原発不明癌のマイクロアレイによる分類

Classification of cancer of unknown primary origin by microarray

倉橋　一成
Issei Kurahashi

指導教員：　大橋　靖雄　教授
Tutor: Prof. Y. Ohashi

東京大学大学院医学系研究科健康科学・看護学専攻　疫学・予防保健学分野
Department of Epidemiology and Preventive Health Sciences,
School of Health Sciences and Nursing, The University of Tokyo

全悪性腫瘍患者の3～5%を占めるといわれている原発不明癌は,悪性腫瘍が発見された時点で多くの転移巣がみられ，丁寧な臨床検査や経過観察を行っても原発巣を特定することが難しい．治療法も一般的に有効とされているものはまだ無いが，ある特徴的な部分集団に対しては予後良好な治療法がある事がわかってきた．しかし未だ存在する予後不良群に対する有効な治療法は無い．その予後不良群に対する治療選択の試みとして，本研究では原発不明癌のマイクロアレイによる有効な癌種分類を検討する．用いるデータは原発巣が特定している通常の癌腫データのみであり，原発不明癌に対する性能はクロスバリデーション（CV）によって擬似的に評価する．癌腫は全16種類, 1,024サンプルであり, Affymetrix社のHuman Genome U133 ArrayかまたはHuman Genome U133 Plus2.0 Arrayで測定されている．解析対象は22,277遺伝子とし, 遺伝子クラスタリングとpartial least squares (PLS)法によって10変数に縮小した．癌種のクラス選択は事後確率最大化法を修正したものを用い，結果の解釈の幅を広げるために情報エントロピー基準を提案した．事後確率最大化法を分類に用いたデータに対して単純に当てはめた結果は正解割合が98.2%であり，10-fold CV法によって評価した情報エントロピー基準の正解率は78.6%であった．この結果により，本研究での次元縮小とクラス選択によって癌種の分類が可能であることが確認できた．また情報エントロピー基準を使うことで結果の解釈に幅が広がり，治療選択を行う上での参考となることが示唆された．

Key words: cancer of unknown primary origin, classification, clustering, microarray, partial least squares

緒言

原発不明癌とは，悪性腫瘍が発見された時点で多数の転移があり，臨床的な諸検査や経過観察を行っても明確な原発部位の同定が困難な癌である．原発不明癌は全悪性腫瘍患者の 3～5%を占め，Median Survival Time（MST）はおよそ6～12ヶ月であり，剖検を行っても多くの症例では原発巣を同定できないと言われている．また抗癌剤の奏効率は20～30%と予後不良である[1]．原発不明癌の定義は様々ある[2]が，現在は肉腫や悪性リンパ腫，悪性黒色腫などは除外され，腺癌や未分化癌，扁平上皮癌などの上皮性腫瘍が対象となる事が多い．

原発不明癌の治療法を確立するために，これまで多くの臨床試験が行われてきた．まず1980年代から5-FU, サイクロフォスフォミドやドキソルビシンなどが試されたが，奏効率が20～30%，MSTは4～9ヶ月程度であった．その後，シスプラチンの開発に伴い臨床試験が行われたが，シスプラチンの有効性を証明できるものではなかった[3),4)]．1990年代にはタキサン，ゲムシタビン，塩酸イリノテカンなどの新しい薬剤が様々な癌腫に導入されたため，原発不明癌に対しても多くの第II相試験が行われた[5)-9)]．これらはこれまでの試験より良い成績を生み，奏効率が30～40%，MSTも12ヶ月を超えるものもあった．中でもプラチナ製剤（カルボプラチン）とタキサン（パクリタキセル）の併用療法の成績が最も有望であったので，現在ではカルボプラチン＋パクリタキセル併用療法が最も頻用されているレジメンであろう．しかし，このレジメンを標準的治療とするにはまだエビデンスが充分でないと思われる．

以上のように原発不明癌に対する確立した治療法は未だ無いと言えるが，治療法がある程度確立し予後良好と考えられている部分集団もある[10)-12)]．例えば腹膜播腫の認められる女性の腺癌は卵巣癌に準じた治療が行われ，頚部リンパ節腫大の存在している扁平上皮癌

は頭頚部癌に準じた治療が行われる．これら予後良好と思われる部分集団は原発不明癌には含めず，これらを除いて臨床試験を行う事が多い．

そのような状況で，マイクロアレイによる新しい癌種分類方法が様々提案されている事に伴って，原発不明癌をマイクロアレイによって分類しようというアプローチがある [13]．マイクロアレイでは何万もの遺伝子発現を同時に測定でき，1 つ 1 つの発現測定は誤差が大きいと思われるが，遺伝子全体の発現プロファイルを探索する事に長けている．様々な癌腫も，ある特定の遺伝子発現よりも全体の発現プロファイルに特徴があるとすれば，これはマイクロアレイによって分類が可能であると思われる．さらに原発不明癌にも原発巣があるとするならば，これもマイクロアレイによってある程度特定が出来，それによってその癌腫に適した治療法を選択する事で生存時間を延長できる可能性がある．以上の事を考慮して本研究では，既に測定されている癌腫発現データを用いて，将来の原発不明癌のアレイ発現データがどの癌種に近いかを推定するクラス選択方法を検討する事を目的とする．実際には原発不明癌のデータは利用できなかったが，クラス選択の性能はクロスバリデーション（CV）によって擬似的に評価する．また本研究で提案した手法は，平成 19 年度厚生労働省科学研究補助金・がん臨床研究事業「原発不明がんの診断・効果的治療の確立に関する研究班」が今後行う臨床試験で使われる予定である．

方法

1．対象

分類のために用いるデータは，公共データベースである Gene Expression Omnibus（GEO）[14]からダウンロードしたものと，近畿大学医学部ゲノム生物学教室から借用したものを用いる．これらのうち Affymetrix 社の Human Genome U133 Array（U133A）か，Human Genome U133 Plus2.0 Array（U133 2.0）で測定されているデータを用いる．測定されている遺伝子数は，それぞれのチップ種で 22,277 遺伝子と 54,675 遺伝子である．U133A で測定されている遺伝子は，全て U133 2.0 でも測定されているので，今回は両方のチップ種で測定されている 22,277 遺伝子を解析対象とした．用いる癌腫とサンプル数，実験施設数は，膀胱癌（49,2），悪性脳腫瘍（103,2），乳癌（81,1），子宮頸癌（61,2），大腸癌（53,4），直腸癌（177,6），胚細胞癌（101,1），頭頚部癌（42,1），肺腺癌（59,2），リンパ腫（5,1），卵巣癌（24,1），膵臓癌（39,1），前立腺癌（169,3），腎臓癌（9,1），胃癌（28,1），甲状腺癌（24,1），であり，合計 1,024 サンプル，16 癌腫，27 施設である． またこれら癌腫と測定されているチップ種の関係を表 1 に示す．本研究で使用する略名を癌腫名に括弧付けで示している．

2．観測値の補正

まず観測値の分散安定化を図るため，生データを逆双曲線正弦関数（asinh 関数）によって変数変換した [15]．

$$H_{gcljk} = \mathrm{asinh}(I_{gcljk} / \bar{I}_{\cdot cljk}) = \log\left\{ I_{gcljk} / \bar{I}_{\cdot cljk} + \sqrt{(I_{gcljk} / \bar{I}_{\cdot cljk})^2 + 1} \right\}$$

ここで I_{gcljk} は遺伝子 g，クラス（癌種）c，実験施設 l，測定チップ種 j，アレイ（サンプル）k の生の観測値であり，H_{gcljk} はこれを asinh 変換したものである．また $\bar{I}_{\cdot cljk}$ はアレイ内平均値である．さらに実験系の効果を削減するために以下の混合効果モデルによって実験施設の効果 b_l と測定チップ種の効果 c_j を推定し，変換値 H_{gcljk} から差し引いた．

$$H_{gcljk} = \alpha_g + \beta_c + b_l + c_j + \varepsilon_{gcljk},$$

$$E(H_{gcljk}) = \alpha_g + \beta_c,\ V(H_{gcljk}) = \sigma_b^2 + \sigma_c^2 + \sigma^2$$

ここで誤差 ε_{gcljk} には $N(0, \sigma^2)$ の正規分布を仮定している．また σ_b^2 は変量効果 b_l の分散，σ_c^2 は変量効果 c_j の分散である．上式は asinh 変換した観測値に，遺伝子効果 α_g と癌種の効果 β_c を固定効果とし，実験施設の効果 b_l と測定チップ種の効果 c_j を変量効果と仮定した混合効果モデルとなっている．以下からは変量効果を除いた $H_{gcljk} - \hat{b}_l - \hat{c}_j$ を用いて解析を行っていく．

3．次元縮小

分類を行うための変数が非常に多い場合は，何らかの方法によって変数を縮小する事によって，計算負荷を軽減することが多い．次元縮小の方法は，シグナルの強い変数のみを用いる方法と，主成分分析（PCA 法）や partial least squares 法（PLS 法）などのように変数の

表 1．癌腫と測定チップ

	U133A	U133 2.0	合計
膀胱癌(bld)	40	9	49
悪性脳腫瘍(brn)	103	0	103
乳癌(brs)	81	0	81
子宮頸癌(crv)	0	61	61
大腸癌(cln)	45	8	53
直腸癌(clr)	71	106	177
胚細胞腫(grm)	101	0	101
頭頚部癌(hed)	0	42	42
肺腺癌(lng)	59	0	59
リンパ腫(lym)	5	0	5
卵巣癌(ovr)	0	24	24
膵臓癌(pnc)	0	39	39
前立腺癌(prs)	150	19	169
腎臓癌(rnl)	9	0	9
胃癌(stm)	0	28	28
甲状腺癌(thy)	24	0	24
合計	688	336	1024
	67.2%	32.8%	

線形変換によって新しい変数を作る方法に大別できる．前者では通常t検定やF検定を行い，p値がある閾値より小さい変数を選び出す．この際，検定を変数の数だけ行うことになるので，検定の多重性を考慮してp値を補正するか，False Discovery Rateという概念によって変数を選び出す事が多い[16),17)]．

しかし今回は混合効果モデルによって実験施設の効果とチップ種の効果を取り除き，それによって癌種の効果を強めている．そのためF検定を行うとF値はどの遺伝子でも大きくなり，P値はほとんどの遺伝子で0.001を下回る．この状況で仮説検定を行うのは無意味と思われるので今回は検定を行うのではなく，PCA法やPLS法による次元縮小を行う事を試みた．しかし変数の数が極端に多いのでそのままPCA法やPLS法を行うことは出来ず，遺伝子を変数としたクラスター分析によって1段階目の次元縮小を行い，その変数を用いて更にPLS法によって2段階目の次元縮小を行った．この遺伝子クラスタリングを行う際，22,277遺伝子を一度に分析することは計算負荷の問題から出来ないので，遺伝子が存在する染色体毎に分けてそれぞれの中で行った．また手法としては重心法を用い[18)]，クラスターとなった遺伝子の平均値をそのクラスターの代表値としている．よって1)遺伝子クラスタリングによって22,277遺伝子を約1,000変数に縮小し，次に2)PLS法によって数変数～十数変数に縮小した．PLS法は結果変数と相関の高い合成変数を作る手法である．次元縮小の際に結果変数の情報を含めているため，４節に示すクラス選択の際に役立つ変数を作成する事ができると思われ，Daiらによって PCA法や sliced inverse regression法よりも性能が優れる事が示されている[19)]．以下，PLS法の説明を行う．

PLS法

観測値の数をN，結果変数と説明変数の数をそれぞれn，mとして，基準化説明変数行列を$\mathbf{X}_0$ $(N\times m)$，基準化結果変数行列を$\mathbf{Y}_0$ $(N\times n)$とする．$\mathbf{Y}_0$と相関の高い変数を$\mathbf{X}_0$から作るため，$\mathbf{X}_0$の線形変換$\mathbf{t}_1=\mathbf{X}_0\mathbf{w}_1$によって$\mathbf{X}_0$と$\mathbf{Y}_0$を以下のように予測する（$\mathbf{w}_1$は$m\times 1$ベクトル）．

$$\hat{\mathbf{X}}_0=\mathbf{t}_1\mathbf{p}_1^T \quad \text{where} \quad \mathbf{p}_1^T=(\mathbf{t}_1\mathbf{t}_1^T)^{-1}\mathbf{t}_1^T\mathbf{X}_0$$

$$\hat{\mathbf{Y}}_0=\mathbf{t}_1\mathbf{c}_1^T \quad \text{where} \quad \mathbf{c}_1^T=(\mathbf{t}_1\mathbf{t}_1^T)^{-1}\mathbf{t}_1^T\mathbf{Y}_0$$

このとき$\mathbf{t}_1$は$\mathbf{u}_1=\mathbf{Y}_0\mathbf{q}_1$との共分散$\mathbf{t}_1^T\mathbf{u}_1$が最大になるように推定し，$\mathbf{w}_1$と$\mathbf{q}_1$はそれぞれ$\mathbf{X}_0^T\mathbf{Y}_0\mathbf{Y}_0^T\mathbf{X}_0$と$\mathbf{Y}_0^T\mathbf{X}_0\mathbf{X}_0^T\mathbf{Y}_0$の第一固有ベクトルになっている（$\mathbf{q}_1$は$n\times 1$ベクトル）．次に$\mathbf{X}_0-\hat{\mathbf{X}}_0$と$\mathbf{Y}_0-\hat{\mathbf{Y}}_0$を用いて同様の計算を行うことを繰り返していき，合成変数であるPLSスコア$\mathbf{t}_i$を順次作っていく．この時，新しく作ることが出来る変数の最大数は$\mathbf{X}_0$の階数に等しい．しかしあまりに多くの変数を作るとover-fittingが起こるため，予測精度は低下する．そのためCVによって，予測誤差が最も小さくなる変数の数を求めるのが一般的である．本研究では，使用データへ当てはめた際の正解割合が95%を超え，さらに４節で述べる情報エントロピー基準での正解率が最も高くなるように変数の数を定めた．また今回の結果変数はクラス（癌種）であり文字変数であるため，クラスを示す2値ダミー変数を16個用意した．

４．クラス（癌種）選択

３節の方法によって縮小された変数を用いて，観測値の分類を行う．異なる群からの観測値を分類するための方法は，クラスタリングやパターン認識の分野で様々な方法が研究されている．古典的な手法としては線形判別分析や，1-nearest neighbor法（1-NN），k-nearest neighbor法（k-NN）などがあり，最近注目を浴びている手法としてはSupport Vector Machine（SVM）やニューラルネットワーク，ベイジアンネットワークなどがある[20)]．後者の手法はマイクロアレイの分野でも多くの適用が成され，適用したデータセットに対する性能の良さは認められている[13)]．しかしこれらの方法は計算アルゴリズムが複雑であり，判別面も線形にならず複雑な非線形面になる場合もある．今回は，判別のために適切な変数選択が出来れば，従来の古典的な方法でも充分に判別が可能であるという立場[17)]で手法を提案する．

1）事後確率最大化法

まず一般的なクラス選択の方法として知られているベイズの定理を用いた事後確率最大化法について説明する．これは観測ベクトル$\mathbf{x}$に対する各クラスcの事後確率$p(c|\mathbf{x})$が最大となるクラスに分類するというものである．各クラスに属する観測値が多変量正規分布に従っていると仮定すると，$p(c|\mathbf{x})$は次のように表せる．

$$p(c|\mathbf{x})=p(\mathbf{x}|c)p(c)\Big/\sum_c p(\mathbf{x}|c)p(c),$$

$$p(\mathbf{x}|c)p(c)=\exp(-0.5D_{\mathbf{x}c}-0.5\ln|\mathbf{S}_c| -0.5M\ln 2\pi+\ln p(c))$$

$$D_{\mathbf{x}c}=(\mathbf{x}-\mu_c)^T\mathbf{S}_c^{-1}(\mathbf{x}-\mu_c),\quad \mathbf{x}=(x_1,...,x_M)^T$$

ここで$p(c)$は各クラスの事前確率であり，推定値としては経験ベイズ流に各クラスのサンプル数割合を用いる．$\mathbf{S}_c$は各クラスの分散共分散行列であり，各クラスで相関構造が異なることを仮定している．$D_{\mathbf{x}c}$はクラスcの平均値ベクトルμ_cからのマハラノビス距離を2乗したものである．またMは観測ベクトル$\mathbf{x}$の次元

である.

2) 観測値の情報量を考慮したクラス選択

事後確率最大化法で用いる事後確率 $p(c|\mathbf{x})$ に関する問題点は，$p(\mathbf{x}|c)p(c)$ が距離 $D_{\mathbf{x}c}$ に関して指数関数的に減少するため，正規分布の仮定が成立しない場合や，相関構造の推定に失敗した場合は $D_{\mathbf{x}c}$ が真のクラスではないクラスに引きずられてしまい，その結果あるクラスの事後確率が極端に高くなってしまう点である．その様な状況への対処法として，正規分布ではなく経験分布を用いることが考えられるが，経験分布を用いると将来の観測値を予測する際の計算量が大きくなってしまう．そこで，正規分布による計算の簡便さを残したまま事後確率を適度なものにするため，次のような計算を行うことを提案する.

$$p'(c|\mathbf{x}) = p'(\mathbf{x}|c)p(c) \Big/ \sum_c p'(\mathbf{x}|c)p(c),$$

$$p'(\mathbf{x}|c)p(c) = \exp\{(-0.5D_{\mathbf{x}c} - 0.5\ln|\mathbf{S}_c| - 0.5M\ln 2\pi + \ln p(c))/M\}$$

ここで $p'(\mathbf{x}|c)p(c)$ は $p(\mathbf{x}|c)p(c)$ の M 乗根となっており，多変量正規分布の確率密度と事前確率を $1/M$ 乗している事に相当する. この補正を加える事で，幾何平均のような頑健性を持ち，観測ベクトル $\mathbf{x}$ の次元が異なっていても事後確率が比較可能となる．さらに観測ベクトル $\mathbf{x}$ の情報量を把握するために，事後確率 $p'(c|\mathbf{x})$ を利用して情報エントロピーを求める [21].

$$E_{\mathbf{x}} = -\sum_c p'(c|\mathbf{x})\log_2 p'(c|\mathbf{x})$$

$p'(c|\mathbf{x}) = 0$ のときは $p'(c|\mathbf{x})\log_2 p'(c|\mathbf{x}) = 0$ である．Fussy-clustering の分野では，事後確率から計算される情報エントロピーの総和をデータ全体の情報量と捉えて，各クラスの最適中心ベクトルを求める際に利用する場合もある [22]．しかし今回は情報エントロピーを個々の観測値が持つ情報量と解釈し，情報量が少ない場合，つまり情報エントロピーが高い場合は次のような基準によって，事後確率 $p'(c|\mathbf{x})$ の高いクラスから順に複数クラスを候補クラスとする.

$$\text{if } \log_2 i_{\mathbf{x}} \le E_{\mathbf{x}} < \log_2(i_{\mathbf{x}}+1) \text{ then}$$
$$\text{candidates } i_{\mathbf{x}} \text{ classes}$$

各 $i_{\mathbf{x}}$ は 1 から分類クラスの総数までの値をとる整数であり，情報エントロピーの値に応じて決まる．$p(c|\mathbf{x})$ の最大化と $p'(c|\mathbf{x})$ の最大化は同値であるため，補正を行った事後確率が最大になるクラスは事後確率最大化法での分類クラスと同じであるが，情報エントロピーは $p(c|\mathbf{x})$ と $p'(c|\mathbf{x})$ のどちらを元にして計算するかで大きく異なる．この候補クラスに真のクラスが含まれていれば正解とみなし，これを情報エントロピー基準と呼ぶ．情報エントロピー値と候補クラス数との関係は，情報エントロピーが 1 未満で 1 クラス，1 以上 1.58 未満で 2 クラス，1.58 以上 2 未満で 3 クラス，2 以上 2.32 未満で 4 クラス，2.32 以上 2.58 未満で 5 クラス，2.58 以上 2.81 未満で 6 クラス，2.81 以上 3 未満で 7 クラスとなっている. この基準によって将来の観測値を予測する際の解釈に幅が広がる．この情報エントロピー基準による候補クラス選択がどのように機能するかを，５節の CV によって確認する.

５．クロスバリデーション（CV）

クラス選択方法の誤分類率を推定するために CV を行う．方法の 1 つである k-fold CV は，データセットを k 個に分け，どれか 1 つを test set とし残りの k-1 つを training set として誤分類割合を計算する事を繰り返し，k 回の平均値を誤分類率の推定値とする方法である. 誤分類率の推定方法には他にも bootstrap 法，.632 bootstrap 法，.632+ bootstrap 法などがある．これらの手法の性能を Efron らがシミュレーションによって比較検討しており, k-fold CV と比べて.632+ bootstrap 法が最も良いとされている [23),24].

しかし今回扱うデータは 1,024 観測値×22,277 変数という大きなデータであり，しかも変数選択も含めて CV を行わないと誤分類率を過小評価してしまう [25]．つまり次元縮小を行うために行った遺伝子クラスタリングと，その変数に基づく PLS 法もその都度繰り返し行わなくてはならない．そのため leave-one out CV 法や bootstrap 法などのように多くの計算回数を必要とする手法は向いていない．そのため今回は誤分類率を推定するために，1)10-fold CV を行った．この際，癌種によって偏りが出ないようにデータを 10 分割した．また新しい観測値が，training set に存在しないクラスに属する観測値である場合の情報エントロピーの挙動を確認したい．そのため，2)あるクラス（癌種）に属する観測値を全て test set とし，それ以外のクラスの観測値を training set とする CV の 2 通り行う．これ以降，2)の方法を leave-type-out CV と呼ぶ． 1)の方法は，将来予測する原発不明癌患者に原発巣が存在する状況に対応している．また 2)の方法は，原発巣が手持ちのデータセットに無い場合や，そもそも原発巣が存在しない場合に対応している.

また今回は，新しい観測値の分類は事前に適切な方法によって実験効果は取り除かれている事を前提として CV を行う．つまり混合効果モデルによる観測値の補正は CV には含めず，次元縮小（遺伝子クラスタリング，PLS 法）とクラス選択のみを CV 計算の対象とする.

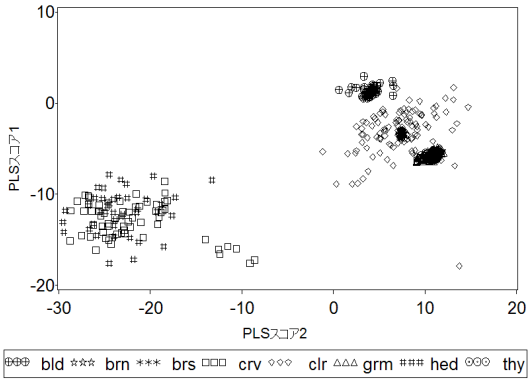

図1. PLS スコア 1, 2 と癌腫の関係 1

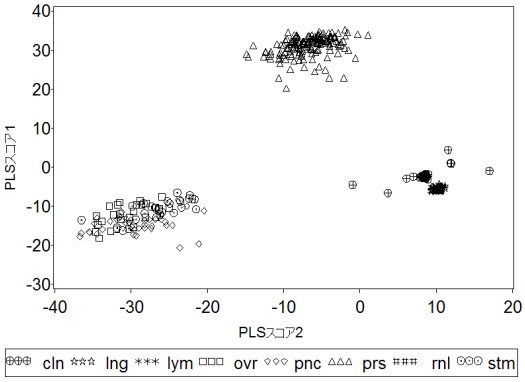

図2. PLS スコア 1, 2 と癌腫の関係 2

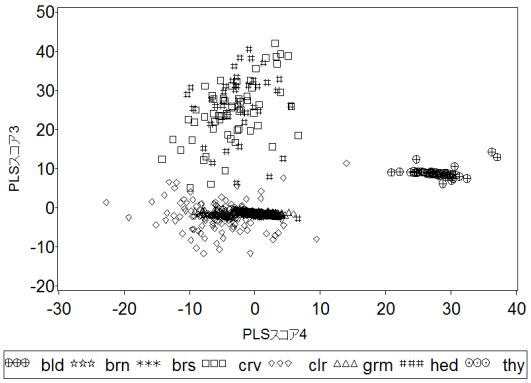

図3. PLS スコア 3, 4 と癌腫の関係 1

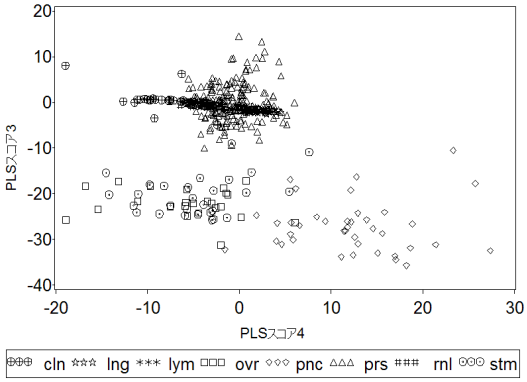

図4. PLS スコア 3, 4 と癌腫の関係 2

表2. クラスタリングの結果

	正解	不正解	合計
膀胱癌(bld)	49	0	49
悪性脳腫瘍(brn)	102	1(grm)	103
乳癌(brs)	75	6(grm)	81
子宮頸癌(crv)	59	2(hed)	61
大腸癌(cln)	53	0	53
直腸癌(clr)	177	0	177
胚細胞腫(grm)	93	8(brn,brs,lng,thy)	101
頭頚部癌(hed)	41	1(crv)	42
肺腺癌(lng)	59	0	59
リンパ腫(lym)	5	0	5
卵巣癌(ovr)	24	0	24
膵臓癌(pnc)	39	0	39
前立腺癌(prs)	169	0	169
腎臓癌(rnl)	9	0	9
胃癌(stm)	28	0	28
甲状腺癌(thy)	24	0	24
合計	1006	18	1024
	98.2%	1.8%	

表3. 10-fold CV 法の誤分類率

	正解	不正解	合計
膀胱癌(bld)	37	12	49
悪性脳腫瘍(brn)	96	7	103
乳癌(brs)	21	60	81
子宮頸癌(crv)	57	4	61
大腸癌(cln)	37	16	53
直腸癌(clr)	166	11	177
胚細胞腫(grm)	94	7	101
頭頚部癌(hed)	30	12	42
肺腺癌(lng)	45	14	59
リンパ腫(lym)	0	5	5
卵巣癌(ovr)	19	5	24
膵臓癌(pnc)	38	1	39
前立腺癌(prs)	141	28	169
腎臓癌(rnl)	0	9	9
胃癌(stm)	24	4	28
甲状腺癌(thy)	0	24	24
合計	805	219	1024
	78.6%	21.4%	

表4. 真のクラスの事後確率の順位

順位	頻度	割合	累積割合
1	449	43.85	43.85
2	187	18.26	62.11
3	79	7.71	69.82
4	78	7.62	77.44
5	71	6.93	84.38
6	71	6.93	91.31
7	39	3.81	95.12
8	15	1.46	96.58
9	1	0.10	96.68
10	3	0.29	96.97
13	5	0.49	97.46
14	12	1.17	98.63
15	5	0.49	99.12
16	9	0.88	100.00

結果

1．使用データへの fitting の評価

まずPLS法によって得られたPLSスコアを癌腫毎にプロットしたものを図1～図4に示す. 方法の３節で述べた基準に沿うと，変数の数は 10 個が最適であった. これらの図により，PLS スコアによって癌腫が分かれていることが分かる. 特に図2で前立腺癌 (prs) の PLS スコア 1 が高くなっており，他の癌腫とは異なるクラスターを示している. さらに膀胱癌(bld)や直腸癌(clr)，大腸癌（cln）なども 1 つのクラスターを成している. また卵巣癌（ovr）と胃癌（stm），膵臓癌（pnc），頭頚部癌（hed）はまとまって存在している．他の癌腫もそこまではっきりとしたクラスターを成しているわけではないものの，ある程度のまとまりは確認できる．また今回示していないPLS スコア 5～10 でも癌腫のまとまりは確認できる．実際のクラスタリングはこの 10 変数を用いるので, 図1～図4で確認できる以上にはっきりとクラスターが形成されているものと思われる．次にこのPLS スコア 1～10 を用いてクラスタリングを行った結果を表 2 に示す．全体の 1.8%である 18 サンプルが誤分類されていた．不正解の列にある括弧は誤分類された癌腫を表している．ただしこの結果は方法の４節で述べた事後確率最大化法の結果である. また図5に癌種毎の情報エントロピーの箱ひげ図を示す. x 軸は癌腫の略名であり順序は表 1 と同じである．癌種によって情報エントロピーの分布が大きく異なっているこ

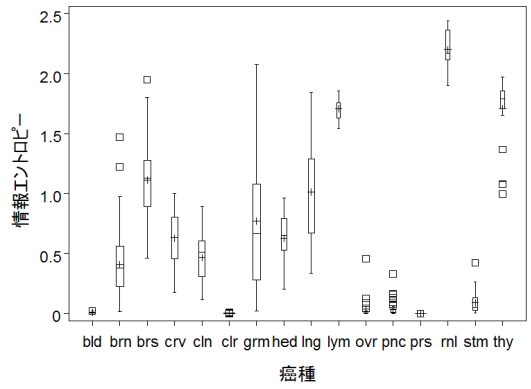

図5. 癌腫毎のエントロピーの箱ひげ図

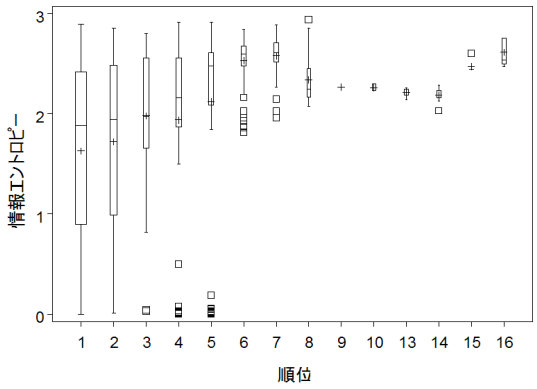

図6. エントロピーの箱ひげ図
(x 軸は真のクラスの事後確率の順位)

とが分かり，もともとサンプル数が少ない腎臓癌（rnl）とリンパ腫（lym）の情報エントロピーが特に高い．また情報エントロピーが 1 より大きい観測値は 149 サンプルあり，情報エントロピー基準を適用するとこれらのサンプルは候補クラスが 2 クラス以上と判定されることになり，誤分類割合は 0.0%となる．

2．クロスバリデーション（CV）

1）10-fold CV 法の結果

次に CV の結果を示す．まず 10-fold CV 法で情報エントロピー基準の誤分類率を推定した結果を表 3 に示す．次に表 4 に，真のクラス（癌種）の事後確率順位を示す．真のクラスの事後確率が他のクラスの事後確率に比べて最も高かった観測値は 449 個（43.9%）であったので，事後確率最大化法では残りの 575 個 (56.1%) が誤分類となってしまう．また図 6 に，各観測値の情報エントロピー分布をこの順位毎に比較した箱ひげ図を示す．真のクラスの事後確率の順位が下がるほど情報エントロピーが平均的に高くなっていることが分かる．図6で情報エントロピー基準を当てはめてみると，順位が 1 の観測値では情報エントロピーが 1 以上の観測値が約 3/4 個であるので，これらの観測値では候補クラスが 2 クラス以上と判断される．これは無駄のように思われるが，順位が 2 の観測値では約 3/4 個以上の観測値で情報エントロピーが 1 以上である．これらの観測値は候補クラスが 2 個以上となるので，候補クラスに真のクラスが含まれ正解となる．真のクラス順位が 3 位の観測値も約 3/4 個の観測値で情報エントロピーが 1.58 以上であるため，候補クラスに真のクラスが含まれ正解となる．このように情報エントロピー基準を用いて候補クラスを複数個用意することによって，正解率を 78.6% (805 個) まで上げる事が可能となっている．

2）leave-type out CV 法の結果

図 7 に leave-type out CV 法の結果を示す．図 5 と比べると図 7 では全体的に情報エントロピーが高くなっている事がわかる．情報エントロピーが 1 を超えている

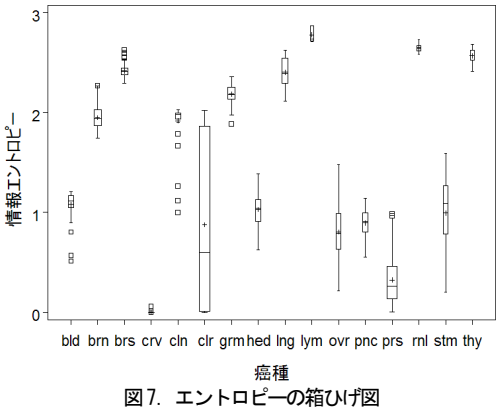

図7. エントロピーの箱ひげ図
(leave-type out CV 法の結果)

観測値は 608 個であり，図 5 の際の 149 個と比べて 4 倍以上になっている．特に直腸癌（clr）で情報エントロピーの大幅な上昇が見て取れる．これは直腸癌を他の癌種に無理に分類しようとすると曖昧な分類しかできないことを意味している．これにより，手持ちのデータセットに真のクラスが存在しない場合は情報エントロピーが高くなることが分かる．

考察

1．データ

使用したデータは，U133A と U133 2.0 という 2 種類のチップ種によって測定されており，測定された実験施設も多種多様であった．この実験施設の効果とチップ種の効果を取り除くために，今回は混合効果モデルを仮定して変量効果として推定した．変量効果を仮定した理由の 1 つは，今回の実験施設やチップ種は，無限にある水準から得られた標本だと考えられるからである．特に実験施設は世界中に多数存在し，利用した実験施設はそれら母集団からの標本である．さらに今回興味のある効果は，実験施設やチップ種に影響されない癌種や遺伝子の効果であったので，混合効果モデルで取り除く事が妥当であったと考えられる．しかし，今回のモデルは asinh 変換後に線形混合効果モデルを仮定したので，実験施設やチップの真の効果が asinh 変

換後のスケールで加法的でない場合は適切に取り除く事が必ずしもできていない可能性もある．

また対象とした癌種については，原発不明癌のうち原発巣が特定できた症例を参考にしつつ，原発不明癌の原発巣として一般的に検討されているものに絞った．特に乳癌に関しては，公共データベース上に多くのデータがあったため，リンパ節に転移があったものに絞った．また肺癌も，腺癌と明記されてあったものに限定した．これによって，より現実の原発不明癌を分類する状況に近づいていると思われる．ただ，他の癌腫は実験データ数が少なかったためそのような選択を行うことは出来なかった．本来であれば，他の癌腫でも転移が見られるものやstageが進行している癌腫に限定することで，より原発不明癌に近づけることが出来たかもしれない．

２．次元縮小

本研究では次元縮小を遺伝子クラスタリング，PLS 法という順序で行った．遺伝子クラスタリングを行う代わりに良く行われる方法は，検定統計量が大きいものから順に変数を選び出す方法である．しかしこの方法を適用すると，実際に使われる変数はほんのわずかしかない．例えば 22,277 遺伝子のうち上位 1,000 遺伝子を使うとすると，約 5%の遺伝子しか利用していないことになる．これは検定統計量が小さい遺伝子はほとんど情報を持っていないという考え方によるものと思われる．しかし，1 つ 1 つの遺伝子が情報を持っていない場合でも，それらの遺伝子の組み合わせによって何らかのパターンが現れてくるという考え方もある．本研究では後者の立場をとり，遺伝子全体の情報を使うために，遺伝子クラスタリングを行って変数の縮小を行った．またこの際に，染色体情報を用いて遺伝子クラスタリングを行ったことで，ある程度生物学的な情報を加味できたと言える．

次に行った PLS 法であるが，癌種毎にクラスターは構成されているもののチップ種の効果が含まれていることは否定できない．子宮頸癌，頭頚部癌，卵巣癌，膵臓癌，胃癌は全てのサンプルが U133 2.0 で測定されているが，図 1 と図 2 ではこれらの癌腫は近いところでばらついていることからこのことが類推される．しかし膀胱癌，大腸癌，直腸癌，前立腺癌も一部のサンプルが U133 2.0 で測定されていることを考えると，チップ種の効果よりも癌種の効果の方が強く出ているとも思われる．もしくは，PLS 法に限らず行列の特異値分解によって次元縮小する手法では 1 つ目の変数に最も多くの情報が含まれるので，PLS スコア 1 がチップ種の効果も表現しており，PLS スコア 2 以降が癌種の効果を表現しているのかもしれない．実際に，PLS スコア 3 を見ると，同じチップ種で測定されている卵巣癌，胃癌と頭頚部癌，子宮頸癌が異なる位置に分布していることが分かる．さらに今回は PLS 法によって 10 変数と比較的少ない変数に縮小した．この理由は，変数をあまり増やすと得られた変数が training set に強く引きずられてしまい，test set の予測がうまく行かなくなってしまうからである．実際に変数を 10 より増やして 10-fold CV を行うと，情報エントロピー基準での誤分類率は増える傾向にあった．また反対に変数を減らすと，今度は事後確率最大化法での誤分類率が増える傾向であった．今回の 10 変数は，情報エントロピー基準と事後確率最大化法との両方のバランスがとれていると言える．

３．クラス選択，クロスバリデーション（CV）

まず遺伝子クラスタリングと PLS 法によって縮小した 10 変数によって癌種の分類を行うとほぼ 100%の正解割合であり，この次元縮小は癌種の違いを充分に捉えていると思われる．次に CV によって将来の観測値に対する誤分類率を確かめたが，通常の事後確率最大化法で行うと 56.1%である（表 4 参照）．fitting を確かめた際の 1.8%と比べると非常に高いが，これは多くの癌種を同時に扱っているためだと思われる．2 種や 3 種の癌腫に絞ってクラス選択を行えば誤分類率はもっと低くなると思われる．しかし本研究の対象は原発不明癌であるため，分類器としてはなるべく多くの癌種を用意しておく必要があった．

本研究ではクラス選択の際に，各観測値が持っている情報量を定量的に測るために情報エントロピーを用いることを提案した．通常のクラス選択では，観測値を事後確率が最大のクラスに分類する．しかしその観測値がクラスに属する確からしさを表している事後確率の大きさ自体を考慮することで，クラス選択の幅が広がると考えた．実際に Fussy-clustering の分野では事後確率の大きさを考慮した分類と解釈を行っている [26]．しかし単純に多変量正規分布を仮定した事後確率を用いると，中心からの距離が少し異なるだけでも事後確率は大幅に変化してしまう．その結果，誤分類が起こっている観測値でも情報エントロピーが非常に低くなる．そのため距離に相当する部分を M で割る（多変量正規分布の確率密度を $1/M$ 乗する）事で，事後確率が 1 つのクラスに引きずられることが無くなり適度なものとなるため，クラス平均から離れている観測値の情報エントロピーが高くなる．今回結果は載せていないが，確率密度に対して補正を加えないと情報エントロピーはどの観測値もほとんど 0 に近くなってしまい，

情報量の差を検出する事は不可能であった．

この情報エントロピー基準を用いる事で，誤分類率を小さくする事が出来ている．この方法は事後確率最大化法と比べて候補クラスの数が増えているので，見かけ上誤分類率が小さくなることは自明である．さらに事後確率最大化法の場合では正解していたサンプルでも，情報エントロピー基準を用いる事で候補クラスが複数発生してしまう場合もある．しかしながら CV の結果から，情報エントロピー基準を適用した場合の候補クラスに真のクラスが存在する確率が 78.6%であると解釈でき，これは事後確率最大化法の正解率である 43.9%よりもかなり高い．実際に推定癌腫を考慮して治療選択を行う場合は，候補クラスに含まれている癌腫を吟味して，誤分類率と治療選択の risk-benefit の trade-off を考慮して決定をすることになると思われる．

結論

本研究による次元縮小，クラス選択によって適切な癌種の分類が可能となることが確認できた．また本研究で提案した情報エントロピー基準を用いることで，治療選択の際の参考になる事が示唆された．

謝辞

本研究を行うにあたって必要不可欠であったアレイデータをご提供下さった近畿大学ゲノム生物学教室の西尾和人教授，荒尾徳三講師，がん臨床試験についてご指導下さいました近畿大学医学部腫瘍内科の中川和彦教授に深く感謝申し上げます．また多くのご指摘，ご指導を下さいました東京大学医学系研究科生物統計学/疫学・予防保健学の大橋靖雄教授，松山裕准教授，伊藤陽一助教，飯室聡特任助教，原田亜紀子特任助教および教室関係者に御礼申し上げます．

文献

1) Pavlidis N, Briasoulis E, Hainsworth J, Greco FA. Diagnostic and therapeutic management of cancer of an unknown primary. *Eur J Cancer* 2003;39:1990-2005.

2) Briasoulis E, Pavlidis N. Cancer of unknown primary origin. *Oncologist* 1997;2:142-52.

3) Woods RL, Fox RM, Tattersall MH, Levi JA, Brodie GN. Metastatic adenocarcinoma of unknown primary site: a randomized study of two combination-chemotherapy regimens. *N Engl J Med* 1980;303:87-9.

4) Goldberg RM., Smith FP, Ueno W, Ahlgren JD, Schein PS. 5-fluorouracil, adriamycin, and mitomycin in the treatment of adenocarcinoma of unknown primary. *J Clin Oncol* 1986;4:395-9.

5) Hainsworth JD, Erland JB, Kalmn CA, Schreeder MT, Greco FA. Carcinoma of unknown primary cite: treatment with one-hour paclitaxel, carboplatin and extended schedule etoposide. *J Clin Oncol* 1997;15:2385-93.

6) Greco FA, Erland JB, Morrissey LH, Burris III HA, Hermann RC, Steis R, et al. Carcinoma of unknown primary site: Phase II trials with decetaxel plus cisplatin or carboplatin. *Ann Oncol* 2000;11:211-5.

7) Greo FA, Burris III HA, Litchy S, Barton JH, Bradof JE, Richards P, et al. Gemcitabine, carboplatin, and paclitaxel for tatients with carcinoma of unknown primary site: A Minnie Pearl Cancer Research Network Study. *J Clin Oncol* 2002;20:1651-6.

8) Culine S, Lortholary A, Voigt J-J, Bugat R, Theodore C, Priou F, et al. Cisplatin in combination with either gemcitabine or irinotecan in carcinomas of unknown primary site: resalts of a randomized phase II study-trial for the French Study Group on Carcinomas of Unknown Primary (GEFCAPI 01). *J Clin Oncol* 2003;21:3479-82.

9) Briasoulis E, Kalofonos H, Bafaloukos D, Samantas E, Fountzilas G, Xiros N, et al. Carboplatin plus paclitaxel in unknown primary carcinoma: a phase II Hellenic Cooperative Oncology Group study. *J Clin Oncol* 2000;18:3101-7.

10) Hainsworth JD, Greco FA. Treatment of patients with cancer if an unknown primary site. *N Engl J Med* 1993;329:257-63.

11) Van der Gaast, Verweij J, Planting AS, Hop WC, Stoter G. Simple prognostic model to predict survival in patients with undifferentiated carcinoma of unknown primary site. *J Clin Oncol* 1995:13:1720-5.

12) Culine S, Kramar A, Saghatchian M, Bugat R, Lesimple T, Lortholary A, et al. Development and validation of a prognostic model to predict the length of survival in patients with carcinomas of an unknown primary site. *J Clin Oncol* 2002;20;4679-83.

13) Tothill RW, Kowalczyk A, Rischin D, Bousioutas A, Haviv I, van Laar RK, et al. An expression-based site of origin diagnostic method designed for clinical application to cancer of unknown origin. *Cancer Res* 2005;65:4031-40.

14) GEO. Available at: http://www.ncbi.nlm.nih.gov/geo/ Accessed January 12. 2008.

15) 倉橋 一成, 伊藤 陽一, 松山 裕, 大橋 靖雄, 西尾 和人. cDNA マクロアレイデータ解析における正規化手法の性能評価. 日本統計学会誌 2007;36:147-63.

16) Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc [Ser B]* 1995;57: 289-300.

17) Simon RM., Korn EL, McShane LM, Radmacher MD, Wright GW, Zhao Y. *Design and analysis of DNA microarray investigations.* New York: Springer;2003.

18) Anderberg MR. *Cluster analysis for applications*. New York: Americ Press;1973.

19) Dai JJ, Lieu L, Rocke D. Dimension reduction for classification with gene expression microarray data. *Stat Appl Genet Mol Biol* 2006;5:1,article6.

20) 甘利 俊一, 麻生 英樹, 津田 宏治, 村田 昇. パターン認識と学習の統計学. 岩波書店;2003.

21) Shannon CE. A mathematical theory of communication. *The Bell System Technical Journal* 1948;27: 379-423, 623-56.

22) Li RP, Mukaidono M. A maximum-entropy approach to fuzzy clustering. *in Proc 4th IEEE Int Conf Fuzzy Syst* 1995;4:2227-32.

23) Efron B, Tibshirani R. Improvements on cross-validation: the .632+ bootstrap method. *J Am Stat Assoc* 1997;92:548-60.

24) Efron B. Estimating the error rate of a prediction rule: improvement on cross-validation. *J Am Stat Assoc* 1983;78:316-31.

25) Simon R, Radmacher MD, Dobbin K, McShane LM. Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *J Natl Cancer Inst* 2003;95:14-8.

26) Kandel A. *Fuzzy techniques in pattern recognition.* New York:John Wiley and Sons;1982.