［ 別 紙 ２ ］

論 文 審 査 の 結 果 の 要 旨

申請者氏名　モハマド マニル ホセイン モラー
Mohammad Manir Hossain Mollah

1. 問題の所在

マイクロアレイは数千ないし数万の遺伝子の発現プロファイルを鳥瞰することを可能とする。もっとも基本的な解析は、条件により発現量が異なる遺伝子を検出することである。調査する遺伝子数が標本サイズを格段に上回るため、発現量が遺伝子間で異なる様子を確率分布で記述し、階層ベイズモデルで表現することにより、検出力を高める方法も開発されてきている。遺伝子間の関連を表現するネットワークを推定する方法も開発されている。しかしながら、これらハイスループットな統計的推定・検定手法は、例外なく遺伝子発現のパターンに対して何らかの仮定をしている。その仮定が成立するときは有効な方法となる。他方、何らかの理由でデータが想定外の発現パターンを含むときは、検出力が低下し、また偽陽性を多く拾う危険性がある。ところが、数多くの遺伝子を分析の対象としているため、個々の遺伝子を詳細に調べるのはほぼ不可能である。データには分析の前提条件を満たさない想定外の発現プロファイルを持つ遺伝子が含まれているのか、含まれているとするとそれはどのくらいの数か、数遺伝子なのか数十遺伝子なのか、あるいは数百、数千のオーダーなのか、これまで調べる術がなかった。

2. β 尤度法と β 荷重

本論文は、グループ間で発現に違いがある遺伝子を検出する方法として、近年頑健な推定法として提案されたβ尤度法を階層ベイズモデルに適用する。遺伝子のβ荷重を計算し、その分布をモデルから予測される分布と対比する方法を提案し、シミュレーションと実データの解析によりこの方法の有効性を証明する。最尤法は、データの背後にある確率構造の統計的モデリングを行い、その前提の下でデータが生成される確率（尤度）を最大にするよう、パラメータを推定する。すなわちデータに最もなじむ確率構造を推定する。本論文で提案するβ尤度法は、回帰分析において対数変換を一般化させたBox-Cox変換に着想を得て、対数尤度を一般化させたものである。微分をとったスコア関数を見ることにより、各遺伝子が異なる重みを持たせた荷重尤度法と同等であることがわかる。遺伝子tの重み（β荷重とよぶ）は尤度f_tをべき乗したf^{β}_tとなる。従って、モデルの想定外の発現プロファイルを持つ遺伝子は尤度が小さく、小さな重みを持つため、解析結果にあまり影響を与えない。すなわち、β尤度法は、異常値に影響されない頑健な推定法であることが期待される。交差検証法により頑健性と検出力をバランスさせるβの値を決める。さらにβ荷重の遺伝子間の分布をモデルが妥当する場合に期待される分布と対比させることにより、異常値を統計的に検出し、モデルの妥当性を診断することが可能となる。

3. シミュレーションによる有効性の検討

標本サイズが 60 の中規模のデータと標本サイズが 20 の小規模なデータについて、2 通りのシミュレーションを行い、提案手法の有効性を検証した。第一のシミュレーションでは異常値による頑健性を提案手法と既存の手法を比較した。偽陽性のコストを勘案した真陽性の検出力を表す AUC および pAUC、グループ間で発現に差のある遺伝子の割合（ここでは混合率という）の推定精度を調査したところ、モデルの前提が成立し、異常値を含む遺伝子がない場合には、どの手法も良好な成績を示した。しかし、異常値を含む遺伝子が存在する場合には、既存のどの手法も AUC および pAUC が減少し、混合率も大幅に過大推定された。これに対し β 尤度法は、中規模のデータ、小規模のデータいずれにおいても、異常値による性能の低下は見られず、混合率も偏りなく推定することが示された。第二のシミュレーションでは、β 荷重の分布によりモデルの妥当性を診断することの可能性を調査した。ガンマ分布に従う発現プロファイルデータに対して、対数正規分布を当てはめる場合について実験を行った。形状パラメータが小さい遺伝子では微小な発現が優先するため対数正規分布で近似することができず、β 荷重が極端に小さくなることが示された。

4. 実データの解析

公開されている 3 つのデータについて分析を行った。第一のデータは頭頸部癌の患者 22 人の癌組織と正常組織を対比したデータで、12625 の遺伝子中 2.2%にあたる 261 の遺伝子が $p<10^{-5}$ で有意に異常な発現プロファイルを持つ遺伝子として検出された。従来法では 95%の確率で差なしと判定するにもかかわらず、β 尤度法は逆に 95%の確率で差ありと判定している遺伝子が 6 つあった。これらは LRP8、S100A8、S100A9、TRIM29、CSTA、ACP5 で、癌との関連が報告されている。いずれも有意に β 荷重が小さく、発現プロファイルをプロットしたところ、異常値が検出された。第二のデータは肺癌患者からの 54675 の RNA 転写産物で、40 人が腺癌、18 人が扁平上皮癌である。β 荷重の分布は予測分布に比し上側と下側に裾が重く、モデルの妥当性が疑われた。そこで平均発現量と分散を調査したところ、有意に β 荷重が小さい遺伝子は分散が大きく、逆に有意に β 荷重が大きい遺伝子は発現に変異がないことが示唆された。後者の遺伝子をはずして再解析したところ、概ねデータがモデルと異常値で説明されることがわかった。第三のデータはシロイヌナズナ Bay0×Sha 組換え近交系 211 株からとられた 22810 プローブの発現および 578 の SFP マーカー遺伝子型の情報である。ここにおいても β 荷重の分布は予測分布と大きくかい離し、異常値を伴う発現プロファイルを数多く検出した。従来法による eQTL 解析は、発現を制御される遺伝子の数を大幅に過小推定する可能性があることが示された。

システム生物学を支える技術の革新に伴い、質的にも量的にもデータ自身の持つ情報量が飛躍的に膨らんできている。トランスクリプトーム、メタボローム、プロテオーム、およびそれらの統合するデータ解析の手法も急速に進歩している。一方で、データの精査がかつてなく困難な問題として突きつけられている。本論文が提案する β 尤度法とそこから派生する β 荷重の分布によるアプローチは、膨大で複雑なデータを二段階接近法により精査する枠組みを提供しており、今後有効性が立証されることが期待され、学問的にも応用的にも貢献するところが大きい。よって審査委員一同は本論文が博士（農学）の学位を受けるに十分な価値があると認めた。