重回帰分析によるバイオマーカ発現量からの生理活性値推定

山森一人[1)] 上口真由美[2)] 吉原郁夫[3)]

Estimation of Physiological Activity Values from Protein Expression Levels with Multiple Regression Analysis

Kunihito YAMAMORI Mayumi KAMIGUCHI Ikuo YOSHIHARA

Abstract

Recent years, many researchers investigate functional foods which are reinforced useful constituents for human. These researches need to measure the physiological activities when a constituent is given to cells. However, it takes much time because the measurement of physiological activities needs many manual operations. Therefore, development of a system which can estimate the physiological activities of functional foods in short time is required. This research proposes a procedure to estimate physiological activity values of functional foods from protein expression levels. For this purpose we employ multiple regression analysis to derive an appropriate estimation function, and it is evaluated by using thirteen protein expression levels and various physiological activities.

Key Words:
multiple regression analysis,physiology activity, foods, protein expression levels,bio-marker

1 はじめに

食品には、これまで一般に 2 つの機能があると考えられてきた。第一の機能は生命を維持する機能で栄養機能と呼ばれ、第二の機能は感覚に訴える機能、つまりおいしさを感じさせる感覚機能である。このほかに第三の機能 (三次機能) として最近注目されるようになったのが高次の生命活動に対する調節機能であり、具体的にはアレルギー反応の低減や免疫能力を高めるなどの生体防御機能、高血圧、糖尿病、腫瘍、先天的代謝異常などを防止し回復する機能、神経活動や消化作用を調節する機能、過酸化脂質生成を抑制して老化を防御する機能などを指す [1)]。

食品の三次機能の評価において、その指標として生理活性がある。しかし、科学的に生理活性の値を測定するには多様な測定器を駆使して複雑な手順を踏む必要があり、収穫時期や産地によってもその値が異なるため、網羅的な測定は不可能である。そこで食品の生理活性をバイオマーカ発現量から間接的に推定する方法が研究されており、佃ら [2)] によりニューラルネットを用いた食品の機能性予測手法が報告されている。

本研究は、統計的手法によるバイオマーカ発現量からの生理活性値推定システムの構築を目的とする。生理活性推定には線形モデル式を用いる重回帰分析を用い、説明変数としてバイオマーカ発現量を用いてその推定精度を評価し、他の手法と比較した。また、バイオマーカの組み合わせを様々に変化させモデル式に最も適したバイオマーカの組み合せについても調査を行うと共に、福島らにより提案された SOM[3)] を応用したバイオマーカ選定法により求められたバイオマーカの組み合わせとの比較検討を行った。

[1)] 情報システム工学科，准教授
[2)] 情報システム工学専攻
[3)] 情報システム工学科，教授

2 重回帰分析の食品機能推定システムへの適用

バイオマーカとなるたんぱく質発現量から生理活性値を推定するモデル方程式として、(1) 式を考える。

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n \qquad (1)$$

このとき、$\hat{y}$ は生理活性推定値，$x_1, x_2, \ldots, x_n$ はバイオマーカ発現量，b_0 は切片，$b_1, b_2, \ldots, b_n$ は回帰係数，n は使用するバイオマーカ数を表す。n 個のバイオマーカ発現量と 1 個の生理活性値をあわせて 1 個のデータセットとすると、m 個のデータセットにおける生理活性の推定値はそれぞれ (2) 式のように表すことができる。

$$\left.\begin{array}{rcl} \hat{y_1} &=& b_0 + b_1 x_{11} + \ldots + b_n x_{1n}, \\ \hat{y_2} &=& b_0 + b_1 x_{21} + \ldots + b_n x_{2n}, \\ &\vdots& \\ \hat{y_i} &=& b_0 + b_1 x_{i1} + \ldots + b_n x_{in}, \\ &\vdots& \\ \hat{y_m} &=& b_0 + b_1 x_{m1} + \ldots + b_n x_{mn}. \end{array}\right\} \qquad (2)$$

また、y_i, e_i をそれぞれ i 番目のデータセットにおける生理活性観測値，観測値と推定値間の誤差とすると、$e_i = y_i - \hat{y_i}$ となる。ここで、全データセットに対して最も誤差が小さくなるように切片 b_0 及び回帰係数 $b_1, b_2, \ldots, b_n$ の値を定めるために、最小二乗法を用いる。最小二乗法の原理は、全てのデータセットにおける誤差 e_i の和を最小にする、すなわち、

$$E = \sum_{i=1}^{m} {e_i}^2 \Rightarrow min, \qquad (3)$$

であり、そのため $b_0, b_1, b_2, \ldots, b_n$ を未知数として E に対し偏微分を行い、その極小値を考える。ある未知数 b_k に対して編微分を行ったすると、極小において、

$$\frac{\partial E}{\partial b_k} = 0, \qquad (4)$$

である。したがって、これを行列式にまとめると、正規方程式である (5) 式が得られる [4)]。

$$\boldsymbol{\alpha} = \begin{pmatrix} \sum 1 & \sum x_{i1} & \ldots & \sum x_{in} \\ \sum x_{i1} & \sum {x_{i1}}^2 & \ldots & \sum x_{in} x_{i1} \\ \vdots & \vdots & \ddots & \vdots \\ \sum x_{in} & \sum x_{i1} x_{in} & \ldots & \sum {x_{in}}^2 \end{pmatrix},$$

$$\boldsymbol{b} = \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_n \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \sum y_i \\ \sum y_i x_{i1} \\ \vdots \\ \sum y_i x_{in} \end{pmatrix},$$

$$\boldsymbol{\alpha} \cdot \boldsymbol{b} = \boldsymbol{\beta} \qquad (5)$$

$det(\boldsymbol{\alpha}) \neq 0$ のとき、$\boldsymbol{\alpha}$ を LU 分解して逆行列を求め $\boldsymbol{b}$ を計算することで、(1) 式の切片 b_0 と回帰係数 $b_1, b_2, \ldots, b_n$ を求めることができる。

3 測定データの前処理

推定に用いるデータ群は、細胞に対して化合物を 3 種類の濃度でそれぞれ与えた場合のバイオマーカの発現量と生理活性値を測定したものである。

本章では、推定に用いるデータ群の測定項目と、そのデータ群の正規化及びデータセットの作成について述べる。

3.1 測定対象化合物

推定に使用するバイオマーカ発現量と生理活性値の測定において、細胞に与えた化合物を以下に示す。

- ポリフェノール
  - フラボノイド系
    - イソフラボン : Genistein, Daizein, Glycitein
    - フラボノール : Kaempferol, Galangin, Quercetin
    - アントシアニン : Cyanidin, Delphinidin,Pelargonidin
    - フラバノール : EGCG, EGC
    - スチルベノイド : Resveratrol
  - クロロゲン酸系 : ChlorogenicAcid
  - クルクミン系 : Curcumin
  - その他 : RosmarinicAcid
- 脂肪酸
  - 共役リノール酸 (CLA), ArachidonicAcid, LinoleicAcid
- 高血圧治療薬 (スタチン類)
  - FluvastatinNa, AtorvastatinCa, Simvastatin, Lovastatin, Pravastatin

- 抗ウィルス薬
  - IFN, Ribavirin
- その他
  - Capsaicin (Alkaloid), LipoicAcid, GABA, BITC

3.2 測定項目

バイオマーカ： 生体指標[5]とも呼ばれ、バイオマーカ発現量とは細胞のたんぱく質増減量のことである。今回用いたデータでは化合物各 3 濃度に対して 13 種類のバイオマーカ Thioredoxin，Survivin，HSP70，XIAP，FADD，TXNRD1，HSP90，MxA，tNOX，NQO1，ERK2，p53，Bcl2 の発現量がそれぞれ 6 点ずつ測定されている。

生理活性： 本研究では抗酸化ストレス活性と抗血管新生活性、細胞増殖抑制活性、抗炎症活性の計 4 種類の生理活性について推定を行う。各生理活性ではバイオマーカ発現量と同様に、化合物各 3 濃度に対して 3〜10 点ずつの生理活性値が測定されている。

3.3 正規化

推定に用いるバイオマーカ発現量は、まず初めに基準たんぱく質である GAPDH の発現量で正規化した後、各化合物ごとに濃度 $0\mu M$ 時のバイオマーカ発現量の平均値によって除算した値を使用する。また、生理活性値についても、各化合物ごとに濃度 $0\mu M$ 時の生理活性値の平均値によって除算した値を使用する。

3.4 データセットの作成

バイオマーカ発現量と各生理活性値のデータは、全て独立に測定されている。そのため、推定を行う前にバイオマーカ発現量の測定値と生理活性値の測定値間で対応付けを行う必要がある。

本研究では各バイオマーカ発現量と生理活性値の間で単回帰分析を行い、二つの値がお互いに対して独立でないことへの背反確率 P を求め、$P \leqq 0.05$ かつ最小となる組み合わせに基いてデータセットを作成した。また、これらの対応付けは４つの活性すべてにおいて行った。

4 生理活性値の推定実験

本章では重回帰分析を用いて実際に抗酸化ストレス活性，抗血管新生活性，細胞増殖抑制活性，抗炎症活性について推定を行った結果について述べる。重回帰分析における切片及び回帰係数の決定には全データセットの内約 8 割を用い、残りを推定率を調べるためのテストデータとして使用する。また推定式はすべて F 検定により検定を行い、テストデータの推定結果は以下に示す絶対誤差を 0.2 と規定したときの推定率として表す。

$$\text{推定率} = \frac{\text{規定誤差を満たすデータセット数}}{\text{テスト用データセット数}} \times 100[\%]$$

4.1 抗酸化ストレス活性推定実験

総データセット数は 540 個あり、その中で 443 個を重回帰分析のパラメータの決定に、残りの 97 個を推定実験に用いた。また、各データセットは全て正規化を行った後、対応付けを行ったものである。

初めに、13 種類全てのバイオマーカーを用いて推定を行ったところ、絶対誤差を 0.2 としたときの推定率は 25.8 %であった。ここで、1 種類から 13 種類までの全てのバイオマーカの組み合せについて重回帰分析を行い、以下に示す任意誤差を最小にするバイオマーカの組み合せを求めた。

$$\text{任意誤差} = \frac{E}{m}$$

表 2 にそのバイオマーカの組み合わせとそのときの任意誤差を示す。表 2 に示したバイオマーカを用いて推定を行ったところ、40.2 %まで推定率を向上させることができた。また、福島らによる SOM を用いた研究[3]より各活性に影響を与えると推察されるバイオマーカの組み合わせを用いた場合、推定率は 20.6 %であった。これらの推定率は他の手法による推定率と共に推定率比較表として表 3 にまとめた。表 3 に示した重回帰分析以外の手法による推定率は[6]を参考にしたものである。

4.2 抗血管新生活性推定実験

総データセット数は 270 個あり、その中で 225 個を重回帰分析のパラメータの決定に、残りの 45 個を推定実験に用いた。また、各データセットは全て正規化を行った後、対応付けを行ったものである。

抗酸化ストレス活性の推定実験と同様に推定を行ったところ、全てのバイオマーカを用いて推定を行った場合の推定率は 20.0 %であった。また、バイオマーカの全ての組み合せについて重回帰分析を行い、任意誤差を最小にするバイオマーカの組み合せを求めた。そのバイオマーカの組み合わせを表 2 に、そのときの推定率を表 3 に示す。表 2 に示したバイオマーカの組み合せを用いて推定を行った結果、28.9 %まで推定率が向上した。SOM による組み合わせについては、抗血管新生活性は福島らの研究 [3] において推定対象に含まれていないため、表 3 において no data と表示している。

4.3 細胞増殖抑制活性推定実験

総データセット数は 540 個あり、その中で 429 個を重回帰分析のパラメータの決定に、残りの 111 個を推定実験に用いた。また、各データセットは全て正規化を行った後、対応付けを行ったものである。

他の推定実験と同様に、まず全てのバイオマーカを用いて推定を行ったところ 92.8 %の推定率が得られた。また、バイオマーカの全ての組み合せについて任意誤差を比較したが、13 種類全てのバイオマーカを用いたとき任意誤差が最小になった。また、他の抗酸化ストレス活性と同様に SOM によるバイオマーカの組み合わせを用いて推定を行った結果、92.8 %の推定率を得た。

4.4 抗炎症活性推定実験

総データセット数は 540 個あり、その中で 409 個を重回帰分析のパラメータの決定に、残りの 131 個を推定実験に用いた。また、各データセットは全て正規化を行った後、対応付けを行ったものである。

他の推定実験と同様に、まず全てのバイオマーカを用いて推定を行ったところ 61.1 %の推定率が得られた。また、バイオマーカの全ての組み合せについて任意誤差を比較したが、抗細胞増殖抑制活性のときと同様に任意誤差がもっとも小さくなるのは 13 種類全てのバイオマーカを用いたときであることがわかった。また、他の活性と同様に SOM によるバイオマーカの組み合わせを用いて推定を行った結果、61.1 %の推定率を得た。

5 生理活性値推定実験への考察

表 1 及び表 3 から、分散が小さく比較的に生理活性値の値がまとまっている抗細胞増殖抑制活性

表. 1 基本統計量

生理活性名	パラメータ決定用データ数	平均	標準偏差	分散
抗酸化ストレス活性	443	1.21	0.67	0.45
抗血管新生活性	225	0.81	0.52	0.28
抗細胞増殖抑制活性	429	0.92	0.13	0.018
抗炎症活性	409	1.03	0.30	0.091

表. 2 任意誤差 E/m を最小にする組み合わせ

生理活性名	使用バイオマーカ群	任意誤差 E/m
抗酸化ストレス活性	Thioredoxin,Survivin,HSP70,XIAP,FADD,TXNRD1,HSP90,MxA,tNOX,NQO1,ERK2,p53	0.092
抗血管新生活性	Thioredoxin,Survivin,HSP70,XIAP,FADD,TXNRD1,HSP90,MxA,tNOX,NQO1,ERK2,p53	0.070
抗細胞増殖抑制活性	Thioredoxin,Survivin,HSP70,XIAP,FADD,TXNRD1,HSP90,MxA,tNOX,NQO1,ERK2,p53,Bcl2	0.013
抗炎症活性	Thioredoxin,Survivin,HSP70,XIAP,FADD,TXNRD1,HSP90,MxA,tNOX,NQO1,ERK2,p53,Bcl2	0.073

表. 3 推定率比較表

生理活性名	手法	推定率
抗酸化ストレス活性	重回帰分析	25.8 % (all) 40.2 % (E/m 最小) 20.6 % (SOM)
	ニューラルネット	37.2 %
	ベイズ識別	80.0 %
抗血管新生活性	重回帰分析	20.0 % (all) 28.9 % (E/m 最小) no data (SOM)
	ニューラルネット	27.6 %
	ベイズ識別	65.5 %
抗細胞増殖抑制活性	重回帰分析	92.8 % (all) 92.8 % (E/m 最小) 92.8 % (SOM)
	ニューラルネット	91.4 %
	ベイズ識別	no data
抗炎症活性	重回帰分析	61.1 % (all) 61.1 % (E/m 最小) 61.1 % (SOM)
	ニューラルネット	55.1 %
	ベイズ識別	65.0 %

や抗炎症活性のような活性については、推定率が高くなる傾向にあることが分かった。

推定率が低かった抗酸化ストレス活性については、測定データに異常に高い生理活性値が存在し、それらが推定率の低下を招いているのではないかと推察される。同様に、最も推定率が低かった抗血管新生活性については、逆に異常に低い生理活性値が測定されており、同様の結果となっている可能性がある。したがってどちらも測定値について担当者と協議し、これらの値への対応を検討する必要がある。

また、表3の結果が測定誤差に因るものでないとすると、推定データには非線形性が含まれている可能性があり、今後は非線形項を含めたモデル式を用いた推定を考慮する必要がある。

6 おわりに

本研究では、重回帰分析による生理活性値推定システムを構築し、それを用いてバイオマーカ発現量から生理活性値の推定実験をおこなった。また、バイオマーカの組み合わせを変えた場合の推定率についても比較評価した。

実験の結果、許容誤差を± 0.2 としたとき、抗酸化ストレス活性では 40.2 %、抗血管新生活性では 28.9 %、細胞増殖抑制活性では 92.8 %、抗炎症活性では 61.1 %の推定率が得られた。

今後はモデル方程式に非線形の項を加え、遺伝的プログラミング等を用いて最適なモデル方程式を探索してゆく予定である。

謝辞

本研究は、独立行政法人科学技術振興機構・地域結集型共同研究事業「食の機能を中心としたがん予防基盤技術創出」の一部として行なわれ、バイオマーカ発現量や生理活性値は宮崎大学農学部及び宮崎県産業支援財団コア研究室にて測定されたものである。関係者各位に深く感謝する。

参考文献

[1] 須見洋行, 食品機能学への招待 – 機能性食品とその効能–, 三共出版株式会社 (1997).

[2] 山森一人, 佃晋輔, 吉原郁夫, “揺らぎを含むタンパク質発現量からのニューラルネットワークによる食品の生理活性値の推定”, MEMOIRS OF THE FACULTY OF ENGINEERING UNIVERSITY OF MIYAZAKI, No. 36, pp. 345–350 (2007).

[3] T. Fukushima, K. Yamamoi, I. Yoshihara and K. Nagahama, “Feature extraction of protein expression levels based on classification of functional foods with SOM ”, The Thirteenth International Symposium on Artificial Life and Robotics 2008, pp. OS12–2 (2008).

[4] 足立堅一, 多変量解析入門, 篠原出版新社 (2005).

[5] 吉川敏一, 大澤俊彦, アンチエイジングと機能性食品 –今なぜバイオマーカーか–, シーエムシー出版 (2006).

[6] S. Togo, K. Yamamori, I. Yoshihara and K. Nagahama, “ Estimating physiological activities of functional foods from protein expression levels using bayesian classifier ”, The Thirteenth International Symposium on Artificial Life and Robotics 2008, pp. OS12–3 (2008).