現量を 15 症例について測定したものであり，15 症例のうち 7 症例に BRCA1 型乳癌，8 症例に BRCA2 型乳癌のラベルがついている．

比較対照は各種統計量に基づく局所 fdr スコアである．行列因子化モデルに基づいて，各遺伝子に対して 5 次元の統計量 $z = (t, \log\sigma, u1, u2, u3)$ を定義し，これに関する局所 fdr を求める．3 章で説明したように帰無分布の与え方に工夫をしながら 2 次元統計量 a1:$(t, \log\sigma)$, a2:$(t, u1)$, a3:$(t, u2)$, a4:$(t, u3)$ に関する局所 fdr を求め，さらにそれらに基づく多次元統計量 c1=a1, c2:$(t, \log\sigma, u1)$, c3:$(t, \log\sigma, u1, u2)$, c4:$(t, \log\sigma, u1, u2, u3)$ に関する多次元局所 fdr を求めた．a1 は標準的モデルに基づいた場合の最適統計量 ODP [5] と同等の性能を示すことが経験的にも [7] 理論的にも [11] 知られている．

まず図 1 に，人工データに対して 2 次元局所 fdr スコアを求める過程を示した．上段に示された推定帰無分布の等値線 (破線) が左右対称であるのに対して，全分布の等値線 (実線) はゆがんだ形をしている．とくに統計量 a2:$(t, u1)$ の分布は大きく傾いており，対立分布における t と $u1$ の相関が大きいことを示している．一方で統計量 a4:$(t, u3)$ の分布はほぼ縦軸横軸の間で相関がない．最終的に得られた局所 fdr の等値線はこれらの特徴をよく反映したものとなっており，相関が強い a2 はそれを利用して多くの遺伝子が fdr < 0.2 の有意領域に入り，相関が弱い a4 では有意遺伝子は少なめに見積もられている．また a4 で等値線がほぼ垂直に引かれていることから，因子統計量 $u3$ はほとんど情報を持っておらず a4 からは統計量 t 単独で算出された局所 fdr と同等の情報しか得られていないことが分かる．

人工データに対する局所 fdr の計算結果に対して適当なしきい値を設けることで，遺伝子を有意／非有意と判定できるが，これとあらかじめ用意した正解と比較して評価した結果を図 2 に示す．しきい値を任意に動かすことで偽陰性・偽陽性のトレードオフバランスを変化させることができるので，ひとつの局所 fdr に対応してひとつの曲線を描くことができる．図 2(A) で 2 次元局所 fdr4 種の ROC 曲線を比較した．ROC 曲線とは，負例（非 DE 遺伝子）中の正解率（横軸）に対して縦軸に正例（DE 遺伝子）中の正解率（縦軸）をプロットしたものであり，曲線が右上に寄っているほど有意性スコアの性能が高いことを意味する．a2 単独の性能が最高であり，以下 a3, a1, a4 の順に性能が高いことが分かった．このことは図 1 の等値線から定性的に分かったのと同様の結果である．これらを累積的に組み合わせて得られた多次元局所 fdr の性能を図 2(B) で比較すると，情報を加えるにしたがって性能が単調に向上していることが分かる．同じ結果を別の視点から見たのが図 2(C) であり，有意性の上位から遺伝子を採用したときに，因子統計量を援用することで真の DE 遺伝子の純度が高まってゆくことがわかる．図 2(D) は FDR 推定値の保守性を示す．実データへの適用時に DE 遺伝子判定に関する正解は得られないため適当な個数の有意遺伝子を検出したときの FDR を的確に推定する必要がある．これに加えて，サイエンスにおける統計的検定の論理では保守的な結論付けのために実際の偽陽性率（false discovery proportion; FDP）が推定値よりも小さい必要がある．結果は，この要請を満たしている．

次に，実データに対して因子統計量に基づく局所 fdr を計算した．図 3 では上記と同様に 4 種類の二次元統計量を計算した結果を示している．とくに a2, a3 において相関が見られ，因子統計量 u1, u2 による情報供給が期待される．右端の図は任意個数の上位遺伝子を有意と判定したときの FDR 推定値を比較したものである．実データでは遺伝子選択における正解が得られないため，推定値での比較しかできないが，因子統計量を加えることで検出力が向上している．

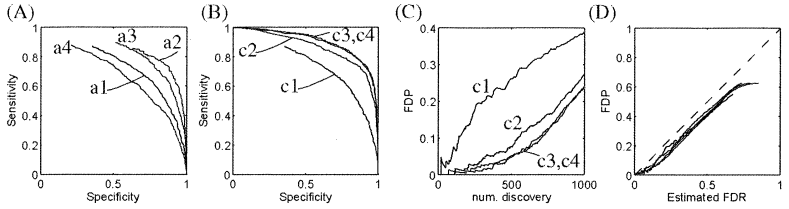

図 2　人工データにもとづく DE 遺伝子検出性能の比較．(A) 各二次元局所 fdr を有意性スコアとして用いた場合の ROC 曲線．(B) 二次元局所 fdr を累積的に用いて得られた多次元局所 fdr の 4 種についての ROC 曲線．(C) 有意と判定された遺伝子数（横軸）に対して，その中に含まれる偽陽性の個数比（縦軸）をプロットしたもの．(B) と同様に，累積局所 fdr を 4 種比較している．(D)FDR の推定値（横軸）と，正解を使用して求めた実現値 FDP（縦軸）を比較したもの．(B) と同様に，累積局所 fdr を 4 種重ねて示しているが，どれも推定値が実現値（対角線）を下回っており，推定値の保守性を示している．

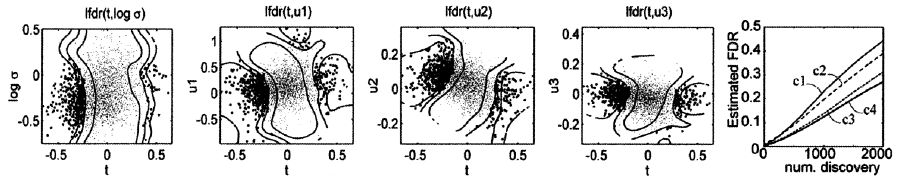

図 3　BRCA データに対する結果．左から 4 つのパネルは二次元統計量の局所 fdr 等高線である．図の見方は図 1 の下段と同様である．右端のパネルは BRCA データにおける累積局所 fdr の比較である．実データでは正解が得られないため推定値のみを示した．図の見方は図 2(C) と同様である．

5. ま と め

遺伝子発現パターンの相関を考慮することで検出力の高い DE 遺伝子検出法を提案した．提案手法は因子統計量に基づく経験ベイズ法に基づいており，検定間の相関関係を因子統計量の形で抽出して積極的に利用することで検出力を向上させつつも，検定の保守性を損なわないことが保証されている．本研究において使用する因子の個数について事前に適当に定めており，保守的用途では小さめの設定で十分であるが，これの自動決定は今後の課題である．また遺伝子発現量データ以外にも，染色体多型や SNP データ解析など検定間の従属関係を積極利用するべき問題は多く，これらへの応用も今後の課題である．

謝　辞

本研究は文部科学省科研費若手研究 (B)19710172 の補助を受けて実施された．

文　献

[1] Y. Benjamini and Y. Hochberg: "Controling the false discovery rate: a practical and powerful approach to multiple testing", J. R. Statist. Soc. B, **57**, pp. 289–300 (1995).

[2] V. G. Tusher, R. Tibshirani and G. Chu: "Significance analysis of microarrays applied to the ionizing radiation response", Proc. Natl. Acad. Sci. USA, **98**, 9, pp. 5116–5121 (2001).

[3] B. Efron and R. Tibshirani: "Empirical Bayes methods and false discovery rates for microarrays", Genetic Epidemiology, **23**, 1, pp. 70–86 (2002).

[4] J. Storey: "The optimal discovery procedure: a new approach to simultaneous significance testing", Journal of the Royal Statistical Society: Series B (Statistical Methodology), **69**, 3, pp. 347–368 (2007).

[5] J. Storey, J. Dai and J. Leek: "The optimal discovery procedure for large-scale significance testing, with applications to comparative microarray experiments", Biostatistics, **8**, 2, p. 414 (2007).

[6] A. Ploner, S. Calza, A. Gusnanto and Y. Pawitan: "Multidimensional local false discovery rate for microarray studies", Bioinformatics, **22**, 5, pp. 556–565 (2006).

[7] E. Perelman, A. Ploner, S. Calza and Y. Pawitan: "Detecting differential expression in microarray data: comparison of optimal procedures", BMC Bioinformatics, **8**, pp. 28+ (2007).

[8] B. Wu: "Cancer outlier differential gene expression detection", Biostatistics, **8**, 3, pp. 566–575 (2006).

[9] R. Tibshirani and T. Hastie: "Outlier sums for differential gene expression analysis", Biostatistics, **8**, 1, pp. 2–8 (2007).

[10] M. Ohira, S. Oba, Y. Nakamura, E. Isogai, S. Kaneko, T. Hirata, H. Kubo, T. Goto, S. Yamada, Y. Yoshida, S. Ishii and A. Nakagawara: "Expression profiling using a tumor-specific cdna microarray predicts the prognosis of intermediate-risk neuroblastomas", Cancer Cell, **7**, 4, pp. 337–350 (2005).

[11] S. Oba and S. Ishii: "Optimal Sufficient Statistics for Multiple Simultaneous Hypothesis Testing", Under Review (2008).

[12] I. Hedenfalk, D. Duggan, Y. Chen, M. Radmacher, M. Bittner, R. Simon, P. Meltzer, B. Gusterson, M. Esteller, M. Raffeld, et al.: "Gene-expression profiles in hereditary breast cancer", New England Journal of Medicine, **344**, 8, pp. 539–548 (2001).