バイオインフォマティクスにおける統計的課題について

江口 真透 （統計数理研究所）

遺伝子解析の技術は大きく向上し，現在では 1 度の実験で数万の遺伝子データを得ることはもはや日常的に使われる技術となった．遺伝子発現を調べるマイクロアレイもその一つである．技術・研究の発展にともない，実際の臨床への応用も盛んに勧められているが，まだまだ明らかにされていない問題点も多い．マイクロアレイはその次元数 p が入手可能な標本数 n よりも極めて大きい．このように立ちはだかる困難な点は‘ 高次元データ小標本問題 ’である．バイオテクノロジーの更なる進展はデータ次元 p をより高くさせる一方で，観測数 n は無制限に大きくすることはできない．この ‘$p \gg n$’ の設定の下で正当化される研究は，統計学の全般の分野に渡ってほとんどなかった．この $p \gg n$ 問題を契機として，統計学や機械学習からの新しい方法論の提案が生まれつつある．

広い意味でのゲノムデータをバイオマーカーと捉え，医学の文脈で疾病のサブタイプや薬剤感受性，奏功性の度合いなどをあらわす表現形との相関研究について考えたい．ゲノムデータから予め治療薬の治療の効果や副作用の程度が分かればいわゆるオーダーメイド医療の実現につながる．典型的なアプローチは判別分析である．ゲノムデータの中から表現形に強く関連するパタンを抽出することが目的だ [6]．ゲノムデータから得られた p 次元の特徴ベクトルを $\boldsymbol{x} = (x_1, \cdots, x_p)$ とし，有限個のカテゴリーで表された表現形のクラスラベルを y と書く．特徴ベクトル $\boldsymbol{x}$ は，はじめに述べられたような SNP のタイプや遺伝子発現量やプロテオームの質量スペクトルのピーク値である．これよりゲノム情報に基づく $\boldsymbol{x}$ から y を高性能で予測できれば医療に様々な福音をもたらす．

マイクロアレイを使った判別分析は 1999 年 Golub *et al.* による急性骨髄性白血病（AML）と急性リンパ性白血病（ALL）という白血病のサブグループ分類によりマイクロアレイの臨床への応用の可能性が提案された後，さまざまな病気のサブグループ分類や，予後の予測などにその応用が試みられている [4]．van't Veer *et al.*(2002)[5] は初期の乳がん患者を対象とした予後予測の方法を提案し，判別に用いる遺伝子として予後との相関の高い上位 70 遺伝子を用いた判別モデルを提案した．その後も乳がんの予後予測に関して Paik *et al.*（2004）[6] による再発スコア，Chang *et al.*(2005)[7] による傷害応答モデル，Sorlie *et al.*(2001)[8] による固有サブタイプ，Ma *et al.*(2004)[9] による Two-gene ratio モデルなどいくつかのモデルが提案された. そして,van't Veer *et al.* の 70 遺伝子を使った遺伝子発現による乳がんの予後予測は MammaPrint としてキット化され, 2007 年マイクロアレイを使った診断キットとしては初となる FDA の承認を得ている．このように臨床への応用が進む一方, その問題点も指摘されている．

医学系のジャ-ナルでは多くの相関研究では非常に初等的な線型判別の方法が採られている．例えば，p 個の遺伝子発現量ごとに 2 つの表現系によって 2 標本に分解して p 個の一様性の検定統計量の内で P 値の小さい順に選んで線型判別関数を構成している．例えば上で比較された 5 つの方法は全てこのような初等的な方法で行われている．しかし同じ表現形の予測に使われたにもかかわらず共通する発現遺伝子がほとんどないことが指摘されている．バイオインフォマティクス系のジャーナルでは機械学習による非線形で高度なパタン認識の方法などが採られている．機械学習では‘ データに基づいて推測する ’ことに代わって‘ データを学習する ’という表現が使われ，パタン認識でも，脳の持つ ‘予測する本能 ’というべき観点から研究が進められている．アダブーストとサポートベクターマシンが急速に適用範囲を拡大し汎用な方法論としてパタン認識の流れを大きく変えてしまった. 統

計学の分野においても，その学習アルゴリズムの統計的性質の検討から深い理解が得られて新たな提案がなされた．特にアダブーストを含むブースト法がゲノムデータのパタン認識には相性が良い．予測のルールは線型判別の場合と基本的には同じであるが関数の埋め込みを利用して異なる幾つかの予測のルールの統合を行う．実際はブースト法はスコアー関数が与えられたときにそれぞれの識別子の線型結合によって単一の判別関数を構成する．結果としてブースト法は t 個の異なる予測のルールの統合を重み付け多数決を行っている．

例えば遺伝子発現 p 個の中で j 番目の遺伝子発現量 x_j に対して $\mathrm{sign}(\beta_{j1}x_j + \beta_{j0})$ $(j = 1, \cdots, p)$ を考えればこれは単一遺伝子発現の予測ルールとなる．ブースト法はこれらを適切にトレーニングデータによって逐次学習をさせてより強力な判別ルールを作ることができる．その最終的な結果を見ると，どの遺伝子がどう予測に関与したかは，その遺伝子に関与する単一遺伝子の識別子を集めると分かるので，判別だけでなく，どの遺伝子が表現形とどのくらい関連しているかについても見ることができる．サポート・ベクターマシンはこの点は弱点となっていて，どの遺伝子が関わっているかは自明ではない．実際にはブースト法は単一遺伝子発現ルールの係数を変えたものを幾つか用意して，これらも統合することによって線形モデル (1) でなく一般化線形モデルの形まで柔軟な非線形性を得ることができる．しかしながら，未だに安定したバイオインフォマティクスの方法は確立したと言えない．この発表では，古典的な線型判別の方法とブースティングの方法と幾つかの比較をしながら最近の統計学の中で発表された成果について紹介しながら現状報告をしたい．

References

[1] Lander, E.; Linton, L.; Birren, B.; Nusbaum, C.; Zody, M.; Baldwin, J.; Devon, K.; Dewar, K.; Doyle, M.; FitzHugh, W. & others Initial sequencing and analysis of the human genome Nature, 2001, 409, 860-92

[2] Venter, J.; Adams, M.; Myers, E.; Li, P.; Mural, R.; Sutton, G.; Smith, H.; Yandell, M.; Evans, C.; Holt, R. & others The Sequence of the Human Genome Science, 2001, 291, 1304-1351

[3] Allison, D.; Cui, X.; Page, G. & Sabripour, M. Microarray data analysis: from disarray to consolidation and consensus Nat Rev Genet, 2006, 7, 55-65

[4] Armstrong, S.; Staunton, J.; Silverman, L.; Pieters, R.; den Boer, M.; Minden, M.; Sallan, S.; Lander, E.; Golub, T. & Korsmeyer, S. MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia Nature Genetics, 2002, 30, 41-47

[5] van't Veer, L.; Dai, H.; van de Vijver, M.; He, Y.; Hart, A.; Mao, M.; Peterse, H.; van der Kooy, K.; Marton, M.; Witteveen, A. & others (2002) Gene expression profiling predicts clinical outcome of breast cancer Nature, Mass Med Soc, 415, 530-53

[6] Paik, S.; Shak, S.; Tang, G.; Kim, C.; Baker, J.; Cronin, M.; Baehner, F.; Walker, M.; Watson, D.; Park, T. & others (2004) A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N Engl J Med, 351, 2817-26

[7] Chang, H.; Nuyten, D.; Sneddon, J.; Hastie, T.; Tibshirani, R.; Sorlie, T.; Dai, H.; He, Y.; van't Veer, L.; Bartelink, H. & others (2005) Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival Proceedings of the National Academy of Sciences, National Acad Sciences, 102, 3738-3743

[8] Sorlie, T.; Perou, C.; Tibshirani, R.; Aas, T.; Geisler, S.; Johnsen, H.; Hastie, T.; Eisen, M.; van de Rijn, M.; Jeffrey, S. & others (2001) Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications Proceedings of the National Academy of Sciences, National Acad Sciences, 98, 10869

[9] Ma, X.; Wang, Z.; Ryan, P.; Isakoff, S.; Barmettler, A.; Fuller, A.; Muir, B.; Mohapatra, G.; Salunga, R.; Tuggle, J. & others (2004) A two-gene expression ratio predicts clinical outcome in breast cancer patients treated with tamoxifen Cancer Cell, Elsevier, 5, 607-616