統計数理(2006)
第 54 巻 第 2 号 405–423

特集「予測と発見」
[総合報告]

遺伝子発現データに基づく予測と推定：言いたいことと言えること

大羽 成征†

(受付 2006 年 1 月 19 日；改訂 2006 年 4 月 5 日)

要　　旨

医学・生物学では，遺伝子発現量に関する大量のデータに対してこれまで様々な方面から統計的解析が試みられてきた．こうした研究はチキンレースのようなものであり,「データから言いたいこと」をきちんと主張したいが,「データから言えること」の境界を超えてはいけない．とくに細胞診断のための教師付き解析では，実際に結論にバイアスを入れて言い過ぎの境界を踏み超えてしまう解析研究が多かった．本解説では，ここまでの歴史の流れを追いながら，主に遺伝子発現量に基づく細胞病理診断の問題において遺伝子発現データの解析者がはまりやすい落し穴について紹介し，我々がその落し穴を避けつつ行ってきた解析研究についての解説を行う．近年ではこうした保守的な考えかたが拡がるとともに，一方では遺伝子発現情報をフル活用しつつなおかつ境界は踏み超えない積極的な手法が現れ始めた．こうした巧妙な手法の紹介を通して今後への明るい展望と残された課題についても議論する．

キーワード：遺伝子発現解析，教師付き学習，教師付き特徴選択．

1. 遺伝子発現データから何が分かるか？

遺伝情報が生物個体の世代間で受け継がれる情報を担っているものであるのに対して，遺伝子発現の測定情報は，生物の各細胞がいままさに生きている生活状態をスナップショットのような形で反映したものである．遺伝子発現データは，複雑な生命現象をシステムの細部に渡って理解することを目的とする現代の分子生物学にとって，重要な情報の宝庫であると考えられ統計的解析の対象となっている．

1.1 遺伝子の発現制御とその計測データ

生物が持つ遺伝情報は鎖状の高分子である DNA を媒体としたデジタルデータであり，世代交代の中でコピーされつつ受け継がれてゆく．DNA 上の情報全体を指してゲノムと呼び，タンパク質の設計図となる情報(遺伝子)や，それを読み出すきっかけを決める情報(プロモータ)がゲノム上に記述されている．遺伝子が読み出される分子反応過程はかなり解明が進んでおり，遺伝子の発現と呼ばれている．シグナル分子が細胞核内に伝えられると，対応するプロモータ上で遺伝子領域の転写と呼ばれる分子反応が始まる．これにより遺伝子情報はメッセンジャー RNA という小さな分子にネガ焼きコピーされる．遺伝子情報を載せたメッセンジャー RNA はリボゾームという細胞内小器官のところまで移動し，そこでメッセンジャー RNA に書かれたデジタル情報どおりのアミノ酸配列を持つポリペプチドが作られる．これが生命活動の各種機

† 奈良先端科学技術大学院大学 情報科学研究科：〒630–0192 奈良県生駒市高山町 8916–5

能を担うタンパク質となるのである．タンパク質が作られるまでの情報の流れに関する以上の基本的な枠組みはセントラルドグマと呼ばれ，その理解についてはかなり信頼されているが，最近では転写によって作られた RNA 自体が各種機能を担う場合や，遺伝子領域以外に関しても転写によって RNA を作る場合など，様々な例外的現象も知られつつある．

このようにして生産された種々タンパク質が互いに協力して細胞の生命現象を担うためには，必要なタンパク質が必要なときに必要な量だけ作られる必要があり，そのためには必要な遺伝子が必要なときに必要な量だけ発現する必要がある．遺伝子の発現のオン・オフを制御するべく，トリガーを働かせたり，はたまた止めたり，動かなくさせたりなど，様々な分子反応が明らかになってきているが，これらをまとめて遺伝子発現制御と呼ぶ(セントラルドグマ・遺伝子発現制御について例えば田村, 2000 などが易しくまとめている)．

マイクロアレイ技術によって様々な情報断片を運ぶメッセンジャー RNA の個数をそれぞれ同時に測定することができるようになった．メッセンジャー RNA の個数は，対応する各遺伝子の発現の回数を間接的に反映していると考えられるため，これを遺伝子発現量と同一視することができる．そこで様々な条件のもとでの，メッセンジャー RNA の個数が測定され，統計的解析の対象となる．

1.2 限られたデータから言えることはどこまでか？

遺伝子発現量の測定データから，何をどこまで知ることができるだろうか？

遺伝子発現は細胞の生化学的状態の違い(表現型)に関する情報をかなりの程度漏れなく反映していると考えられている．細胞の生化学的状態が変わるときはたいてい，それに対応するシグナル分子が遺伝子発現のプロセスを変化させているからである．そこで，細胞環境の各要因や細胞内での各遺伝子の発現のオン／オフの間の関係を，因果関係のネットワークとして捉え，そのネットワーク構造や働きを調べようとする試みがある．これまでにベイジアンネットワークモデル(Friedman, 2004)，ペトリネット(Doi et al., 2004)などのモデルが提案され，応用が試みられているがこうしたネットワークモデルに対して細胞の様々な内外環境のもとでの遺伝子発現の網羅的観測のデータを適用すれば，生命現象の分子基盤の主要なものほとんどを知り得るのではないかと期待されてきた．

しかし，サンプルを調製しマイクロアレイを計測する時間・金銭的コストの問題から，一つの研究プロジェクトの中で解析対象とすることのできるマイクロアレイの枚数は数十から数百程度に限られるのが通常であり，遺伝子間の複雑な因果関係／相関関係の全てを同時に十分な精度で推定するのに必要なだけの情報量はなかなか得られそうにない．得られる限りの情報に基づいて言えることを探すならば，知りたいことの焦点を絞らねばならない．ここに，本論の中心テーマ「言いたいことと言えることとの間の境界線」が現れる．

癌などの細胞診断を遺伝子発現プロファイルによって行う研究では，例えば遺伝子発現のパターンによって癌の悪性度を予測する問題が扱われてきた．こうした研究では各症例から取得した細胞標本のそれぞれについて各症例の臨床経過に関する情報が得られるため，これと遺伝子発現量プロファイルデータとを併せて教師付き解析をすることができる．たとえば癌の悪性度を教師ラベルとした教師付き解析では,「悪性度との関連性が統計的に有意である遺伝子を選び出す」ことや「新規症例について細胞診断を行い悪性度を推定する」ことが主な目的となる．こうした研究においても，各遺伝子の発現と癌の悪性度との間でネットワークをなす因果関係の詳細が「知りたいこと」であるのに変わりは無いのだが，あえて詳細までは「言えない」こととして諦めて，いかに「言えることの境界ギリギリ」を主張するかが問題の焦点となった．

例えば,「遺伝子発現量での細胞診断による悪性度予測は，統計的に有意な正解率を持つ」ということは，最低限言いたいことであったし，さらには「細胞診断の正解率は 80% である」等

の主張もしたい．提案する新手法が過去に知られていたものよりも有効なのだ，と主張したい場合も多いだろう．ネットワーク推定と比べた場合に教師付き解析の重要な特徴は，クロスバリデーションや汎化誤差推定など各々の主張について，それがデータから支持されているのか否かを判定し，信頼性を検証するための分かり易い道具立てが存在することである．

ただ，遺伝子発現に関する教師つき解析では,「遺伝子選択」と「ラベル予測」の二つの組合せかたを誤ることで言い過ぎの境界を踏み超えてしまう落し穴が存在する．

本論では，ここまでの歴史の流れを追いながら遺伝子発現データの解析者がはまりやすい落し穴について紹介する(2 章)．また落し穴を避けつつ行ってきた解析研究の一例として我々が神経芽腫の予後予測について行っている研究を解説する(3 章)．また，最近になって出てきた落とし穴にはまっていそうではまっていないぎりぎりの線上を歩く巧妙な手法をいくつか紹介する(4 章)．

近年になり，マイクロアレイのほかにも様々な新しい測定技術と新しい種類のデータが出現している．SNP (一塩基多型)とは DNA の遺伝子領域における一塩基レベルの個体差のことであり，遺伝病の原因を構成していると考えられている．アレイ利用比較ゲノム解析(aCGH)とは，細胞核内の染色体コピー数異常の測定法である．正常な体細胞で 2 コピーであるものと比較して，癌などの染色体疾患の原因もしくは結果として増幅や欠損がみられることが知られており，aCGH ではそれらを検出し疾患と関連付けてゆくことが重要な研究課題とされている．電気泳動法や MALDI-TOF などの質量分析法では，タンパク質やその断片の質量スペクトルを計測することができ，こうして得られるデータはプロテオミクス情報と呼ばれる．様々な反応の過渡をなすタンパク質の正確な状況はメッセンジャー RNA の量を測るだけでは間接的にしか分かったことにならない場合があるため，こうした別種の測定方法も必要であるとされ様々な試みが続けられている．

これら全ては結局は，生命現象と生体分子反応間の因果関係のネットワークを知ることを究極の目的としているという意味で，遺伝子発現解析と同様の目的を持つ別個の方法として位置付けることができ，またそこで解かねばならない問題の構造も非常に類似している．本論では遺伝子発現解析を主な対象として述べるが，ここで述べる事柄の多くは，これらの新しい種類のデータに対しても相変らず重要性が高いことにも注意しておく．

2. 言い過ぎの境界

教師なしクラス発見と教師付きクラス分類の二つは機械学習における主要な対象分野であり，遺伝子発現に基づく細胞診断の研究においても重要な役割を担っている．しかしこれらの不適切な使用が横行し，その結果として「言い過ぎの境界」を踏み超えた結論に至ってしまった研究が数多く出版され，反省されている．Nature 誌では「遺伝子発現解析では統計学が重要であるがあやしい議論が多いため，健全な結論を得るためには統計学者によるレビューが欠かせない」と警鐘を鳴らす記事が掲載された(Tilstone, 2003)．医学系で有名な Lancet 誌では，これまでの研究に関する再現性検討の研究を積極的に掲載している．Ntzani and Ioannidis (2003)はそれまでに行われた癌の予後予測に関する 84 の研究を集めてきて，きちんとクロスバリデーションや独立テストがなされている研究が少ないことを指摘した．また，それらが行われているものについても，クロスバリデーションによって示された性能が独立テスト性能と比較して大幅に良い値を出しているものが多いことを指摘して，過学習による過大評価の可能性を示唆している．Michiels et al. (2005)はそのなかで公開データ 7 種を選び出したうえで手法を揃えて再解析を行い，そのうちの 5 種類についてチャンスレベル未満の予測性能しか見出せなかったと結論付けている．言い過ぎ研究の存在は大変に恐れられていると言っていい．

この章では，言い過ぎの境界に関する重要なトピックをいくつかとりあげたい．

2.1 クラスタリングはクラス発見の根拠になりえるか

90 年代後半は夢が先行する時代であった．遺伝子発現量の行列に対して，行方向，列方向のそれぞれで階層型クラスタリングを行い，その結果を元に上手な並び替えを行い，赤黒緑で色分けした画像で可視化を行う Eisen プロット(Eisen et al., 1998)という表示方法があるが，貴重な情報がぎっしりと詰まった宝箱のように輝かしく見えた．この研究が出てきた頃はネットワーク推定の研究がまだ始まったばかりであり,「発現量解析」という言葉は，ひとつひとつの遺伝子や遺伝子群の動きを Eisen plot の上で肉眼で追いながら想像するということ以外の何物でもなかった．最大の現実的な目標は遺伝子もしくは症例に関するクラス発見であった．

Eisen Plot はデータを網羅的に可視化する手法として優れており今でもさかんに使用されているが，実は階層型クラスタリングの結果から積極的に言えることは少なく，見えた印象を元にデータの裏の構造について「可能性を示唆する」程度のことしかできないとされている．これは階層型クラスタリングはモデルを持たないためである．遺伝子発現量データにおけるクラスタについて何かを裏付けるためには，まずクラスタが有るとはどういうことか無いとはどういうことか，その両方に関するモデルがあらかじめ必要である．両方あってはじめて，データに基づいて片方を否定して片方を採用する論理的な根拠を得ることができる．構造の可能性を示唆する以上の意味を Eisen Plot に持たせる議論がなされれば，残念ながら言い過ぎである．

一方でパラメトリッククラスタリングの枠組みでは，観測におけるノイズの混入過程や，細胞内での遺伝子発現の分布の生成過程，とくにその中に観測データ生成過程においてクラスタ構造がどのような影響を与えるか，についてのモデルをあらかじめ含んでおり，それを推定することで結果としてクラスタリング結果を得る．モデルが妥当であるという前提を認めれば，結果として得られたクラスタリング結果が妥当であるか妥当でないかについてデータに基づいて定量的に評価し，議論することが可能になる．

我々はこれまでに，主成分分析で得られた低次元特徴空間における正規混合分布モデル(Muro et al., 2003)，オリジナル空間における制約付き混合主成分分析モデル(Yoshioka et al., 2002; 吉岡・石井, 2002)，方向ベクトル成分のみに着目した混合 von Mises-Fisher 分布モデル(大羽・石井，2004)など，様々なパラメトリックモデルを用いて，様々なクラスタリング法を提案している．これらは全て遺伝子発現量データにおけるクラスタの妥当な生成モデルを目指した試みである．

しかし，モデルがあることによって妥当性の議論が原理的に可能であることと，それがモデルとして実際に妥当であることとは別の話である．考えなければならない問題は階層的になっており，前提のもとでどのクラスタが妥当であるかの問題と，前提自体が妥当であるかの問題は別である．データに基づいて前提自体の妥当性を強く示した研究は私の知る限り存在せず，まだ結論が出そうな気配は無い．

さらに悪い知らせもあり，様々なパラメトリック手法のもとでクラスタリングによって得られたクラスタ構造の再現性を調べた結果，モデルによらず再現性はかなり低いというネガティブな結論に達した研究がある(Garge et al., 2005)．

パラメトリック法・ノンパラメトリック法ともに，クラスタリングを行ううえでベクトルパターン間の距離を定義する計量の問題は無視できない．階層化クラスタリングではユークリッド距離もしくはベクトル間の角度(相関係数と等価)で測ることが多いが，とくに各細胞の特徴を表す数千次元のベクトルパターン間を比較する際には，その高次元さゆえにどの 2 つのパターン間もほとんど似ていない中で類似性を比較することにならざるを得ない．また，類似性を評価する要素として各次元(各遺伝子に対応する)を採用するか否かによって，ベクトルパ

ターン間の距離の定義自体にいくらでも変動を生じさせ得る．距離定義の変動に伴って，クラスタリング結果も大きく変わってしまう．

以上で見てきたようにクラスタリング結果だけを根拠にしたとき，言い過ぎにならないようにクラス発見を主張することはまだ困難であるといわざるを得ない．

クラスタリング結果だけを根拠にクラス発見を主張できないのであれば，それ以外のところに根拠を求めねばならない．例えば，癌細胞の遺伝子発現パターンを調べる研究では，各試料について悪性度などの症例ラベル情報がついている．そこでクラスタリング結果と，ラベルとを比較することによって，クラスタリング結果の正当性を示せばよい．

ところが，そもそも症例に悪性度高／低のようなラベルが既についており，最終的にラベル予測の良し悪しに基づいて結論を導き出すのであれば，これを教師ラベルとして用いた教師付き解析を行うことができる．教師なしで行うのは回り道である．

以上のように教師無しでクラス発見を強く主張できるのはどういう状況か？に関して，我々はまだ十分な答えを持っていない．一方で教師ラベルを用いた場合には望みがありそうである．そこで，教師ラベルを用いた場合にどこまで言えるのかを，次に考えてみよう．

2.2 教師ラベルからどこまで言えるか？

遺伝子発現データに関する教師付き解析の目的には，有意遺伝子を選択する問題と教師ラベルを予測する問題とがあるが，両者は互いに密接な関係を持っており完全には分けられない．

ラベル予測を行う際に，ラベル予測に関係のない遺伝子の発現量情報はノイズ源となるため，遺伝子選択法を前処理として用いて遺伝子を絞った上で，既存の教師付き解析手法を適用する，という合わせ技が有効である．しかし，この遺伝子選択プロセスの位置づけを誤ることに基づく「言い過ぎ」が横行した．「言い過ぎ」の原因は情報漏洩に基づく選択バイアスとテスト例が少ないことによるバリアンスの二つである．典型例を挙げながら順に見ていこう．

2.2.1 情報漏洩に基づく選択バイアス

まずは以下のような解析が行われたとしよう．これは正しいやり方と言えるだろうか．

例 1．(1) 遺伝子選択を行って有意な遺伝子を少数選び出し，選び出された有意遺伝子のリストを示す．
(2) 選ばれた遺伝子を用いて教師付き分類器を構成する．このときクロスバリデーションを行って分類性能を示す．

これは Nature に掲載された van't Veer et al. (2002) の教師付き解析の要約であり，一見しただけでは分からないかもしれないがクロスバリデーションが不完全になっている典型例である．遺伝子選択にも教師ラベルデータが使用されているためクロスバリデーション結果を示すさいに教師ラベルデータが隠されておらず，そのせいで分類性能を高く示し過ぎてしまうのである．これを情報漏洩と言う．遺伝子選択と分類器構成とをあわせた過程全体で教師ラベルが使われているので，情報漏洩を起こさないためには，この過程の全体を対象としてクロスバリデーションを行う必要がある．

ところでせいぜい 100 や 200 のサンプル数を対象としてこうした解析を行っている場合，クロスバリデーションの各々の過程において，遺伝子選択法によって選ばれてくる少数の有意遺伝子の顔ぶれは毎回異なる．教師付き遺伝子選択によって選び出される遺伝子リストの信頼性は，その程度のものなのであるということにも注意と覚悟が必要である．

では次の例は正しいだろうか．

例 2．(1) 遺伝子選択を行って有意な遺伝子を有意性の高い順に N 個選び出し，選ばれた

遺伝子を用いて教師付き分類器を構成する．このとき，クロスバリデーションを行って分類性能を示す．

(2) $N = 1, 2, 3, \ldots, 1000$ に対して示されたクロスバリデーション性能を比較し，クロスバリデーション性能を最大とさせた N を最終的に採用する．このときのクロスバリデーション性能の最大値を，分類性能として示す．

今度の例では遺伝子選択の過程もクロスバリデーションの対象になっているため，例 1 の問題は生じていない．しかし，「$N = 1, 2, 3, \ldots, 1000$ の候補の中からの選択」を行うところで，選択の判定基準として使われた分類性能の値は，教師ラベルデータを全て使用することで得られている．気付きにくいが，(1)のクロスバリデーションでは，教師ラベルデータを全て使用している．教師ラベルデータを全て使用して多数の候補の中から一つを選択したならば，ここで情報漏洩に基づく選択バイアスが生じるのである．

遺伝子選択のプロセスもモデル選択のプロセスも，教師ラベルを用いる過程は全て教師付き解析の一環であることを意識する必要があって，そうでないと簡単に情報漏洩を起してしまう．2000 年代前半頃のものでこうした例に似た研究論文を目にしたならばこのあたりをあえて強調して意識した記述があるもの以外は，ほとんどが不完全なクロスバリデーションとなっているのでぜひ慎重に読んでほしい．

テスト例が少ないことによるバリアンス

さて，例 1，例 2 のように，仮にクロスバリデーション時点でバイアスを載せてしまったとしても，テスト用のブラインドデータを別個に用意して独立なテストを行えばそこでは選択バイアスは生じない．そのため，独立テストはとても重要なものと考えられている．標準的には，使用可能な全データの 1/2 で学習，1/4 でモデル選択，1/4 でテストという配分による独立テストを行うことが常識的とされてきた(Hastie et al., 2001, p. 196)．

しかし，それで本当に十分と言えるであろうか？

例えば珍しい癌の症例などではテストデータとして新規に用意できるデータの個数をあまり多く用意できない．そのさいには症例数が少ないことによるバリアンスが大きくなってしまう．バリアンスが大きければ，偶然に良い結果・悪い結果が出てくる可能性も大きくなり，出版バイアスすなわち，たまたま良い結果が出たものだけが論文として採択されて出版されているというバイアスが無視できなくなってしまう．

独立テスト用にどれぐらいの個数のサンプルを用意すれば十分であるか，主張したい内容にもよるので一概には言えないが，それだけで信頼性を保証するためには，10 や 20 のテストデータだけでものを言うのは危険が大きいと思われる．テスト用のブラインドデータとして，使用可能な全データの 1/2 程度もしくはそれ以上用いることも最近では増えてきているようだ．ただし，もしもクロスバリデーションが正しく行われておりバイアスが乗っておらず，独立テストでも似た性能が示されていたならば，両者合わせて全体としての信頼性は高まる．その意味でも，仮に最終的な独立テストが得られる場合にも，クロスバリデーション時に無駄なバイアスを乗せるべきではないのである．なお，使用可能なデータ数によっては，次章で述べる Leave Two Out 法のような二重クロスバリデーションも選択肢として検討に値するだろう．

3. 遺伝子発現量の教師つき解析

この章では，癌に関わる遺伝子発現量に対する教師付き問題の一つの例として，我々が神経芽腫に対して中川原 章氏(千葉県がんセンター研究所長)の研究グループと共同で行っている解析研究について紹介する．

神経芽腫は小児の腹部腫瘍であり，小児において白血病に次いで多い癌である．悪性度の高い早期進行例がある一方で，自然寛解してしまう悪性度の低い例もあることが知られている．悪性例では早くから厳しい治療を施すことが必要である一方，治癒が期待できるケースでは，治療を緩和することで厳しい化学療法などによる副作用(腎障害などの後遺症)を低減することができるため，あらかじめそれらを適切に判別する方法が望まれている．これまでに，MYCN 遺伝子の増幅の有無や，腫瘍組織グレード分類を用いた判別法などが有効とされて使われていたが，複数要因に基づく判断が矛盾するなど，判別が困難なケースも稀ではなかった．

我々の研究の目標は遺伝子発現解析によって，この状況を突破することであった．これを目指して，神経芽腫組織から抽出した 5,340 個の遺伝子を載せたマイクロアレイチップを開発し 136 症例の遺伝子発現量を計測し，さらに各症例の診断情報や予後情報とのクロス解析を行った．またその結果を利用して，対象遺伝子数を 200 まで絞ってコストを低減した新チップを開発し，これに基づく診断システムを構築した．

我々の論文(Ohira et al., 2005)では以下の結果についてまとめている．

- マイクロアレイによる 2 年時生存/死亡の判別器を構成し，9 割弱の正解率を得ることができたこと
- その内訳を見てみたところ，既存マーカーによっては判別のできない「中間予後群」と呼ばれるグループに対しても，有意な正解率で予後良/予後悪を分けることができていたこと
- さらに実用的診断システムの構築のために対象遺伝子の個数を 200 遺伝子に絞った低コストチップによる判別性能も，思わしいものであったこと

また，統計手法の観点からは以下の点が重要であったと考えている．

- pair wise 法による遺伝子選択を前提として判別器を構築したこと
- 判別性能を偏りなくなおかつ低い分散で示すための Leave Two Out 法を選択したこと
- 2 値判別の確率値出力を考慮したこと

論文中では，その理由に関する考察などに関して紙面と対象読者の都合で述べられなかった事項が多かった．以下では，多少細かくなるが議論の補足を行っておきたい．

3.1 遺伝子選択における pair wise 法の意義

この研究では，診断コストを低くするチップの実用化を狙っており，低コストチップにおいて対象とする遺伝子数を少なく絞るために遺伝子の重要性を評価してランク付ける必要があった．重要遺伝子をランク付けして選択する問題はマイクロアレイという測定手法が出てきた最初期からある重要な問題であるが，その目的や基準はひとつではない．

一般に良く知られている第一の目的は，各遺伝子の発現量が教師ラベルに対して統計的に有意であるか否かということである．また遺伝子発現量パターンに基づく教師ラベル識別性能を考えるにあたって，各遺伝子発現量の寄与を考えたい場合には，単一遺伝子としての重要性のみならず遺伝子の組合せも評価する必要がある．

我々は以下のように，複数の基準を順番に用いて徐々にスクリーニングするヒューリスティクスを用いて遺伝子の重要度の順位をつけた．

(1) 教師ラベルを用いずに遺伝子発現の全分散を計算し，その大きなものから順に 1000 遺伝子をスクリーニングする．
(2) t-score に基づいて上位 500 遺伝子をスクリーニングする．
(3) 500 遺伝子間で pair wise 法ランキングをつける．

ステップ(1)の目的は，遺伝子発現量の繰り返し測定における再現性誤差の標準偏差(約 0.3)と比べて同程度未満の標準偏差しか持たない遺伝子を，あらかじめ排除しておくことである．ステップ(2)において，t-score とは遺伝子毎に発現量の群内平均の差と群内標準偏差の平均との比をとった統計量，いわゆる S/N 比のことであり，その絶対値が高いものほど遺伝子の統計的有意性が高いと考えられる．ステップ(3)の pair wise 法は，遺伝子のペア毎に有意性スコアをつける．具体的には，二つの遺伝子のペア毎に発現量に基づく二次元特徴空間をつくって教師ラベルの線形判別を行い，そのときの判別性能を各ペアのスコアとする．各遺伝子について自分が参加したペアのスコアの最大値を，各遺伝子のスコアとして，ランキングをつける．この論文で取り扱ったデータでは，ラベル間の頻度に偏りがあったため，判別性能を sensitivity と specificity の調和平均で評価するという独自の工夫を行ったが，本質的には Bo and Jonassen (2002)と同じ方法である．ステップ(1), (2)が必要であったのは，ステップ(3)で必要とされる計算量が，候補遺伝子数の 2 乗に比例するため，その節約が必要になるためである．

さてこうしてつけられたランキングに基づいて，上位遺伝子 M 個を選び，WV 法(weighted vote 法；変量間の相関をゼロと仮定することで単純化した線形判別分析；ナイーブベイズ法)による判別を行ったところ判別性能と遺伝子数 M との関係は図 1 のようになった．なお，判別性能は Leave One Out によって計算しており，各 Leave One Out ごとに遺伝子選択をしなおしていることに注意しておく．図中で実線は pair wise 法による遺伝子ランキングを用いたもの，破線は t-score の絶対値を遺伝子ランキングに用いた場合を重ねて描いている．pair wise 法のオリジナル論文(Bo and Jonassen, 2002)では，同程度の判別性能が pair wise 法によれば比較的少数の遺伝子によって得られるとされていたが，我々が得た結果においても果たして実際にそのようになっていることが分かる．

しかしこの結果を得るのは簡単ではなかった．pair wise スコア計算時の 2 次元データに対する線形判別器には，クラス間で等しい分散を仮定しており，なおかつ共分散がゼロであることを仮定したものを使っている．これは最大限の簡単化を施したものであるが，これよりもちょっとでも複雑なもの(たとえば分散がクラス間で異なるなど)を使ってしまうと学習データに対する判別性能が高くなる一方で，Leave One Out データに対する判別性能が低くなってしまった．

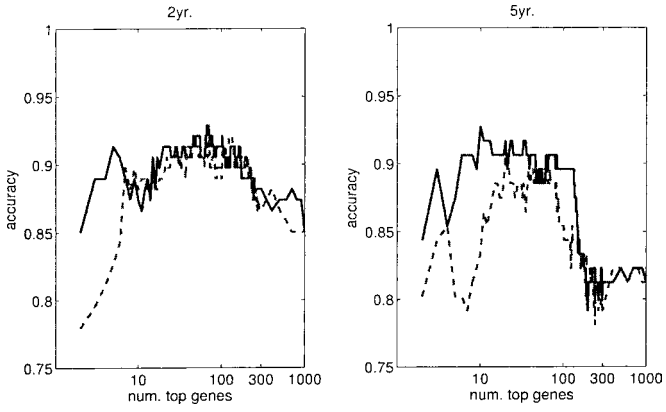

図 1. 遺伝子選択に基づく，Leave One Out 性能．縦軸は予後予測精度の Leave One Out 推定値，横軸はそれを推定するのに使用した遺伝子の個数である．破線は単一遺伝子毎のスコアで遺伝子選択した場合，実線は pair wise のスコアで遺伝子選択した場合を示す．左のパネルは 2 年時，右は 5 年時予後の予測精度を示す．どちらの場合も pair wise スコアのほうが，少ない遺伝子で高い性能を示す傾向があることが分かる．

これはここでの特徴選択の過程自体が既に過学習を起こしていたものと解釈できる．

WV 法という最も単純な仮定に基づく線形判別分析でさえ遺伝子選択法次第で容易に過学習してしまうものであることを考えると，pair wise 法の発展として例えば triplet wise 法や，もっと一般的な組合せ最適化法を考える方向には慎重になるべきだろう．この問題の本質は，学習モデルの学習容量の問題が変数選択の過程において起っているところにある．次元が高い問題の場合教師付き遺伝子選択という過程自体が潜在的な複雑さを大きく持っており，十分な制限を加えておかないと簡単に過適応をきたすのである．計算量爆発のような分かりやすい困難ばかりではない．

3.2 判別性能推定における Leave Two Out 法の意義

新規症例に対する判別性能の高さを示すことが，この論文における最も重要な主張点であった．2 章で述べたとおり，これまでのこの分野では「言い過ぎ研究」が多過ぎたせいで反省気分が拡がっていたこともあり，判別性能推定の根拠の示し方には心を砕く必要があった．

我々が採用した方法(Leave Two Out 法)はモデル選択過程などで頻繁に用いられる Leave One Out 法によるクロスバリデーションを，モデル選択過程の内側と外側の二重で行う方法である．具体的には以下のような過程を踏む．

(1) 全データからテスト用データを一個抜き，残りを外側学習用データとする．
(2) 外側学習用データに基づいて分類器を構成する．
  (2-1) 外側学習用データから内側テスト用データを一個抜き，残りを内側学習用データとする．
  (2-2) 内側学習用データに基づいて遺伝子ランキングをつける．
  (2-3) 内側学習用データ，遺伝子ランキング，モデル(複数候補)に基づいて，分類器を学習し構成する．
  (2-4) (2-3)で構成された分類器(複数候補)で内側テスト用データのラベルを予測する．
  (2-5) (2-1)–(2-4)を外側学習データの個数ぶん繰り返してラベル予測性能を評価する．
  (2-6) 複数候補分類器のなかで最も予測性能のよかったものを選択．
  (2-7) 外側学習用データに基づいて遺伝子ランキングをつけ，(2-6)で選ばれたモデルに基づいて分類器を構成する．
(3) (2-7)で構成された分類器でテスト用データのラベルを予測する．
(4) (1)–(3)を全データ個数ぶん繰り返してラベル予測性能を評価する．

外側ループ(1)–(3)だけを見れば，これは内側ループで行われている遺伝子選択，学習，およびモデル選択の過程の全てをひっくるめた全過程を一つの学習過程と考えて，これに対してただ一回の Leave One Out を行っていることと等価であるため，通常の Leave One Out 法と同様にバイアスがほぼゼロであると同時に，バリアンスの小さい判別性能推定が期待できる．また，通常の Leave One Out と同程度の信頼性が置けるのだということを読者が難しい議論無しに理解できる，というのも無視できない利点であった．

一般に機械学習に基づく教師付きラベル予測では，経験誤差最小化と，モデル正則化の二つの基準の間でバランスをとる必要があり，そのバランスを設定するためのパラメータ(正則化パラメータ)を決定する必要がある．(なお，ニューラルネットの誤差逆伝播アルゴリズムにおける早期学習打ち切り法の学習エポック数や，この研究で使用した WV 法における使用遺伝子個数は，厳密には正則化パラメータとは呼べないかもしれないが，ここでは同一視しておいて問題はない．)

したがって，判別性能を客観的に示し主張するためには，経験誤差最小化を行い，モデル正

則化パラメータの設定を行う，というプロセスの全体に対して，新規ブラインドデータを用いたバリデーションが不可欠である．汎化誤差を予測するための方法として，他に例えば様々な情報量規準が提案されているが，新規データでのバリデーションの代わりにはならない．どのような情報量規準も何らかの形で理想化された前提条件を含んでおり，そうした前提条件が目の前の実データに対して厳密に(もしくは十分に)当てはまる絶対の保証が得られることは一般には期待できないためである．

クロスバリデーション性能を最大化するモデル選択をし，なおかつ最終的な結論を出す時に選択バイアスを一切入れないことを保証するためには独立データに対するテストが必要である．ところが 2 章でも述べたように独立テスト用のデータ数が少なければ今度はバリアンスが無視できなくなる．

Leave Two Out 法では，外側の Leave One Out が独立テストの代わりである．テストに使用される実質的なサンプル数が増えるため，バリアンスが減ることが期待できる．直接テストされる対象が「ひとつの完成された予測器」ではなく「予測器を構成する過程」である，という違いに注意しなければならないが,「予測器を構成する過程」を注意深くパッキングしているおかげで選択バイアスも入らない．

ただし，Leave One Out や k-fold クロスバリデーションによる判別性能評価のバリアンスが実際のところどの程度であるのか，例えば k-fold クロスバリデーションの最適な k はいくらか，などの問題について，私の知る限りいまだ結論は出ていないようである．比較的最近の関連研究によれば，これは難しい問題であるようだ．Bengio and Grandvale (2004)によって，判別性能評価のバリアンスが判別誤差の評価方法や学習アルゴリズムにも依存するため，任意の場合にバリアンスの不偏推定量を得ることが不可能であることが示されている．さらに最近でも，Markatou et al. (2005)がバリアンス推定に取り組んでいるが，これが使える条件は必ずしも広くないようだ．全データ数が限られているという条件下において我々が現在考え得る限り最も堅実なソリューションとして Leave Two Out 法を採用したが，以上の状況から見て，これが真に最適である保証は残念ながら未だ無いと言わざるを得ない．

3.3 出力形式について

医師による臨床の現場での治療方針判断の参考になるように細胞診断の結果を適切な形で出力したいというのは，細胞診断の最終段階に関する重要な論点である．この段階でデータから知り得ることをなるべく情報を落とさずに提供したいと考えるとき，いったいどこまでが可能であろうか．また知り得る限りのことを出力しようとすることは本当に妥当であろうか．

予後予測を予後良/予後悪の二値分類問題に還元して学習した場合，新規データに対する判別結果は，スカラー実数値の判別関数として得られる．これを閾値処理すれば二値出力が得られるが，各症例が閾値に近い予後良/悪の境界線上の症例であるのか，閾値を大きく越えた予後良もしくは予後悪の典型的な症例であるのかなど，判別関数の値そのものにも臨床診断上の意味を見出し得る．そこで我々は，判別関数の出力値をロジスティック回帰によって確率の値に変換してから出力する，確率出力システムを採用した．

しかし判別出力の出し方については，様々な観点からまだ考えなければいけない問題が残っている．論文出版後に浮上した問題も含めて論じたうえで，今後の課題を考えてゆこう．

3.3.1 確率出力の問題

確率出力の解釈は難しい．

天気予報の降水確率が 30%であると知ったとき，人は傘を持っていくか否かの判断をどのように行っているだろうか．個人個人で傘を持っていく手間，降られたときの不快さ，季節，気温などさまざまな要因をきめ細かく評価して最終的には直感的に決定しているのであれば，個

人の判断において確率値を有効に解釈して活用していると言えるだろう．しかし心の中で確率値に閾値を決め，自動的に二値化してしまうこともでき，そのほうが早い．臨床医は，患者やその家族の社会生活環境やものの考え方までも含めた複数要因を総合的に考えて迅速な判断を下さなければならない状況にいる．連続値としての確率値出力は情報過多であるため，適切な閾値を決めて欲しいという意見が実務家であるユーザー周辺には根強くあるようだ．

確率出力の解釈は，よく考えてみると実は原理的にも難しい．ロジスティック回帰モデルは，変数の値が分かればラベル出力の「真の確率」が決まる，ということを仮定しており，最尤推定によるフィッティングの結果として得られた確率値は，理想的にはその「真の確率」の推定値，を意味する．「真の確率」とは，まったく同じ観測を持つサンプルを無限個持ってきたとしたときに，ラベルが 0 もしくは 1 となる頻度の比率，すなわち頻度論的確率のことである．しかしそれの推定値は，果たして頻度論的な確率だろうか，ベイズ的主観確率だろうか．筆者はどちらでもないと思うが，強いて言えば後者に近そうだと思われる．では，例えばそれはベイズ的最適行動決定のための事前確率として使って良い値だろうか．明快な回答は私の知る限り存在しない．

確率値を出力することの理論的な有効性が何らかの形で示せたとしても，臨床現場での有効活用は別の問題として残るのである．

第二は確率出力の誤りの評価が難しいという問題である．

本当は悪性である症例を，誤って 100% 良性であると予測するリスクは，60% 良性であると予測するリスクよりも大きいだろう．その差をつけることにこそ確率出力の意義があると言えるのだが，しかし現実問題としてどの程度の量の差をつけるのが良いだろうか．この重みのつけかたには自由度が残り，その決定基準はユーザーの任意ということになるが，現実問題として基準は無い．

また確率値に関して回帰問題と同様にバイアス・バリアンスのジレンマがあるが，確率値推定におけるジレンマには特殊な性質があることが分かっている．仮に真の確率 μ と，推定された確率の値 $\hat{\mu}$ の自乗誤差の期待値をリスク関数として定義すると，これはバイアスとバリアンスの和の形で書ける．

$$E[(\hat{\mu}-\mu)^2] = E[(\hat{\mu}-E[\hat{\mu}])^2] + E[E[\hat{\mu}]-\mu]^2 \tag{3.1}$$

ここにバイアスとバリアンスのジレンマが生じるが，二値分類ではもうひとつ，「確率値出力 $\hat{\mu}$ を適当な閾値 θ で二値化したときの正解率」

$$E\left[I\left(\mathrm{sign}(\hat{\mu}-\theta) = \mathrm{sign}(\mu-\theta)\right)\right] \tag{3.2}$$

という値が性能を評価するうえで重要な要素となる．ただし，$\mathrm{sign}(x)$ はスカラー値 x の符号，$I(a)$ は命題 a が真ならば 1，偽ならば 0 をとる指標関数である．この点に関して，Friedman (1996) が「確率値を二値化したときの正解率を大きくすること」と，「確率値推定のバイアスを小さくすること」とが矛盾することを示した．例えば，実際に 0.5 を閾値としたときの正解率を大きくしようして判別関数を最適化すると，バイアスが大きくなり，すべての確率値が 0.5 近辺ぎりぎりの値になってしまうか，もしくは，逆に 0 近辺と 1.0 近辺ぎりぎりの値しかとらなくなってしまい，せっかくの確率値の意味が失われてしまう．

つまり，二値分類問題に対する出力として，二値的出力の正解率だけが欲しいものなのであれば確率的出力に色目を使わないほうがよいようだ．確率値が欲しい場合にも，本当に連続値の事後確率が欲しい場合は少ないと思う．むしろ予後悪中良のような三値，四値に量子化することで分かりやすくバイアス・バリアンスともに適当な大きさに抑えられた予測が得られるかもしれない．知りたいことを全て知ろうとすることはできないのであるから，知りたいことを

あらかじめ絞っておくことが重要である，ということのようである．

3.3.2 それ以外の出力

各例ごとでの予測の信頼性の指標として，例えばデータの新奇性は重要なファクターとなり得る．これまでに学習で得られたどのパターンにも似ていない入力が来たとすれば，出力の推定の信頼性は低いものと考えたくなるだろう．前節では判別関数に基づく判別境界からの距離を考慮した確率出力について詳しく議論したが，データパターンが新奇である場合に関しては前提の崩れ(癌を対象として良性悪性を区別しようとしているのに，症例が癌細胞でない，測定が失敗しているなど)を想定しなければならなくなるため，臨床応用上も別の取り扱いが必要になるだろう．

予後の良し悪しの予測を二値分類の枠組みではなく，生存時間解析の枠組みで行う試みもある．例えば比例ハザードモデルに基づく手法(Gui and Li, 2005; Li and Luan, 2005)などでは，各症例について悪性度をハザードスコアで出力して示した．前節で述べたように臨床現場で使用できるかどうかは別の問題として残るが，データから知り得ることをなるべく情報を落とさずに提供したいという立場からは積極的に試してみるべき手法であると思われる．

多値分類の枠組みも，実際の細胞診断の上で応用範囲が広そうである．「診断時には得られないがフォローアップによって得られる多値分類ラベル」があって，それと予後との間に関係があるのであれば，多値の分類ラベルを当てにいって，それに基づいて診断を行うという方向もあるだろう．

さらにこれらを混ぜ合わせる方法もいろいろ考えられるであろうが，前節の最後に述べたように，知り得ることは限られているため，あらかじめ知りたいことを絞ってゆく努力が常に必要とされるであろうことに注意をしておく．

4. 境界線を押し上げる

4.1 Semi-supervised Learning

2 章において教師無し学習と教師付き学習の応用と，その関係について述べた．近年では，教師ラベル付きデータ点と教師ラベル無しデータ点を併用した場合のデータ解析法(semi-supervised learning)も機械学習の理論・応用における重要なテーマとなっている．ウェブデータマイニングなどでは，巨大サンプル数のラベル無しデータを参照しながら，比較的少数のアノテーション情報付きデータをもとにモデルを学習させたいニーズがある．それに対してマイクロアレイデータ解析では，少し異なる観点から教師無しと教師有り状況を混合することが有効である．

Bair and Tibshirani (2004)は，semi-supervised 手法という用語を使って，以下のような解析を行っている．この論文で提案している手法は，

(1) 症例のラベルを使って教師付き遺伝子選択をして遺伝子を絞る
(2) K 平均法 $K=2$ によるクラスタリングを行う
(3) (1)(2)をあわせた過程のクロスバリデーションを行う

というものである．

これは一見したところとくに新しいところのない簡単な方法であって，なおかつある程度経験を積んでいる者から見れば，「言い過ぎ解析」の典型例そっくりである．教師付き遺伝子選択を行ったうえで K 平均法クラスタリングを行えば，ほぼ教師ラベルどおりのクラスタリング結果が得られるのは当たり前である．したがって，もしも「クラスタリングの結果として発見したクラス分類が，既知の分類と相関した！」などと結論付けたらトートロジーである．査読者も迷ったことであろう．

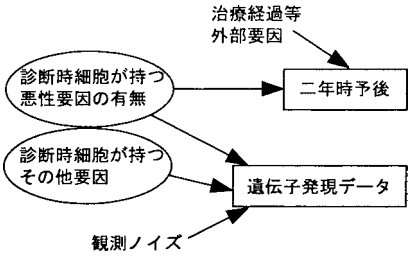

図 2. 予後関連情報のうち，遺伝子発現量データからの推定によって知り得るのは，診断時細胞が持つ悪性要因の有無までである．

しかし「教師付き遺伝子選択」と「教師無しクラスタリング」とを合わせた全体がひとつの「教師付き分類」の過程となっているのであり，クロスバリデーションが(1)と(2)の解析プロセス全体に対してかかっていることもあって，この解析は実は落し穴にはまっていない全く正当なものだ．

「教師無しクラスタリング」によるクラス発見の意味は，この手法では一味違う意味になる．上でトートロジーと呼んだとおり,「教師付き遺伝子選択」に基づいて,「教師無しクラスタリング」を行えば，クラスタは教師ラベルと同じものが得られるのが当然であり，そうなったとしても積極的な意味は無い．しかし，もしも仮にクラスタリング結果に教師ラベルと異なるものが出てきたとすれば，逆にそこに有効な意味が生じてくる．ここは注目すべき点である．

さらに，教師付きよりも教師無しの分類結果のほうが的確な結果をもたらしている可能性もある．この点をもう少し詳しく考えてみよう．

細胞自体を観察することによって潜在的に知り得る「悪性度の違い」と，実際に観測される「予後の悪さ」との間の因果関係は直接的なものではないと考えられる．遺伝子の発現量から知り得ることは，癌の悪性度を増す要因のなかでもデータ取得時点で細胞に現れている限りのものだからである．図 2 は，その事情を簡単なグラフで表現したものである．例えば癌に関する 2 年時生存/死亡の 2 値ラベルを考えるとき，これはデータ取得時点以後の治療経過その他の環境条件にも依存するであろうし,「2 年時点」というのは恣意的な時期設定であるから，それを遺伝子が完璧に反映することは考えにくい．そこで発現による細胞診断が必要になるような問題では，教師付き学習によって 100%近い正解率を得るような状況は考えにくい．教師ラベル情報は,「細胞診断によって知り得る限りの知りたいこと」とは少しずれているはずで，それがもし遺伝子の発現量に反映されているならば，教師有りよりもむしろ教師無しで見るほうが的確である．

なおこの論文では，別法として,(2)の過程で，$K=2$ の K 平均法を用いる代わりに，主成分 1 次元のみによる主成分分析を行い，その主軸への投影を判別関数とした判別分析を行う，という手法も提案していた．つまり教師ラベルを用いて遺伝子選択を行った後で教師無しで 1 次元の主成分分析を行うため最終的に得られた 1 次元の主成分は教師ラベルを間接的に反映し，教師付き学習における判別関数に対応するものとなる．これは通常の線形判別分析と通常の主成分分析の中間をとったものとなっている．K 平均法と同様の意味で注目に値するが，彼らが調べた限りでは教師ラベルの正解率だけを見たときには多少見劣りのする結果が得られたようだ．

このように semi-supervised 状況を巧妙に使うことによってクラス発見に関して統計的に言えることが増えてきそうである．

4.2 多重検定における最適統計量

もうひとつ注目すべき研究を紹介する．多重検定における遺伝子選択基準について最適性を定義し，実際に最適な基準(ODP，Optimal Discovery Procedure)を提案するというもので，解析実務者向け(Storey et al., 2005)と，理論家向け(Storey, 2005)の 2 バージョンのテクニカルレポートが公開されている．

多重性の無い検定における最適性については，検定の強さという概念があり Neyman-Pearson の補題によって裏付けられているが，ODP はこれを多重検定に拡張する．

各遺伝子 i に対してその複数細胞サンプルにおける発現量 $X_i=(x_{i1},\ldots,x_{iM})$ が得られているとき，これに基づいて遺伝子 i が帰無仮説 H_0 に従う非活性遺伝子であるか，もしくは対立仮説 H_1 に従う活性遺伝子であるかを判定する問題を考えよう．対立仮説および帰無仮説が尤度関数 $f(X_i)$ および $g(X_i)$ で定義されるとき，尤度比 $h(X_i)=g(X_i)/f(X_i)$ を統計量としてしきい値処理することで最強な判定基準が得られることを，Neyman-Pearson の補題は保証している．ここで最強とは，判定基準「$h(X_i)>\lambda$ ならば有意」を使用し，第一種過誤の確率が適当な値 α になるように λ を決定したときの第二種過誤の確率 β が最小，すなわちこの判定基準を下回るような他の判定基準が存在しない，ということを意味する．

多くの検定では対立仮説および帰無仮説の尤度関数が未知パラメータを持ち以下のように書ける場合

$$p(X_i|H_1)=g(X_i|\theta),\qquad p(X_i|H_0)=f(X_i|\phi) \tag{4.1}$$

を考える必要がある．たとえば H_1「X_i の期待値は 0 でない」，H_0「X_i の期待値は 0 である」という対立・帰無仮説を，「X_i の分布は正規分布である」という前提のもとで比較するとき，H_1 における期待値の具体的な値 μ や，正規分布の分散の値 σ^2 は未知パラメータとして残る．この例は $\theta=\{\mu,\sigma_1^2\}$，$\phi=\{\sigma_0^2\}$ とすることに対応する．このとき，パラメータ θ,ϕ を最尤推定値 $\hat{\theta},\hat{\phi}$ で置き換えたバージョンの尤度比を統計量として使用することで，サンプル数 $M\to\infty$ において漸近的に最強であるような検定統計量が得られる．

$$\begin{aligned} &h(X_i)=g(X_i|\hat{\theta})/f(X_i|\hat{\phi}),\\ &\hat{\theta}=\arg\max_{\theta} g(X_i|\theta),\hat{\phi}=\arg\max_{\phi} f(X_i|\phi)\end{aligned} \tag{4.2}$$

さて，これを多重検定の場合に拡張する前に，これまで有意遺伝子選択を目的として行われてきた多重検定に関する研究を概観しておく．これまでに得られた成果は大きく分けて，多重性に基づく困難を克服するための研究と，多重性を積極的に利用するための研究との両面がある．

多重性に基づく困難とは，各個検定における有意性指標として p 値が使えないということに尽きる．最も保守的なものとしては Bonferroni の補正，すなわち p 値に多重性(検定対象遺伝子数)を掛け算したものについて 5% や 1% の有意水準で評価する，というものがあるが，多重検定全体でひとつひとつ有意か否かを決定した結果を集計した偽陽性比率(FDR, False Discovery Rate)を見て，それをコントロールするべきであるとする手法(Storey and Tibshirani, 2003)が便利とされ，受け入れられてきており，またその具体的な計算方法についての研究が進められている．

多重性を積極的に利用する研究では，パラメトリック尤度のパラメータを各遺伝子毎に考えねばならない状況，

$$p(X_i|H_1)=g(X_i|\theta_i),\qquad p(X_i|H_0)=f(X_i|\phi_i) \tag{4.3}$$

において，異なる遺伝子 $i=1,2,\ldots$ の間で，パラメータ θ_i,ϕ_i の値がなんらかの関係を持って

いることを利用する．これによって得られた検定による FDR が，遺伝子それぞれ独立に有意性の度合を計算する場合に比べて改善したならば，多重性を利用できたことになる．これまでに，各遺伝子毎の群内分散推定を縮小推定で安定化することで効果を上げたもの（Tusher et al., 2001），ベイズ的階層モデルを用いてパラメータの事前分布を階層的にさらに推定して用いることで効果を上げたもの（Efron and Tibshirani, 2002; Lonnstedt and Britton, 2005），パラメトリックな検定統計量のパラメータを適応的に調整することで効果を上げたもの（Mukherjee et al., 2005），などがある．

Storey（2005）の ODP は，こういう背景を踏まえて提案されたものであり，これまでモデルを入れたりパラメータを調整してみては良かった悪かったと言っていたところに，一般的な最適性の基準を持ち込んだところが画期的だと思われる．ODP は，Neyman-Pearson の補題における最強性の定義を多重検定に拡張したものであり，期待擬陽性数（EFP，Expected False Positives）を固定したうえで，期待真陽性数（ETP，Expected True Positives）を最小にする判定基準を，ODP と定義する．さらにこれは以下のようにして陽な形で得ることができる．

ODP の補題．全遺伝子 $i=1,\dots,N$ に対して共通の統計量，

$$S_{\mathrm{ODP}}(X)=\frac{\sum_{j\in G_1} g(X|\theta_j)}{\sum_{j\in G_0} f(X|\phi_j)} \tag{4.4}$$

に対して，適当な値 λ をしきい値として作った判定基準「$S_{\mathrm{ODP}}(X_i)>\lambda$ ならば遺伝子 i は有意」に基づく多重判定は，「これと EFP を同じにするような他のどの判定基準にも ETP において負けることがない」という意味で最適である．ただし，G_1 は対立仮説が成り立つ遺伝子の集合，G_0 は帰無仮説が成り立つ遺伝子の集合を指す．

証明．この補題の証明は，Neyman-Pearson の補題に帰着させることによって得られる．詳細は Storey（2005）を参照．

多重検定を実際に使用する場面では，G_1, G_0 が不明であり（これが既知ならば遺伝子選択をする必要はない），また対立仮説・帰無仮説が成り立つ遺伝子における尤度関数の真のパラメータ θ_j, ϕ_j も未知であるため，観測によって得ることのできる情報のみを用いた近似が必要である．例えば以下のような近似が可能である．

$$\hat{S}_{\mathrm{ODP}}(X)=\frac{\sum_j g(X|\hat{\theta}_j)}{\sum_{j\in \hat{G}_0} f(X|\hat{\phi}_j)}, \tag{4.5}$$

ここで各遺伝子 j における真のパラメータ θ_j もしくは ϕ_j の代わりに最尤推定量 $\hat{\theta}_j=\arg\max_\theta g(X_j|\theta), \hat{\phi}_j=\arg\max_\phi f(X_j|\phi)$ を用いている．右辺分子の和については対立仮説が成り立っているいないに関わらず全ての遺伝子で和をとり，分母の和については通常の尤度比検定によって帰無仮説を棄却されなかった遺伝子 $\hat{G}_0=\{j|g(X_j|\hat{\theta}_i)/f(X_j|\hat{\phi}_i)>\epsilon\}$ について仮に対立仮説が成り立っているものとみなして和をとっている．

彼らはこれを人工データ及び，実データに適用して有効性を示した．全体的な傾向としては，候補遺伝子数が多く多重性が高いときや，対立仮説が成り立つ場合のモデルが少数のパターンに偏っているなど対称性が悪いときには，旧来手法に比べて ODP は FDR 固定のもとで選択される遺伝子数を倍近くにするなど，劇的と言ってよいほどの性能改善を示した．そうでないときには旧来手法に近い性能を示すこともあったが，常に同等以上の性能を示した．

私の理解によれば，ODP による遺伝子選択は，前節で述べた広い意味での semi-supervised 状況を積極的に利用することによって性能を高めている．つまり，教師ラベルの裏に，ラベルとは少し違った複数（しかし遺伝子数よりは少数）の要因があって，遺伝子はそれを通して教師

ラベルとの相関を持っているのである．そのため，ラベル自体と比較した近さのみならず，まわりの遺伝子との近さを使うことで遺伝子発現の活性が精度高く当たるようになっているのである．教師ラベルを使用しつつも，教師無しの意味で周辺の情報を上手に援用している点が semi-supervised 的である．もしも全遺伝子がそれぞれ完全に独立にラベルと相関を持っていたならば，多重検定を完全に独立に使用する旧来手法との違いはほとんど生じない．

5. むすび

知りたいことをデータに反映させて読み取ろうとすることを「革新的」，データから言えることだけに限定して言おうとすることを「保守的」と呼ぶならば，こうした「保守」と「革新」との間の矛盾と軋轢は，個人の中で，研究グループのメンバー同士で，各論文の査読プロセスで，論文誌の編集方針で，世界の研究全体の流れの中で，など様々なレベルで起こり続けている．「革新的」な主張をしようとすれば，常に言い過ぎの危険がある．意図的な捏造と，意図しない言い過ぎとは全く異なるものであるが，科学そのものへの信頼性へのダメージという意味では，よく似た効果を持ち得る．しかし「保守的」な主張に終始するだけでは科学は進まない．第二章で述べたように，言い過ぎの境界を超えてしまう解析結果が Nature，Science といったインパクトの高い論文誌に掲載され続けてしまったことの最大の理由は，生物学者・医者にとってそれこそが知りたいことであったからである．「革新的」であり過ぎてもいけないが，「保守的」であり過ぎるのもまずい．言い過ぎの境界をきっちりと理解し意識することは必要であるが，それを前提としてあえて可能性レベルの議論を行うことも重要であろう．

本論ではとくに「保守的」であることが強く求められる応用分野として，遺伝子発現に基づく臨床的細胞診断と，そこでのクラス発見，クラス分類に関する研究を紹介した．その一方で遺伝子制御ネットワークの推定の研究は，比較的「革新的」であることが許され，また望まれる応用分野であるはずであった．しかし，本論ではほとんど述べられなかったが，近年になって遺伝子個々間のネットワーク推定を一旦棚上げにして，遺伝子群をまとめてモジュールと呼んだうえで，それらの間の関係を捉えようとする動きがさかんである．これは，精細度を粗くしてもいいから，得られる結果の信頼性を高めようという保守方向へのゆりもどしであると解釈できる．どちらの分野においても，言い過ぎの境界がどこにあるかについて信頼のおける処方箋が必要とされており，研究の焦点になっているのである．

第四章では，最近になって言い過ぎの境界をじりじりと押し上げる力が台頭しつつあることを示した．遺伝子発現解析のようにサンプルの個数よりも変量（遺伝子）の個数がはるかに多い問題では，変量間での情報の冗長性の取り扱いが肝要である．掬い上げられていない情報がまだそこに残っており，「言い過ぎ」と「言えること」との間の境界を上方修正する余地がまだまだ大いにあるらしいということがこれらの研究から示されたと考えている．

第一章の終わりに述べたように，特徴次元が高すぎる問題は遺伝子発現データ以外にも山積している．遺伝子発現解析の手法開発に対して今もなお続けられている努力は，大きな普遍性を持つものだと思われる．統計学に課せられた責任と寄せられる期待が非常に大きいものであることをデータ解析者の立場から伝えることができれば，本論の目的は達せられたものと思う．

参 考 文 献

Bair, E. and Tibshirani, R. (2004). Semi-supervised methods to predict patient survival from gene expression data, *PLoS Biology*, **2**, 511–522.

Bengio, Y. and Grandvalet, Y. (2004). No unbiased estimator of the variance of k-fold cross valida-

tion, *Journal of Machine Learning Research*, **5**, 1089–1105.

Bo, T. and Jonassen, I. (2002). New feature subset selection procedures for classification of expression profiles, *Genome Biology*, **3**, R17.

Doi, A., Fujita, S., Matsuno, H., Nagasaki, M. and Miyano, S. (2004). Constructing biological pathway models with hybrid functional Petri nets, *In Silico Biology*, **4**, 271–291.

Efron, B. and Tibshirani, R. (2002). Empirical bayes methods and false discovery rates for microarrays, *Genetic epidemiology*, **23**, 70–86.

Eisen, M., Spellman, P., Brown, P. and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns, *Proceedings of the National Academy of Sciences USA*, **8**, 14863–14868.

Friedman, J. H. (1996). On bias, variance, 0/1-loss, and the curse-of-dimensionality, *Data Mining and Knowledge Discovery*, **1**, 55–77.

Friedman, N. (2004). Inferring cellular networks using probabilistic graphical models, *Science*, **303**, 799–805.

Garge, N. R., Page, G. P., Sprague, A. P., Gorman, B. S. and Allison, D. B. (2005). Reproducible clusters from microarray research: Whither?, *BMC Bioinformatics*, **6**, Suppl 2:S10.

Gui, J. and Li, H. (2005). Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data, *Bioinformatics*, **21**, 3001–3008.

Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Element of Statistical Learning*, Springer, Berlin.

Li, H. and Luan, Y. (2005). Boosting proportional hazards models using smoothing splines, with applications to high-dimensional microarray data, *Bioinformatics*, **21**(10), 2403–2409.

Lonnstedt, I. and Britton, T. (2005). Hierarchical Bayes models for cDNA microarray gene expression, *Biostatistics*, **6**(2), 279–291.

Markatou, M., Tian, H., Biswas, S. and Hripcsak, G. (2005). Analysis of variance of cross-validation estimators of the generalization error, *Journal of Machine Learning Research*, **6**, 1127–1168.

Michiels, S., Koscielny, S. and Hill, C. (2005). Prediction of cancer outcome with microarrays: A multiple random validation strategy, *Lancet*, **365**, 488–492.

Mukherjee, S., Roberts, S. J. and van der Laan, M. J. (2005). Data-adaptive test statistics for microarray data, *Bioinformatics*, **21**, Suppl. 2:ii108–ii114.

Muro, S., Takemasa, I., Oba, S., Matoba, R., Ueno, N., Maruyama, C., Yamashita, R., Sekimoto, M., Yamamoto, H., Nakamori, S., Monden, M., Ishii, S. and Kato, K. (2003). Identification of expressed genes linked to malignancy of human colorectal carcinoma by parametric clustering of quantitative expression data, *Genome Biology*, **4**, R21.

Ntzani, E. E. and Ioannidis, J. P. A. (2003). Predictive ability of DNA microarrays for cancer outcomes and correlates: An empirical assessment, *Lancet*, **362**, 1439–1444.

大羽成征，石井　信(2004). **高次元超球面上のカーネル密度推定とモード検出，電子情報通信学会技術研究報告** NC2004-28, **104**, 1–6.

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

Storey, J. (2005). The optimal discovery procedure: A new approach to simultaneous significance testing, *UW Biostatistics Working Paper Series*, Working Paper 259, http://www.bepress.com/uwbiostat/paper259/.

Storey, J. D. and Tibshirani, R. (2003). Statistical significance for genomewide studies, *Proceedings*

of the National Academy of Sciences USA, **100**, 9440–9445.

Storey, J., Dai, J. and Leek, J. (2005). The optimal discovery procedure for large-scale significance testing, with applications to comparative microarray experiments, *UW Biostatistics Working Paper Series*, Working Paper 260, http://www.bepress.com/uwbiostat/paper260/.

田村隆明(2000). **『新 転写制御のメカニズム』, 羊土社, 東京.**

Tilstone, C. (2003). DNA microarrays: Vital statistics, *Nature*, **424**, 610–612.

Tusher, V. G., Tibshirani, R. and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response, *Proceedings of the National Academy of Sciences USA*, **98**, 5116–5121.

van 't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A. M., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T., Schreiber, G. J., Kerkhoven, R. M., Roberts, C., Linsley, P. S., Bernards, R. and Friend, S. H. (2002). Gene expression profiling predicts clinical outcome of breast cancer, *Nature*, **415**, 530–536.

Yoshioka, T., Kawase, N. and Ishii, S. (2002). Correlation-based cluster analysis using mixture of constrained PCAs, *Genome Informatics*, **13**, 256–257.

吉岡 琢, 石井 信(2002). **制約付き混合主成分分析による時系列データのクラスタリング, 第** 5 **回情報論的学習理論ワークショップ**(IBIS2002), 196–201.

Proceedings of the Institute of Statistical Mathematics Vol. 54, No. 2, 405–423 (2006) 423

Prediction and Estimation from Gene Expression Data Analysis: What We Want to and be Able to Conclude from Data

Shigeyuki Oba

Nara Institute of Science and Technology

Gene expression profile data is becoming increasingly important in clinical and biological research, and statistical analyses are applied from many aspects. In such research, it is necessary to draw appropriate conclusions from data, and at the same time avoid drawing inappropriate conclusions. In recent years, many supervised analyses on cell diagnosis of human diseases, e.g., cancers have tended toward overstatement. This review paper introduces some of the pitfalls into which ordinary analysers of gene expression profile data have tended to fall. It also introduces our research on cancer diagnosis as an example of supervised analyses that have carefully avoided such pitfalls. On reflection to many researches including overstatements, conservative researches have been increasing, however, in recent years, some aggressive and clever methods are appearing which approach the border and keep in the border. We also discuss future tasks in this field based on recent ideas.

Key words: Gene expression data analysis, supervised learning, semi-supervised learning, supervised feature selection.