様式 C-19

科学研究費補助金研究成果報告書

平成２１年５月２２日現在

研究種目：基盤研究（C）
研究期間：2006～2008
課題番号：１８５００１３２
研究課題名（和文）標的蛋白質を用いたパターン認識による癌診断システムの開発

研究課題名（英文）Development of a cancer diagnosis system by pattern recognition techniques with use of target proteins

研究代表者
浜本　義彦（HAMAMOTO YOSHIHIKO）
山口大学・大学院医学系研究科・教授
研究者番号：９０１９８８２０

研究成果の概要：
現在、社会問題となっているＣ型肝臓癌の診断を目的として、肝臓癌診断に有用な蛋白質を高精度で標的蛋白質として同定し、標的蛋白質群を用いて肝臓癌の診断を８５％の精度で達成した。この研究成果を更に発展させることにより、患者一人ひとりに合った治療を行うテーラーメイド医療を実現することができ、更に無駄な医療費を大幅に削減することも期待できる。

交付額

（金額単位：円）

	直接経費	間接経費	合　計
2006 年度	1，800，000	0	1，800，000
2007 年度	800，000	240，000	1，040，000
2008 年度	900，000	270，000	1，170，000
年度			
年度			
総　計	3，500，000	510，000	4，010，000

研究分野：情報工学
科研費の分科・細目：情報学・知覚情報処理・知能ロボテックス
キーワード：パターン認識，癌，プロテオーム，診断

１．研究開始当初の背景

分子生命科学の急速な発展を受けて、癌研究の分野でマイクロアレイを用いた遺伝子発現解析に関する数多くの研究がなされている。しかし、遺伝子発現情報はメッセンジャーRNA量を定量化したもので、最終的な生産物である蛋白質が生命現象の大部分を担っていること、また薬剤の直接の標的分子が蛋白質であることからも、その限界が指摘されている。そのため、近年、蛋白質そのものを対象とするプロテオーム解析が注目されている。実際、プロテオーム解析は、第３期科学技術基本計画（中間報告）でもテーラーメイド医療実現の基盤技術として、改めて重要視されている。

プロテオーム解析では、一般に、２次元電気泳動により生体組織内の蛋白質をゲル画像上でスポットとして単離し、最終的には質量分析とデータベース検索により注目スポットが如何なる蛋白質であるかを調べることになる。そのため、癌医療に有用なスポット、すなわち癌に深く関与する蛋白質を見出すことが重要となる。しかしながら、マイクロアレイによる遺伝子発現解析に比べ、現状の蛋白質発現解析は国内外とも患者単位で

個々の蛋白質を対象とする個別レベルの解析であり、有用蛋白質の探索において網羅性が著しく欠けている。更に蛋白質の検出感度にも問題があり、一般に大量発現している蛋白質が検出され易く、微量ながらも重要な蛋白質のほとんどが見逃されている。

このようにプロテオーム解析には多大な期待がかけられているものの、臨床応用とは大きなギャップがあるのが現状である。

２．研究の目的

本研究では、これまで遺伝子発現解析で実績をあげている肝臓癌を対象とし、トランスレーショナルリサーチ（基礎から臨床への橋渡し研究）としての観点から、パターン認識技術による臨床応用可能な肝臓癌診断システムの開発を目的とする。このためには、標的となる癌診断に有用な蛋白質の同定とそれに基づく癌診断システムの構築が課題となる。

そこで、画像解析により２次元電気泳動からのゲル画像を癌部と非癌部で対比させて癌診断に有用なスポット（蛋白質）を検出し、パターン認識技術により候補となる蛋白質の中で診断に有用な標的蛋白質群を特徴とみなして肝臓癌診断システムを構築する。

３．研究の方法

（１）有用蛋白質の同定

①スポット検出

スポット検出を高速化するためには、まずラスタースキャンによりゲル画像から濃度の極大点となるスポット候補点を探す。次に、スポット候補点を中心とする探索領域を設定し、その探索領域内で濃度が高い領域をスポットとする。この際、探索を容易にするため、濃淡強調を行う。この濃淡強調を行った探索領域に対して、ガウス型関数のフィティングを行う。このときゲル画像の歪みを考慮して、当該探索領域を０度から１３５度まで４５度刻みで回転させ、ガウス型関数フィティングを行う。次に、スポットとノイズを分離する。これには歪度を用い、ガウス型関数の分散と歪度が、予め定めたしきい値内であれば、スポットとして検出する。

②特異的スポットの検出

片方のゲル画像にしか存在しない特異的スポットを検出するために、局所的アフィン変換を用いたスポット双方向対応付けを行う。今、ゲル画像Ｒとゲル画像Ｓがあり、それらのスポットを対応付ける問題を考える。スポット双方向対応付けとは、まず、局所的アフィン変換を用いてゲル画像 Ｒ のスポット群の位置を修正してゲル画像Ｓのスポット群に対応づける。このとき、ゲル画像Ｒの複数のスポットの中には、ゲル画像Ｓの特定のスポットに重複して対応付けられるものがある。そこで、逆に、ゲル画像Ｓからゲル画像Ｒへ、局所的アフィン変換を用いてスポット群の対応付けを行う。ここで、両方の対応付けで一致したスポット対を、ゲル画像間で対応付けられたスポットとする。

次に、スポット群から対応付けられたスポット対を削除し、残った、つまりゲル画像間で対応付けられず、片方のみのゲル画像に存在するスポットを、特異的スポットとして検出する。特異的スポットには、癌特異的なスポットもあれば、正常特異的なスポットもある。これらの特異的スポットが、肝臓癌を診断する上で有用な候補蛋白質として考えられる。

（２）肝臓癌診断システムの設計

本研究では、教師あり学習におけるパターン認識に基づいた標的蛋白質の同定及び識別器の学習により肝臓癌診断システムを構築する。図１にその枠組みを示す。教師あり学習は、正解となるクラスラベルが付記されたサンプルを用いる学習である。ここでのクラスは、肝臓癌のクラスと非癌（正常）のクラスを意味する。臨床医によって与えられたクラスラベル付きのサンプルを利用できるサンプル集合とする。このサンプル集合から、訓練サンプルとテストサンプルを取り出し、まず訓練サンプルを用いて標的蛋白質の同定と識別器の設計を行って診断システムを構築し、次にテストサンプルを用いて診断システムの診断性能を評価する。

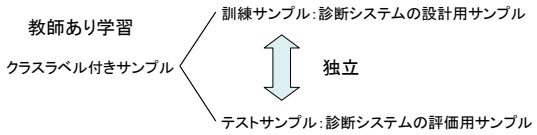

図1　教師あり学習の枠組み

図２は、診断システムの設計と評価の概要を示している。診断システムの設計は、標的蛋白質の同定と標的蛋白質を用いた識別器の設計からなる。

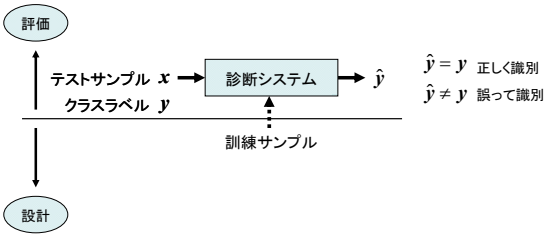

図2　診断システムの設計と評価

まず、標的蛋白質の同定は、パターン認識における特徴選択法によって行われる。特徴選択法は、特徴評価関数とその最適化法によって定められる。候補蛋白質の中から癌診断に有効な蛋白質を同定するため、候補蛋白質の有用性を評価する特徴評価関数を定めることが必要である。特徴評価関数には様々なものが提案されているが、以下の Fisher 比が最も簡単で、かつ有効であることが知られている。

Fisher 比は２クラス間の分離性を評価するもので、一種の統計的距離である（図３参照）。いま、蛋白質 i の Fisher 比は以下のように定義される。

$$F(i)=\frac{(\mu_1(i)-\mu_2(i))^2}{P(1)\sigma_1^2(i)+P(2)\sigma_2^2(i)}$$

$P(1)$：クラス ω_1 の事前確率

$P(2)$：クラス ω_2 の事前確率

ここで、Fisher 比の値が大きい程、その蛋白質は識別力に富み、癌の診断に有用である、と解釈される。

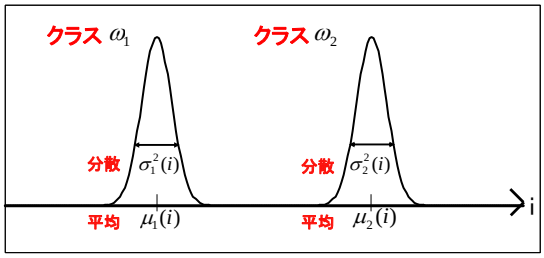

図3　Fisher比の説明

診断システムに用いられる識別器には様々なものがあるが、Bayes 識別器とニューラルネットワーク識別器がよく知られている（図４参照）。Bayes 識別器は、分布を仮定するパラメトリック識別器と分布を仮定しないノンパラメトリック識別器に大別される。ここでは構造が簡単な Fisher 線形識別器と k 最近傍識別器を採用する。

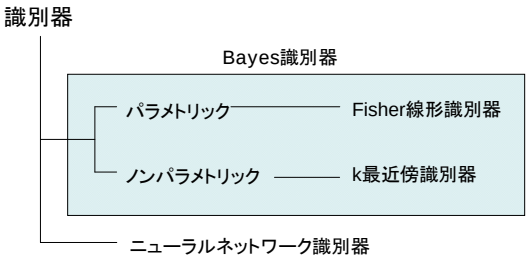

図4　代表的な識別器の分類

まず、Fisher 線形識別器は、以下のように設計される。各クラスに対して訓練サンプル集合 $\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N\}$ が与えられているものとする。この訓練サンプル集合を用いて、以下の式により各クラスの平均ベクトルと共分散行列を推定する。

平均ベクトル　$\boldsymbol{\mu}=\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{x}_i$

共分散行列

$$\boldsymbol{\Sigma}=\frac{1}{N-1}\sum_{i=1}^{N}(\boldsymbol{x}_i-\boldsymbol{\mu})(\boldsymbol{x}_i-\boldsymbol{\mu})^T$$

Fisher 線形識別器は

$$f_1(\boldsymbol{x})=\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)^T\left[\frac{1}{2}\Sigma_1+\frac{1}{2}\Sigma_2\right]^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)$$

$$f_2(\boldsymbol{x})=\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_2)^T\left[\frac{1}{2}\Sigma_1+\frac{1}{2}\Sigma_2\right]^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2)$$

を用いて、線形識別関数

$$f_1(\boldsymbol{x})-f_2(\boldsymbol{x})+\alpha=0 \qquad \alpha:\text{Cut off}$$

で表すことができる。

次に、k 最近傍識別器は、以下のように設計される。この識別器は、分布を仮定せず、入力パターンに最も近い３つサンプルの中で多数が属するクラスへ入力パターンを帰属させるものである。

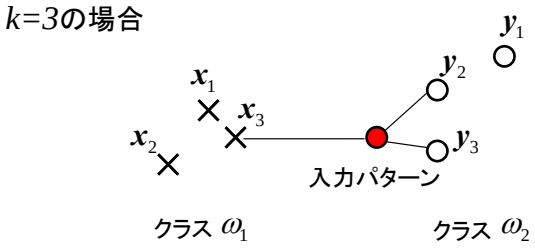

図5　k最近傍識別器の例

入力パターンに最も近い３つのサンプルはクラス ω_1 の x_3、クラス ω_2 の y_2 と y_3 であり、多数となるのはクラス ω_2 のサンプルであり、従って入力パターンはクラス ω_2 のパターンと識別される。

最後に、ニューラルネットワーク(3 入力の例)は、以下のように設計される。

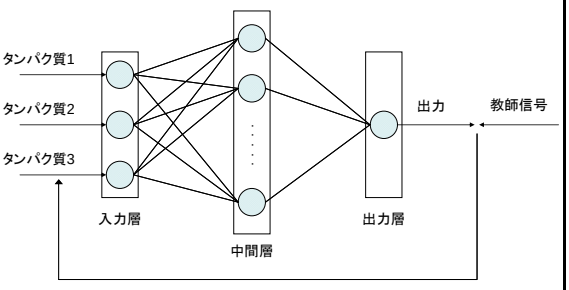

図6　誤差逆伝播法（BP法）による学習

ニューラルネットワーク識別器は、中間層により非線形な識別面を構成できるもので、高い汎化能力が期待できる。ここでは、中間層のユニット数を変え、重みの学習には誤差逆伝播法を用いた。図６に３入力のニューラルネットワーク識別器の構成を示す。

設計された識別器は、正解となるクラスラベルをシステムに隠した状態でパターンが入力され、システムはそれの帰属するクラス名を出力する。診断システムの出力したクラス名とパターンに付記されたクラス名とが一致した場合、診断システムは正しく識別したと言え、不一致の場合は誤った識別を行ったことになる。この誤った識別の確率が誤識別率となる。

識別器の性能評価は誤識別率によってなされるのが直接的である。ここでは誤識別率の推定として、再代入法と leave-one-out 法を組み合わせた改善法を採用する。再代入法は訓練サンプル集合とテストサンプル集合を同一とするため、真の誤識別率よりも誤識別率の推定値が小さいほうへ偏るという欠点を有する。これは訓練サンプル集合とテストサンプル集合の独立性がないためである。これを解決する手法として leave-one-out 法が提案されたが、今後は推定値が高い方へ偏ってしまう欠点がある。そこで、これらを組み合わせた誤識別率の推定法として

誤識別率＝（再代入による誤識別率＋leave-one-out 法による誤識別率）/２

を考える。これにより、負と正の偏りを相殺し、偏りを低減することができる。

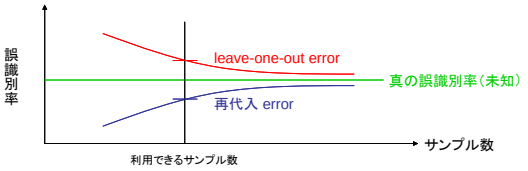

図7　サンプル数と誤識別率との関係

４．研究成果

肝臓癌の１６サンプルと非肝臓癌の２６サンプルを用いて肝臓癌の診断シミュレーション実験を行った。

実験１：Fisher 比による識別力の評価

４つの候補蛋白質の Fisher 比を計算し、それを表１に示す。蛋白質 A と蛋白質 B が有望であることが判明した。

表１　Fisher 比による評価

	Fisher 比	順位
蛋白質A	0.1394607	1
蛋白質B	0.1394565	2
蛋白質C	0.1007275	3
蛋白質D	0.0069805	4

実験２：蛋白質 A と蛋白質 B を用いた診断性能

次に蛋白質Ａと蛋白質Ｂを用いた識別器の診断性能を調べた。その結果を表２に示す。３層ニューラルネットワーク識別器で高い診断性能を得ることができた。なお、診断率は、100−誤識別率（%）で与えられる。

表2 肝臓癌の診断性能

識別器 / 性能	Fisher	3-NN	ニューラル
感度(%)	0	66	72
特異度(%)	98	69	92
診断率(%)	61	68	85

以上の結果から、候補蛋白質の中で蛋白質Ｂと蛋白質Ａが肝臓癌診断の標的蛋白質として有望であり、これらを用いた３層ニューラルネットワークにより

感度　**72**%
特異度　**92**%
診断率　**85**%

を達成することができた。今後、サンプル数を増加させた評価実験、識別器の高精度化などの更なる改良を行えば、肝癌診断システムの診断性能が実用レベルに達するという感触を得た。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計０件）

〔学会発表〕（計３件）
①澤村美貴子、浜本義彦、少数サンプル下における Bayes 誤識別率を用いた特徴選択法の評価、情報処理学会第 71 回全国大会 473-474(2009).
②林孝哉，坂本和史，浜本義彦，中村和行，岡正朗、二次元電気泳動画像を対象とした前処理に関する一考察、439-440(2007).
③林 孝哉，坂本 和史，中村 和行，岡 正朗，浜本 義彦、スポット検出のためのフィルタリングによるゲル画像の処理、平成 18 年度 電気・情報関連学会中国支部第 57 回連合大会講演論文集、144-144(2006).

〔図書〕（計０件）

〔産業財産権〕
○出願状況（計０件）
○取得状況（計０件）

〔その他〕
ホームページ等
http://www.ir.csse.yamaguchi-u.ac.jp/

６．研究組織
(1)研究代表者
浜本　義彦（HAMAMOTO YOSHIHIKO）
山口大学・大学院医学系研究科・教授
研究者番号：９０１９８８２０

(2)研究分担者
平林　晃（HIRABAYASHI AKIRA）
山口大学・大学院医学系研究科・准教授
研究者番号：５０２７２６８８
内村　俊二（UCHIMURA SHUNJI）
山口大学・大学院医学系研究科・助教
研究者番号：５０２０３５５０
中村　和行（NAKAMURA KAZUYUKI）
山口大学・大学院医学系研究科・教授
研究者番号：９０１０７７４８
蔵満　保宏（KURAMITU YASUHIRO）
山口大学・大学院医学系研究科・准教授
研究者番号：５０２８１８１１
飯塚　徳男（IIZUKA NORIO）
山口大学・大学院医学系研究科・准教授
研究者番号：８０３３２８０７

(3)連携研究者　なし