4 バイオ統計学

柳川　堯
久留米大学バイオ統計センター
yanagawa_takashi@kurume-u.ac.jp

バイオ統計学がなければコンピュータからの結果はただのゴミにすぎない．遺伝情報は本質的に確率関数であり，データはバイアスとバラツキに支配されているからである．本稿では，シンプソンのパラドクス，オーバーフィティング，検定の多重性等に焦点をあてバイオ統計学の意義と価値および重要性について解説する．

バイオ統計学とは

バイオ統計学（バイオスタティスティクス）の源流は1662年に刊行された「死亡表に関する諸観察」（J. Graunt 著）に発している．1800年代になってメンデルの法則や進化論が提唱され，それを計量的に検証するためF. Galton やK. Pearson が回帰の概念，相関係数，カイ二乗統計量など，少し遅れてR. A. Fisher が実験計画法，分散分析，最尤推定法などの概念を導入してバイオ統計学は一気に発展し花を咲かせた．Biometrika という数理統計学の超一流学術誌があるが，元をたどればGalton とPearson が動物遺伝学者W. F. R. Weldon を加えて1900年初頭に創刊したバイオ統計学のフォーラム誌にほかならない．バイオ統計学は数理統計学の1つの大きな源流でもあった．

バイオ統計学は，いつの時代でもその時代最先端の科学の問題に取り組む科学者と協調しつつデータと対決して，その底にひそむ法則性の認識や，予測を目的とする科学的方法論を切り開くとともにその数理を解明してきた．その事情はいまも変わらない．いま，ポストゲノム時代の中で遺伝子の機能解析が大きな科学的焦点になっている．30億個のDNA 配列の中に約3万個の遺伝子が配置されている．その遺伝子の中の特定のものの突然変異が遺伝的疾患を引き起こす可能性を持つ．疾患を持つ患者の家系とそうでない家系からデータをとり原因遺伝子を探索する逆問題，あるいは糖尿病など多因子疾患と呼ばれている疾患にはいくつかの複数の遺伝子がかかわっているが，それらの遺伝子をマイクロアレイデータから特定する問題などをめぐって世界がしのぎを削っている．遺伝子の探索・特定には，コンピュータによる莫大な計算が必要である．しかし，それだけではどうしようもない，バイオ統計学がなければコンピュータからの結果はただのゴミにすぎないというのが，日本以外の国，特に米英欧での世界の常識である．その理由は何であろうか．本稿では，この理由に焦点をあてバイオ統計学の意義と価値およびその現代的重要性について解説する．

その有効性と有用性

シンプソンのパラドクス

表-1 は，血統1と2のラットを層別して遺伝子マーカM1 とM2 と作用あり，なしの関係を要約した2つの2×2表である．血統1では，M1 とM2 の中での作用ありの割合はそれぞれ P(作用あり | $M1$) = 80/96 = 0.83，P(作用あり | $M2$) = 160/320 = 0.50．他方，血統2では P(作用あり | $M1$) = 50/502 = 0.16，P(作用あり | $M2$) = 10/462 = 0.02 である．血統1，2ともに遺伝子マーカM1 の中での作用ありの割合の方が，M2 の中での作用ありの割合より高い．

表-2 は，同じデータを血統を無視して要約した表である．(1,1) セルの数値130は，表-1の2つの層の対応する(1,1) セルの数値の和として得られる．他のセルの数値も同様である．表-2から $M1$ と $M2$ の中での作用ありの割合を算出すると P(作用あり | $M1$) = 130/598 = 0.22，P(作用あり | $M2$) = 170/782 = 0.22 となり．両者は一致する．いいかえれば，表-2からは，作用あり，なしは，

	血統 1			血統 2		
	作用あり	作用なし	計	作用あり	作用なし	計
M1	80	16	96	50	452	502
M2	160	160	320	10	452	462

表 -1　血統 1，血統 2 で層別した分割表

	作用あり	作用なし	計
M1	130	468	598
M2	170	612	782

表 -2　血統 1，血統 2 を無視した分割表

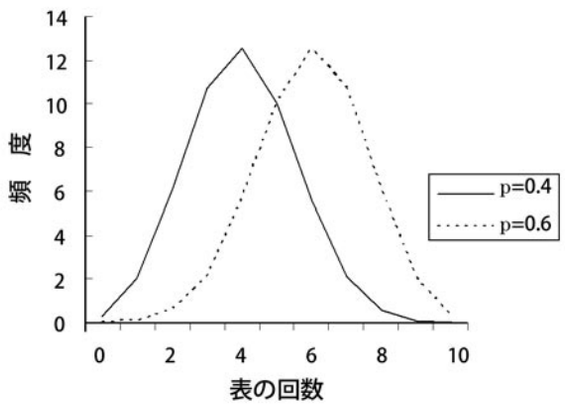

図 -1　$p = 0.4$ と 0.6 の 2 つのコイン投げ

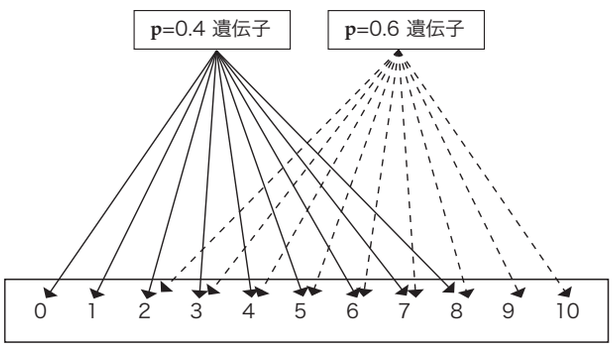

図 -2　$p = 0.4$ 遺伝子と $p = 0.6$ 遺伝子

遺伝子マーカに依存しないという結果が得られる．ラットの血統を考慮に入れて解析するか，しないかによって結果ががらりと変わる．このような現象はシンプソンのパラドクスと呼ばれている．因子によっては層別してもしなくても同じ結果を与えるものもあるし，上とは逆に層別すると関係ないが，層別しなければ関係が出てくる場合もある．一体，どのような場合に層別しなければならず，どのような因子なら層別が無視できるのか．バイオ統計学は長年このような問題について考察してきた．要因と反応のどちらにも関係する因子で層別せよというのが正解であるが，ことはそれほど単純ではない（たとえば文献1）参照）．データが持つこのような根源的問題やパラドクスに正当に対処せず，コンピュータを長時間走らせてもアウトプットされるのは見せかけの結果，つまりゴミにすぎない．

●オーバーフィティング

遺伝情報は本質的に確率関数である．したがって，たとえゲノムデータといえバラツキがある．最も単純な例を挙げれば，遺伝子データとは本質的に，おもてが出る確率が$p = 0.4$ のコインを10 回投げたときおもてが出る回数のようなものである．いま，$p = 0.4$ の遺伝子と$p = 0.6$ の遺伝子をデータに基づいて判別したいとしよう．**図 -1** は，$p = 0.4$ と$p = 0.6$ のコインをそれぞれ10 回投げるときおもてが出た回数を記録する実験を50 回繰り返したときの結果である．ヨコ軸はおもてが出た回数，タテ軸は頻度である．図から，バラツキのため，おもてが出た回数が大きくオーバーラップしている様子が分かる．**図 -2** は，図-1 を別の観点から描いたものである．$p = 0.4$ 遺伝子と$p = 0.6$ 遺伝子を判別する問題は，どちらの遺伝子からのデータか分からないデータが1つ得ら

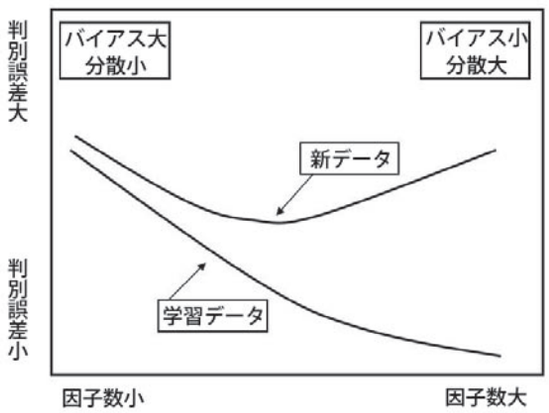

図-3　学習データと新データの判別誤差

れたとき，図-2の下の方から上の方に矢印を逆にたどってどちらの遺伝子からのデータであったかを判別する逆問題である．この逆問題は，たとえば7をデータとして得たとき$p = 0.4$遺伝子と$p = 0.6$遺伝子の両方に行く道がある，つまり対応が1対1ではない．1対1対応がないこのような問題を解くには統計学特有の考え方が必要である．その考え方をみてみたい．

まず，統計学ではバラツキを持つデータに基づく判定は必ず間違いをおかすことに留意する．次に，間違いをおかす確率を最小にするルールを求める．具体的にみてみよう．数理統計学の習慣に従って，データのバラツキを確率分布でとらえ，確率変数という概念を導入してデータを確率変数の実現値とみなす．いまr次元データを小文字ベクトル$x = (x_1, x_2, \ldots, x_r)$で表し，対応する確率ベクトルを大文字$X = (X_1, X_2, \ldots, X_r)$で表す．$Y$を$p = 0.4$遺伝子なら0，$p = 0.6$遺伝子なら1の値をとる2値確率変数とする．$X = x$を所与としたときの$Y = 1$の条件付き確率を$p(x) = P(Y = 1 \mid X = x)$とおく．$p(x) > 0.5$なら$Y = 1$，そうでなければ　$Y = 0$とする方式で判別するとする．この判別は間違いを必ずおかす．その間違いの大きさをerror $= E[L(Y, p(X))]$で評価する．これを判別誤差という．ただしEは確率変数XとYに関する期待値，LはYと$p(X)$の隔たりを表す関数，たとえば次のようなクロス・エントロピー関数である．

$$L(Y, p(X)) = -2\left(I(Y = 0)\log(1 - p(X)) + I(Y = 1)\log(p(X))\right)$$

さて，N個のr次元データ$x_1, x_2, \ldots, x_N$と対応するYの実現値$y_1, y_2, \ldots, y_N$が与えられたとする．このようなデータを教師つき学習データという．このとき判別誤差を最小にする$p(x)$は判別誤差の推定関数

$$\overline{error} = \frac{1}{N}\sum_{i=1}^{N} L(y_i, p(x_i))$$

を最小にする$p(x) = \hat{p}(x)$で与えられる．さて，ここからが本番である．実は，判別誤差は$p(x)$を複雑にすればするほど小さくできるのである．たとえば，$x = (x_1, \ldots, x_r)$をr次元ベクトルとして，$p(x)$をロジスティックモデルとよばれる次の関数のクラスに属するとするとき，rの数，つまり考慮に入れる因子x_jの数を増やせば増やすほど判別誤差は小さくできる．

$$p(x) = \exp(\beta_1 x_1 + \cdots + \beta_r x_r)\big/\left(1 + \exp(\beta_1 x_1 + \cdots + \beta_r x_r)\right)$$

したがって，判別誤差を小さくすることを目的としてきたが，それが満たされたからといって喜べないのである．なぜか．問題は，新しいデータの判別に対してこのルールがどの程度の大きさの判別誤差を持つかを評価することであって，ルール作りのもとになった教師つきの学習データに対する評価ではないからである．つまり学校での成績が良くても社会に出たときの成績が必ずしも良いとは限らないということである．**図-3**に，ロジスティックモデルの場合に，タテ軸に判別誤差，ヨコ軸に因子数を与え，学習データに対する判別誤差と新データに対する判別誤差の模式図を与えた．図より新データに対する判別誤差は，ある点までは因子数を増やせば減少するが，その点を超えれば逆に増加する様子が分かる．バイアスと分散のトレードオフが起こっているからである．この現象をオーバーフィティング（overfitting）

という．オーバーフィティングは，機械学習による遺伝子の探索において特に注意が必要である．きわめて小さい例数にもかかわらず考慮される遺伝子の個数がとても多いからである．バイアスが大きいモデルからの結果はゴミにすぎない可能性が大きい．オーバーフィティングに対処するには，図-3で新データに対する判別誤差が減少から増加に移るときの点を因子数に持つモデルを用いればよい．この点を見つける一般的な方法として，たとえばAIC（Akaike Information Criterion），BIC（Bayes Information Criterion），CV（Cross Validation）などモデル構築技法と呼ばれるさまざまな技法が開発されている（文献2)参照)．

●統計的検定とその多重性

健常人とある薬剤投与グループを比較し，薬剤投与によってどの遺伝子が動いたか，を調べる研究では遺伝子ごとにグループ間の比較が行われる．このとき，帰無仮説H_0：遺伝子A_1は動いていない，を対立仮説K_1：遺伝子A_1は動いた，に対比させて有意水準αの統計的検定が行われることが多い．ここで，有意水準αの検定とは，帰無仮説が正しいにもかかわらず対立仮説が正しいと間違って判定する，第1種の誤りと呼ばれる誤りの確率をαに設定しておき，対立仮説が正しいにもかかわらず帰無仮説が正しいと間違って判定する確率を最小にする検定方式のことである．たとえば，200個の遺伝子について，有意水準5%の検定を適用して両グループを比較すると，200回の検定を互いに独立な1セットの検定とするときの有意水準は$1-(1-0.05)^{200}=0.999965$である．動いていない遺伝子を動いている（偽陽性）と誤判定する確率を99.9965%も見込んだ比較を行ってもひっかかってくるのはゴミばかりということになる．この問題は統計的検定の多重性と呼ばれている．統計的検定をよく理解して適用しなければこのような愚をおかす．なお，多重性の対処の仕方はいろいろ提案されているがBonferroni法と呼ばれる方法では，上の場合$1-(1-\alpha)^{200}=0.05$を満たすα，すなわち0.000256を毎回の検定の有意水準と定めればよい．

バイオインフォマティクスとの接点

本章では，バイオ統計学の分野で，いま我が国で取り組まれているゲノムデータ解析に関していくつかの話題を，何が問題とされどのような研究が行われているかに焦点をあて駆け足で紹介する．いずれも急速に発展しつつある先端的話題で詳細は著者の論文やWebサイト等を見ていただきたい．

●マイクロアレイデータ

一口にマイクロアレイデータといってもさまざまなデータがある．たとえばcDNAアレイデータとは，1枚の基盤上に数千から数万の遺伝子（プローブ遺伝子）を異なるスポットとして固定させ，調べたい検体をふりかけ，検体に含まれる遺伝子たち（ターゲット遺伝子）とプローブ遺伝子とをハイブリッドさせることによって，数千のターゲット遺伝子の発現量を同時に観察して得られたデータのことであるが，そのcDNAアレイデータだけでも測定法はDNAチップ，マイクロアレイ，マクロアレイの3種類に大別できる．統計的データ解析を行うとき最も重要なのは，データの信頼性である．先進的研究者はいち早くマイクロレイデータを用いて医学研究をはじめたが，データの信頼性まで問うことは少なかった．マイクロアレイが当初高価であったこともあり同じ検体を数回繰り返し測定してみるなどのこともほとんど行われなかった．近年，ようやくマイクロアレイデータは測定段階のさまざまな過程で測定にバイアスとバラツキが生じること，測定法によって精度が大きく異なることが認識されるに至った．データをとる前に，どのような要因がデータにバイアスとバラツキを与えているのかを検討し，実験計画をたてて質のよい，信頼度が高いデータをとる必要がある．また，そのような努力を行った上でなおデータに混在するバイアス・バラツキをデータの変換や規格化を行って調整する必要がある．これらに関する研究はバイオ統計学が最も得意とする研究であって，大谷ら[3]，伊藤ら[4]によって世界に先んじる研究成果が上がっている．

●隠れマルコフモデル

バイオインフォマティクスの本質は，データからの情報の抽出にある．対象とされるデータには，疾患を持つ患者の家系とそうでない家系からとられたデータ，マイクロアレイデータ，タンパク質データなどさまざまなデータがあるが，それらはいずれもお饅頭の外側を測定したようなものでしかない．遺伝子が表現型を次世代に受け渡す以上その底には必ず一定の法則性があるはずであるが，その法則性はお饅頭の中に隠れていてみえない．隠れマルコフモデルは，目に見えないお饅頭の中をマルコフ連鎖という数学モデルでモデル化しておき，さらに観測できるお饅頭の外側との関係をモデル化してこの法則性を抽出することを狙った汎用モデルであり，おそらく遺伝子データの解析に最もよく使われているモデルである[5]．隠れマルコフモデルは，音声の認識など情報処理の世界ではよく知られたモデルであるので詳細は割愛する．

ベイシアンネットワーク

糖尿病などの多因子疾患の場合にはよく知られているが，他の多くの疾患や表現型の場合にも単独の遺伝子だけではなくいくつか複数の遺伝子がかかわっていると推察されている．マイクロアレイは1,000～20,000個の遺伝子の発現量を同時に測定することを可能とした．いま，4個の遺伝子G_1，G_2，G_3，G_4が図-4のような関係にあることとマイクロアレイで測定される$G_1, \ldots, G_4$の遺伝子発現量$X_1, \ldots, X_4$の同時密度関数が

$$f(x_1, x_2, x_3, x_4) = f_4(x_4 \mid x_2, x_3)\, f_3(x_3 \mid x_2)\, f_2(x_2 \mid x_1)\, f_1(x_1)$$

のように分解されることを1対1に対応させる．ただし，$f_4(x_4 \mid x_2, x_3)$はx_2，x_3を所与としたときのX_4の条件付き密度関数である．目的はデータから密度関数の分解を行って図-4のような図を求めることである．x_{ij}をi番目の個体から測定されたj番目の遺伝子発現量として$X_{i1}, \ldots, X_{ip}$の同時密度関数が次のように表されると仮定する．

$$f(x_{i1}, \ldots, x_{ip} \mid \theta, G) = \prod_{j=1}^{p} f_j(x_{ij} \mid p_{ij}, \theta_j),$$

ただし，$f_j(x_{ij} \mid p_{ij}, \theta_j)$は「$p_{ij} = X_{ij}$以外の$x's$」を所与とした時の$X_{ij}$の条件付き密度関数である．$G$や$\theta$は以下で明らかにされるパラメータである．$p_{ij}$のどの要素が真に条件付き密度関数に寄与しているか判定できれば図-4のようなグラフが描ける．そこでX_{ij}をノンパラメトリック加法モデルで次のようにモデル化する．

$$X_{ij} = m_{j1}(p^j_{i1}) + \cdots + m_{jq_j}(p^j_{iq_j}) + \epsilon_{ij}$$

ただしϵ_{ij}は平均0，分散σ_j^2に従う誤差である．$m's$を次のようにBスプライン関数で表す．

$$m_{jk}(p^j_{ik}) = \sum_{h=1}^{M_{ik}} \gamma_{hk} B^j_{hk}(p^j_{hk})$$

このとき，条件付き密度関数f_jは次のように与えられる．

$$f_j(x_{ij} \mid p_{ij}, \theta_j) = \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left(-\frac{\left(x_{ij} - m_{j1}(p^j_{i1}) - \cdots - m_{jq_j}(p^j_{iq_j}) \right)^2}{2\sigma_j^2} \right)$$

θ_jは右辺に含まれるすべてのパラメータを表す．N個体からアレイデータが得られた時，以上の設定から尤度関数が構成できる．そこで上述のAICを拡張した統計的モデル選択基準を適用して$f_j(x_{ij} \mid p_{ij}, \theta_j)$に真に寄与している$p_{ij}$を選択する．これで原理的に遺伝子ネットワークが描けるが，実際には約3万個の遺伝子に対してデータはたかだか100個体くらいからしか取られないので，たとえ計算ができても明らかにオーバーフィティングである．ところで，人の遺伝子の95%以上は他の生物の遺伝子と同じであるといわれている．井元ら東京大学医科学研究所のグループは事前の知識や，他の生物から得られた遺伝子間の関連性に関する知識を事前情報として組み込むベイジアンネットワークと呼ばれる手法を開発し素晴らしい成果を上げている．事前情報の組み込みは，オーバーフィティングを避ける良い方法でもある．この方法は，事前情報を事前確率密度関数$\pi(\theta \mid \lambda, G)$でとらえ，データ$\{x_{i1}, \ldots, x_{ip}\}_{i=1}^{N}$が与えられた時のネットワーク$G$の事後確率

$$\pi(G) \int \prod_{i=1}^{N} f(x_{i1}, \ldots, x_{ip} \mid \theta, G)\, \pi(\theta \mid \lambda, G)\, d\theta$$

の最大化を目的として上述の方法と同様な考えでネットワークを構成する．詳細は文献6），7）を参照されたい．

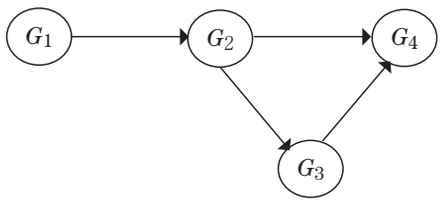

図-4　遺伝子ネットワーク

今後の展望

科学の最先端は，鋭敏に切り込むことができる新しい道具の創造と手を携えて，切り開かれる．道具の作り手としてバイオ統計学に対する期待はとても大きい．疾患や副作用に関連した遺伝子が探索・特定されれば治療法や薬剤開発の道が開ける．本稿では触れなかったが，個々人の体質を遺伝子レベルで調べ，使う薬の種類や量を選択する「オーダーメイド医療」も今後の大きな課題である．これらが進展すれば，新薬や新しい治療法の効果や副作用を調べる臨床試験が我が国でも盛んに行われることになる．臨床試験デザインの開発，効果や副作用

の評価はバイオ統計学の伝統的な守備範囲であったし，今後もバイオ統計家の活躍なしではやっていけない分野である．証拠に基づく医療の流れは奔流となり，バイオ統計学の重要性はますます高まっていくことと思われる．経済産業省が2010 年を目処に25 兆円のバイオ産業育成計画を推進していることからも明らかなように，先進各国はバイオ関連産業を国家発展の命運にかかわる次世代産業ととらえている．アメリカ合衆国には50 以上の大学にバイオ統計学科が存在しているが，その多くは10 年ほど前からポストゲノム時代に対応できる人材の養成に力を注いでいる．また，NSF（National Science Foundation）はMathematical and Statistical Science を米国科学政策6 重点領域の1つに掲げている．バイオ統計学がその主体であることは言うまでもない．

以上のほかにバイオ統計学は，地球温暖化にかかわる環境問題，環境汚染物質のヒトや動植物に対するリスク評価にかかわる問題などにも深くかかわっている．そのような分野では，情報機器の発展によって1日数ギガバイトのデータが蓄積されることもざらである．このような巨大データから情報を抽出することは人類にとって初めての経験であり，その方法論の開発がバイオ統計学に大きく期待されている．

人材の養成

バイオ統計学に対する新しい時代の巨大な要請に対して，我が国のバイオ統計学人材養成は始まったばかりである．平成15 年度科学技術振興調整経費によって久留米大学大学院医学研究科に国内で初めて定員10 名の修士課程「クリニカルバイオスタティスティクスコア人材養成ユニット」が設置された．平成17 年4 月には博士過程（定員5 名）も開設予定である．このユニットはバイオインフォマティクスを重視したバイオ統計学人材養成に重点がおかれている．ほかにも，製薬企業のサポートによって3 年前に北里大学大学院薬学研究科に臨床統計コース，2 年前に東京理科大学大学院工学研究科に医薬統計コースが設置されている．これらのコースでは，臨床試験に重点をおいた講義が行われている．また，時代の方向性を鋭く捉えた教授たちの個人的努力によって東京大学医学研究科，広島大学原爆放射能研究所，九州大学大学院数理学研究院などでバイオ統計学の博士号が数名ずつ授与されている．人材養成にはお金と時間がかかる．しかし，その重要性および今後の発展性を考えるとき重点的に早急に対策を立てていくことが次世代に対する私どもの責務であろう．

参考文献

1）甘利俊一，狩野　裕，佐藤俊哉，松山　裕，竹内　啓，石黒真木：多変量解析の展開II 章，III 章，岩波書店，東京（2002）．
2）Hastie, T., Tibshirani, R. and Friedman, J. : The Elements of Statistical Learning, Springer, Tokyo（2001）.
3）大谷敬子，大瀧　慈，佐藤健一，西山正彦，檜山圭子：遺伝子発現強度データの解析，2004 年度統計関連学会報告集，pp.180-181（2004）.
4）伊藤陽一，大橋靖雄：マイクロアレイデータを用いた任意の用量反応パターンをもつ遺伝子の探索手法の提案，2004 年度統計関連学会報告集，pp.113-114（2004）.
5）Koski, T. : Hidden Markov Models for Bioinformatics，Kluwer Academic Publishers, Netherlands（2001）.
6）Imoto, S., Goto, T. and Miyano, S.：Estimating Genetic Networks and Functional Structures between Genes by Using Bayesian Networks and Nonparametric Regression, Pacific Symposium of Biocomputing 7 , pp.175-186（2002）.
7）玉田嘉紀，坂内英夫，井元清哉，片山俊明，宮野　悟:マイクロアレイデータと遺伝子に進化情報に基づく遺伝子ネットワーク推定，2004 年度統計関連学会報告集，pp.391-394（2004）.

（平成17年1月25日受付）