論 文 内 容 の 要 旨

博士論文題目

統計的手法による遺伝子発現情報からの細胞状態の同定に関する研究

氏　　名　　行縄　直人

（論文内容の要旨）

近年，マイクロアレイや定量的 PCR 法などの mRNA 定量化技術により，細胞サンプルにおける包括的な遺伝子発現情報を得ることが可能となり，細胞の状態と遺伝子発現を直接結び付けて解析する，いわゆるトランスクリプトーム解析が行われるようになった．本論文では，トランスクリプトーム解析における諸問題に対して，頑健な解析を行うための統計的手法に関して議論する．

まず，包括的 mRNA 測定技術の一つである，アダプタ付加競合 PCR (ATAC-PCR)法により得られる蛍光量データの特徴とその補正法について報告する．これまで問題となっていたアダプタ長依存の測定バイアスの解明を主眼とし，ATAC-PCR 法で得られたデータの詳細な解析を行った．解析結果に基づき蛍光ピーク値に関する観測モデルの定式化を行ない，ノイズ項のパラメータの推定量の導出と，それらを用いたピーク値補正法を提案した．この手法を，アダプタノイズ解析のために特化した採取されたピークデータに適用し，アダプタ依存ノイズのパラメータを求め，次いで，実データに対しバイアス補正の適用を試み，その有効性を確認した．

次に，生きた細胞における遺伝子発現ダイナミクスの解析を目指し，遺伝子発現プロファイルの時系列に対する解析法について述べる．ここでは，状態空間モデルに基き，ノイズプロセスに白色ガウシアンを仮定した線形ダイナミカルシステムモデルを考え，変分ベイズ法による推定とモデル選択を行うための新たな手法を提案した．本手法を出芽酵母細胞周期に関する公開データセットに適用したところ，従来手法で選択されたモデルと比較し，より単純かつ尤もらしいモデルが選択された．また，この結果得られたモデルパラメータは，生物学的な考察と良く一致した．人工データへの適用も行い，ノイズを含む時系列データに対する有効性が示された．

最後に，遺伝子発現からの癌の病理診断を想定した，新たな多クラス識別法について述べる．本手法では，多クラス識別問題を一対一ペアや一対残りペアなどのラベルの任意の組み合わせから成る 2 値分類問題群に分解し，各問題での判別結果を統合することによって最適な識別結果を得る．各 2 値分類問題における真の分類確率がクラス所属確率をパラメータとした確率モデルによって生成されると考え，これを 2 値分類器によって得られた分類確率の推定値から推定する方法，さらに 2 値分類器の重みを推定する方法を導いた．本手法を人工データおよび甲状腺がん分類問題をはじめとした実データに適用し，従来のヒューリスティクスによる投票法と同等以上の性能を達成することを示した．さらに，この分野で提案されてきたいくつかの多クラス識別法との比較を行い，本手法の優位性および性質を明らかにした．

氏　名	行縄　直人

（論文審査結果の要旨）

細胞の状態を，遺伝子転写産物量のパターンベクトルとして捉えた遺伝子発現プロファイルは，遺伝子機能や細胞の表現型に関与する遺伝子の同定，また，逆にそれを用いた表現型予測など幅広い解析の基礎となる重要な情報源である．遺伝子発現プロファイルは，各種ノイズや，得られるサンプル数に対する特徴空間の次元の高さなど，従来の統計的解析の対象にはない特徴的な性質を有しているため，これに特化した統計的解析手法が求められている．本論文では，遺伝子発現解析における，計測データの補正，時系列発現データからの特徴抽出および腫瘍組織の判別解析の 3 つの異なる課題のそれぞれに対し，確率モデルに基づく新たな解析手法の提案および検証を行い，その有効性について議論している．本論文の主な成果は以下のように要約される．

1. 高効率 mRNA 測定技術の一つである，アダプタ付加競合的 PCR (ATAC-PCR) 法の蛍光ピーク値の補正法を開発した．アダプタ長依存の測定バイアスの補正行うために，蛍光ピーク値に関する観測モデルの定式化を行ない，ノイズ項のパラメータの推定量の導出と，それらを用いたピーク値補正法を提案した．本手法を実際の実験データに適用し，バイアスの軽減および，キャリブレーション精度の向上を検証した．
2. 遺伝子発現時系列データに対する特徴抽出法として，線形ダイナミカルシステムモデルに基づく解析法を提案した．多数の遺伝子発現の変化が少数の因子に駆動されるモデルを基に，さらに，従来の解析法では従来の手法では考慮されていなかった，システムノイズおよび観測ノイズを導入したモデルとその推定法を提案した．出芽酵母の細胞周期における遺伝子発現データなどを用いて，モデル推定を試み，データの適合度，データ生成の内部状態，および観測系列の生成行列に関して検討し，ノイズを含む高次元の時系列データから特徴抽出法としての有効性を示した．また，モデル選択基準を用いて，自動的に生物学的考察に基づく経験的な解析指標と同等のものが得られることを確認した．
3. 遺伝子発現プロファイルによる癌サブクラスの診断を目的として，新たな多クラス識別法を提案した．多クラス識別問題をラベルの任意の組み合わせから成る二値分類問題群に分解し，各問題での判別結果を統合することによって最適な識別結果を得るものであり，クラス所属確率を隠れ変数とした確率モデルの推定を行い，さらに，2 値分類器ごとの重みを任意のロス関数に基き最適化する．これにより，従来手法で問題となっていた，最適な 2 値分類器群の選択問題の解決を試みた．本手法を人工データおよび各種腫瘍分類問題に適用し，その有効性を示した．

本論文の成果は，特殊な性質を持つデータを扱う遺伝子発現解析の種々の段階の問題において，確率モデルの有効性を示しており，バイオインフォマティクスならびに情報科学における寄与が多大である．よって，博士（工学）の学位論文として価値のあるものと認める．