様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成 25 年　5 月　8 日現在

機関番号：　６２６０３
研究種目：基盤研究（C）
研究期間：　2010～2012
課題番号：　22500269
研究課題名（和文）　ゲノムデータに対して有効性が高い多重検定法の開発

研究課題名（英文）　Efficient multiple testing methods for genome data

研究代表者
藤澤　洋徳（FUJISAWA HIRONORI）
統計数理研究所・数理・推論研究系・准教授
研究者番号：　00301177

研究成果の概要（和文）：　Ｐ値を推定する際に，標本数の少なさをカバーするために，他の遺伝子のデータを加えることがしばしば行われている．しかしながら，その妥当性は，必ずしも検証されていなかった．本申請研究では，その妥当性が保証されるＰ値推定の条件を数理的に整理できた．その条件のもとでの最適な検定も導出できた．数値実験でも実データ解析でも，劇的に優れていることが検証できた．平均の同等性だけでなく分散の同等性についても検討した．

研究成果の概要（英文）：　The P-value has often been estimated using the observations on other genes to overcome the small sample size. We have investigated necessary conditions to ensure the validity. Under such conditions, the uniformly most powerful test has been obtained. This test showed a good performance by simulations and data analyses. We have considered the equality of variance as well as the equality of mean.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2010 年度	1,300,000	390,000	1,690,000
2011 年度	1,000,000	300,000	1,300,000
2012 年度	1,000,000	300,000	1,300,000
年度			
年度			
総　計	3,300,000	990,000	4,290,000

研究分野：統計科学
科研費の分科・細目：統計科学
キーワード：遺伝子発現差解析．有意性検定．Ｐ値．

１．研究開始当初の背景

統計科学の方法は，通常は，繰り返し実験の回数(n)が説明変数の数(p)よりも多い場合（n>p）を想定していた．ところが，ゲノムデータは，繰り返し実験の回数(n)よりも説明変数の数(p)が圧倒的に多い場合，いわゆる「n<<p」問題を提示したのである．統計科学の分野に，その逆の状況に対応する方法を，強く要請した．

そのようなゲノムデータの一つである遺伝子発現データを研究当初のターゲットと

したい．遺伝子発現データでは，繰り返し実験は数回なのに，説明変数にあたる遺伝子数が数百から数千という状況である．遺伝子発現データにおいては，遺伝子発現差が存在する遺伝子を発見するという遺伝子発現差解析が主要な目的の一つとして行われ続けている．

その目的を達成するために，各遺伝子に対して，適当な検定統計量に基づいて有意性検定を行うことが，しばしば行われている．これは，一般的な観点から言うと，統計科学における平均差の同等性検定に対応する．そして，有意であると判断された遺伝子を，遺伝子発現差が存在する遺伝子であると考える．

このときに，有意かどうかを判定するために，データに基づいてＰ値を推定する必要がある．ここに問題が起きる．繰り返し実験の回数が少ないために，通常のＰ値の推定方法では，その信頼性に疑問が残るのである．

各遺伝子に対するデータだけを利用する限り，この問題を克服することはできない．繰り返し実験の回数が少ないため，何らかの事前情報がない限り，本質的に精度を上げることはできない．しかし，遺伝子発現データには，ある種の特殊性がある．それは，遠い遺伝子同士のデータは，ほぼ独立と考えられるという点である．そこで，ある遺伝子のＰ値を推定するときに，他の遺伝子のデータを「うまく」援用することができれば，繰り返し実験の回数に対応する情報が増えて，推定の精度を上げることができるかもしれない．「n<<p」問題は，遺伝子数(p)が多すぎることで，通常の統計的方法の適用を困難にした．しかしながら，遺伝子発現データにおける遺伝子発現差解析においては，逆に，遺伝子数が多いことを逆用する試みが行われている．

代表的な方法が Tusher ら(2001) によって提案された．彼らは古典的な並べ替え技法を援用することを考えた．通常，並べ替えの技法を利用した検定は，着目している仮説（ターゲットとなる遺伝子）に対応するデータに並べ替え技法を適用してリサンプルを行い，そのリサンプルに基づいてＰ値が計算される．しかし，遺伝子発現データでは，繰り返し実験の回数が少ないために，それだけでは，十分な数のリサンプルは得られない．彼らは，着目していない仮説（他の遺伝子）に対して得られたリサンプルも使うことで，飛躍的にリサンプルの数を増やした．つまり，遺伝子数が多いことを，逆用したのである．このアイデアは現在でも注目を浴び続けている（たとえば Scheid and Span (2007) や Southworth et al. (2009) など）．

ただし，並べ替え技法の単純な適用によるリサンプルの増加には，大きな問題点がある．並べ替え技法は，仮説が一つの時には，その妥当性は保証されており，ある意味での最適性も保証されている．しかし，単純な適用では，違う遺伝子同士のリサンプルの意味が，必ずしも同じではないという問題点が内在しているのである．違うリサンプルを一緒にした場合には，Ｐ値の推定には，必ずしも妥当性が保証されない．

この問題点を克服する方法が，Pan (2003) によって提案された．この研究では，他の遺伝子のデータを援用しても妥当性が失われないように，検定統計量の形に工夫が施さており，さらに，並べ替えの方法にも工夫が施されている．そのような工夫によって，妥当なＰ値の推定方法が提案された．ただし，それらの提案の仕方は，非常にアドホックであり，適当な意味での最適性が保証されているという訳ではない．

２．研究の目的

本研究では，他の遺伝子のデータを援用しても妥当性が保証されるＰ値の推定方法を，アドホックに与えるのではなく，適当な意味で最適な方法として与えることを試みたい．これまでのアドホックな議論に統一的な観点を与えて，その本質を明らかにして，ある意味での最高到達点を理論的に与えたい．そのような研究は，これまでに一つもないように思われる．その結果として，アドホックな感覚だけでは発見しにくいような，より安定的でよりパワフルな方法をも提案できればと考えている．妥当性が保証される条件を，扱いやすい形で，かつ，現実的な要請に合うように書き下すことがポイントであると考えている．

ここまでは，普通の有意性検定，より明確には，平均の同等性検定に着目していた．ところが，最近になって，申請者個人に，ある生物学者から，平均差ははっきりしないが，分散差が確実にありそうな例を提示された．遺伝子発現データでは，分散の同等性検定は，これまで，全く議論されていないように思われる．特に，平均と分散では，単純な最強力検定においても，理論にギャップがある．平均の同等性検定で培われた方法のアイデアを利用して，このギャップを埋め，そのようなデータに対する新しいＰ値の計算法をも提示したい．

ところで，基本的には，遺伝子発現データに着目して話をしてきたが，上述の内容は，他のタイプのゲノムデータにも適用可能性は高いと考えている．たとえば，SNP やプロテオームのデータなどでも，平均の同等性検定は行われている．そのようなデ

ータも，同様に「n<<p」問題を内在しており，そのようなタイプのデータへの拡張可能性も考えて行きたい．

３．研究の方法

平成 22 年度

Pan (2003) によって提案されたＰ値の推定方法では，ある検定統計量がアドホックに提案され，その検定統計量に関して，ある種の制限された並べ替え技法が適用されている．さらに，妥当性を保証するために，データの分布に対称性が仮定されていた．そのような想定の下では，Ｐ値を推定するとき，他の遺伝子のデータを援用しても，その推定方法は妥当性が保証される．

申請者は，最近に，次の結果を得た．まずは，ある種の制限された並べ替え技法を適用できる統計量の中で，データの分布が対称であるときに，他の遺伝子のデータを援用してもＰ値の推定の妥当性が保たれる検定統計量のクラスを考えた．もちろん，そのクラスには，Pan によって提案されたアドホックな検定統計量が含まれる．そのクラスの中で，検出力が最も高い有意性検定は何になるかを考えた. その結果は, Pan (2003) の検定統計量を少しだけ改良したものとなった.

実は, Pan (2003) の検定統計量の形を見たときから，得られたタイプの検定統計量の方が，ある意味で検出力が高いであろうと予想していたのだが，その通りの結果が得られた．また，同時に，Pan (2003) がアドホックに提案していた標本サイズの割り当て方が，最も検出力を高くするということも証明できた.

しかしながら，データの分布は対称とは限らないし，ある種の制限された並べ替え技法は，過去の通常の並べ替え技法の概念から少し離れている．特に，データの分布の対称性は，必ずしも保証されないので，そのようなときにでも，Ｐ値の推定に妥当性が保証される方法が欲しい.

データの分布に対称性を仮定することによって, 数理的な議論は容易になっていた. しかしながら，データの分布に対称性を仮定せずに，愚直に，他の遺伝子のデータを援用してＰ値を推定しても妥当性が保証されるという条件を，数理的にうまく書き下すことはできるであろう？ その中で最適な有意性検定は，どのようなものになるのであろうか？ 並べ替えの技法を, ある種の制限された制約の下ではなく，より普通の使い方に近づけることはできないだろうか？ まずは, これらのテーマを追い求めてみたい.

データの分布の対称性を外すことで，自由度が大きくなるので，妥当性を保証する検定統計量のクラスが広くなると予想される．より広いクラスでの議論を行えば，より検出力が高い方法が得られないかと期待している.

また，数理的に最適性が保証されたとしても，その方法が，数値的に，どの程度優れていて，かつ，想定外の状況においても安定的であるかどうかは定かではない．現実的な遺伝子発現データに対して，想像される状況において，様々なシミュレーションなどによって，そのようなことを確認したい.

平成 23 年度以降

平均の同等性検定の理論的研究が終わり, その有用性が確認された後は，まずは，現実の遺伝子発現データへ適用してみたい．さらに，遺伝学研究所の研究者たちとも議論を進めたいと考えている．申請者は，イネの研究者とマウスの研究者という二種類の研究者と密接に議論できる状態にあるので, 何かのタイプのデータに偏ることなく, 汎用的なコメントが手に入れられると期待している. 何かしらの問題点が出てくれば, その問題点を克服できるような方法論を構築したい.

平均の同等性検定の話が終わった後は，分散の同等性検定に関しても，平均の同等性検定で培われたアイデアが，どこまで拡張可能かを確認したい．これについては，研究計画にも書いたように，全く過去の研究がないと思われる.

並べ替えの技法は，基本的には，平均の同等性検定において，その最適性が示されており，その検定に使うために提案されている．しかし，申請者は，並べ替えの技法を，単なるリサンプルと考えれば，Ｐ値の推定という意味では，数値計算的には問題は起こらないと考えている．つまり，とりあえず最適性を忘れれば，単なる推定の意味では妥当性は保たれるであろうと想像している．まずは，この点を確認したい.

研究のコアとなるのは，対象が平均から分散に変わるとき，モーメントに関して線形性が失われるので，それを，どう克服するかである．つまり，平均の同等性の検定と同様に，他の遺伝子のデータを援用してもＰ値の推定の妥当性が保証されるという条件を，数理的にどう書き下すかである．平均の同等性ほどクリアではないが，分散の同等性に関しても，ある種のクラスでは最強力検定が導かれている．そのような性質を使って，このテーマにも積極的にチャレンジしたいと考えている．特に，ある生物学者からの積極的な依頼があるので，データ解析的にも，非常に先駆的な研究にな

る可能性が高い.
平均の同等性検定が行われているゲノムデータは遺伝子発現データに限らない. SNP やプロテオームデータでも類似の検定が行われている. データにはそれぞれの特徴があり, それらの特徴に合わせた最適な検定方法が提案できないだろうか, というテーマについても取り組んでみたい.

４．研究成果

まずは平均の同等性検定を対象にした. データの背後にある分布が対称な場合には, 本研究申請の前に結果を得ていた. ある種のクラスでは, ある検定統計量が最適であると分かっていた. 本申請研究では, データの背後にある分布が対称でなくても, 二つの母集団において, ある種の同一性を持つクラスでは, ある検定統計量が最適であると分かった. 後者の検定統計量は前者の検定統計量よりも自由度が一つ小さかった. 実験の繰り返し回数は小さいので, 自由度が一つ違うというのは, 大きな差異をもたらす. 実際に, 後者の新しい検定統計量に基づくと, 過去のアドホックに提案された手法に比べて, 検出力という意味でも, Ｐ値推定という意味でも, 劇的に優れていることが検証できた. 実際のデータ解析でも導出した手法の優位性は明らかとなった.
その話の延長として分散の同等性検定に関してもチャレンジした. 他の遺伝子のデータをうまく援用してＰ値を妥当に推定できる検定統計量については幾つか思い付いた. その方法の妥当性もある程度は確認できた. ある程度のクラスの中では統一的に扱うことは可能となった. しかし, 平均の同等性検定のように, 非常に汎用的なクラスにまで広げられなかったのは残念であった.

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計１件）
① Fujisawa, H. and Sakaguchi, T. (2012). Optimal significance analysis of microarray data in a class of tests whose null statistic can be constructed. TEST, Vol.21, 280-300, DOI 10.1007/s11749-011-0243-5, 査読有.

〔学会発表〕（計９件）
① 藤澤洋徳：ゲノムデータのための統計的手法の開発, 新領域融合プロジェクト外部レビュー, 2012.10.31, 東京
② 藤澤洋徳：ゲノムデータ解析のプロジェクトに携わって, 統計サマーセミナー , 2012.8.5, 熱海
③ 藤澤洋徳：遺伝子発現差解析におけるＰ値推定と最適な有意性検定, 統計科学セミナー, 2012.1.13, 福岡
④ Fujisawa, H.: Optimal significance analysis of microarray data in a class of tests whose null statistic can be constructed, International Conference on Advances in Probability and Statistics, Hong Kong, China, 2011.12.30
⑤ Fujisawa, H.: Optimal significance analysis of microarray data in a class of tests whose null statistic can be constructed, 7th IASC-ARS Joint 2011 Taipei Symposium, Taipei, Taiwan, 2011.12.17
⑥ 藤澤 洋徳: 遺伝子発現差解析における並べ替え技法によるＰ値推定, 融合「遺伝機能」プロジェクト会議, 2011.10.20, 三島
⑦ 藤澤洋徳: マイクロアレイデータを利用した遺伝子発現差解析において帰無統計量が構成できる検定の中で最適な有意性検定, 名古屋統計セミナー, 2011.5.20, 名古屋
⑧ 藤澤洋徳, 坂口隆之: 遺伝子発現差解析における並べ替え技法による P 値推定と最適な有意性検定, 科研費研究集会「生物情報解析の理論的基礎とその応用」, 2010.12.7, 東京
⑨ 藤澤洋徳, 坂口隆之: 遺伝子発現差解析における並べ替え技法によるＰ値の推定, 統計関連学会連合大会, 2010.9.7, 東京

６．研究組織
(1)研究代表者
藤澤　洋徳（FUJISAWA HIRONORI）
統計数理研究所・数理・推論研究系・准教授
研究者番号： 00301177
(2)研究分担者 なし
(3)連携研究者
二宮　嘉行（NINOMIYA YOSHIYUKI）
九州大学・数理学研究院・准教授
研究者番号： 50343330
坂口　隆之（SAKAGUCHI TAKAYUKI）
大分県立看護科学大学・看護学部・助教
研究者番号： 10436496
高田　豊行（TAKADA TOYOYUKI）
国立遺伝学研究所・系統生物研究センター・助教
研究者番号： 20356257