GMDHによるタンパク質発現量からの食品生理活性値の推定

山森一人[1)] 岩崎敬太[2)] 吉原郁夫[3)]

Estimation of Physiological Activity Values from Protein Expression Levels with GMDH

Kunihito YAMAMORI Keita IWASAKI Ikuo YOSHIHARA

Abstract

Today, many people pay attention to prevent cancer. Many researchers focus on some kinds of foods which give good effect for human body to prevent cancer. These researches try to be clear the physiological activities of foods, that means how much effect does a food have for human, when human takes a food. But, there are many kinds of foods in the world, and their physiological activity also depend on the way of cooking process. So it is difficult to measure physiological activities directly. We have been proposed a method to estimate physiological activities from protein expression levels that can measure easier than that of physiological activities. In this paper, we try to make a model to estimate physiological activities from protein expression levels by using of Group Method of Data Handling(GMDH).

Key Words:
Group Method of Data Handling, physiology activity, bio-marker, protein expression levels

1 はじめに

今日の日本における死亡率の約3割はがんが占めており、その傾向は2015年まで変わらないと推定されている。その原因として人々の食生活が大きく影響しており、特に食の欧風化に伴うカロリーの摂取過剰があげられてきた[1)]。そういった中で、近年では食品を通じたがん予防が注目されつつある。このため、食品のがんに対する効果の測定が進められているが、その効果は産地や品種、加工法によっても異なる上、測定自体も複雑な手順を踏む必要があり、1つ1つ測定していくことは現実的ではない。

一般に、食品が持つ生物に対する影響は生理活性と呼ばれている。こうした生理活性は細胞内のタンパク質により測定されており、ある生理活性には特有のタンパク質(バイオマーカ)発現量が増減することが知られている。また、バイオマーカ発現量はELIZA等[1]を用いることで生理活性よりも比較的容易に測定することができる。

本研究では、精製した食品成分や医薬品をがん細胞に作用させた時のバイオマーカ発現量と生理活性を測定してモデル式を構築し、様々な成分を含む食品抽出物を細胞に作用させた時のバイオマーカ発現量からその生理活性値を推定することを目的とする。

このようなバイオマーカ発現量から生理活性値を間接的に推定する手法としては、久野ら[2)]によりニューラルネットワークを用いる手法などが提案されている。本研究は、予測やモデリング、システム同定などに幅広く応用されており、非線形モデルを作ることのできるGroup Method of Data Handling(GMDH)という手法を用いて、バイオマーカ発現量から生理活性値の推定を行った。

[1)]情報システム工学科准教授
[2)]情報システム工学科学生
[3)]情報システム工学科教授

[1]抗原抗体反応を利用してタンパク質を検出する方法。

本研究で用いる各測定データは宮崎県産業支援財団のコア研究室によって測定された。しかし、測定されたデータだけではデータ数が少なく、GMDH のモデル構築用データとしては不十分である。そのため、測定値からモデル構築用データを合成する手法を用いた。このことにより、本研究で用いるデータが生物由来のデータであるためにばらつきが生じやすく、不安定なものになりやすいことに対応した。

2 Group Method of Data Handling(GMDH)

2.1 GMDH によるモデル化

Group Method of Data Handling(GMDH) とは、Ivakhnenko[3] が 1968 年に提案した手法であり、簡単な非線形式を組み合わせて複雑な非線形モデルを自己組織化的に構成してゆく手法である。GMDH の一般的な特徴としては、

- 数少ない入出力データで非線形システムのモデリングを手軽に行うことができること、
- システムの構造に関する先験的な情報を必要とすることなく、モデル構造の自己選択が可能であること、

が挙げられる [4]。 一般に、GMDH の伝達関数は 2 入力 1 出力であり、入力変数を x_i と x_j とすると、最も簡単な伝達関数は (1) 式に示した 2 次式となる。

$$G(x_i, x_j) = a_0 + a_1 x_1 + a_2 x_2 + a_3 x_1 x_2 + a_4 x_1^2 + a_5 x_2^2. \quad (1)$$

ここで、(1) 式の a_k $(k = 0, \cdots, 5)$ は係数を表しており、これらは次節で述べる最小二乗法によって決定される。

図 1 は基本的な GMDH によるモデル構築の概要を図示したものである。以下で GMDH によるモデル構築アルゴリズム [5] について説明する。

Step 1：システムの入力 (説明変数) を $x_1, x_2, \cdots, x_m$、出力 (被予測変数) を y とし、m 個の説明変数から二つを取り出す全ての組み合わせを考える。従って、説明変数の組み合わせは $S={}_mC_2$ 通りである。また、それぞれの組み合わせに対し y を最もよく近似するように伝達関数 $G(x_i, x_j)$ の係数 a_k $(k = 0, \cdots, 5)$ を決定する。

$$\begin{aligned} u_1 &= G_1(x_1, x_2), \\ u_2 &= G_2(x_2, x_3), \\ &\vdots \\ u_s &= G_s(x_{m-1}, x_m). \end{aligned}$$

ここで出力された $u_1, \cdots, u_s$ の中から、出力 y との二乗誤差の小さい順に p 個を選択する。以下では、選択された p 個を $u_1, \cdots, u_p$ と記す。

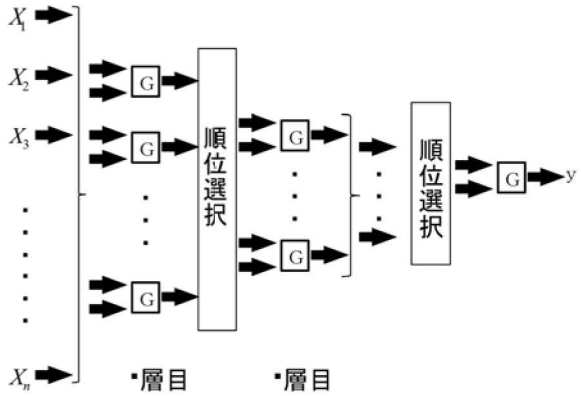

図 1 GMDH によるモデル構築の例

Step 2：Step 1 によって順位選択された $u_1, \cdots, u_p$ を改めて説明変数とみなし、同じ操作を繰り返す。

$$\begin{aligned} v_1 &= G_1(u_1, u_2), \\ v_2 &= G_2(u_2, u_3), \\ &\vdots \\ v_t &= G_t(u_{p-1}, u_p). \end{aligned}$$

Step 1 と同様に、出力された $v_1, \cdots, v_t$ の中から、出力 y との二乗誤差が小さいものから順に q 個を選ぶ。

Step 3: 同様にして、これ以上高次の組み合わせを作っても、出力 y との二乗誤差が小さくならなくなるまで、同じ操作を繰り返す。

以上のように、GMDH のアルゴリズムは各中間層で出力 y との二乗誤差が小さくなる説明変数の組み合わせを選択していき、出力 y との二乗誤差が変わらなくなるまで順位選択を続ける手法である。伝達関数は 2 入力であるので、出来上がるモデル式は図 2 のように説明変数を葉とし、各 $G(x_i, x_j)$ をノードとする完全二分木となる。

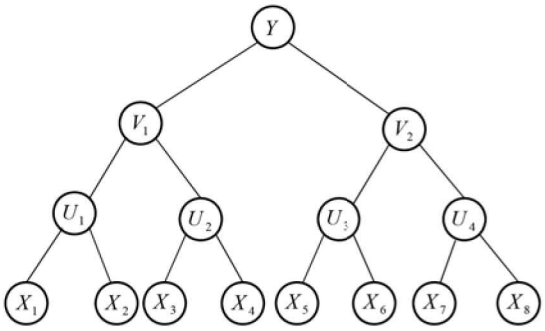

図 2　GMDH で構築されるモデルの例

3　測定データの前処理

3.1　測定データ

バイオマーカ発現量から生理活性値の推定を行うにあたって、宮崎県産業支援財団コア研究室で測定されたバイオマーカ発現量と生理活性値のデータは、30 種類の化合物を HepG2 細胞 (ヒト肝ガン由来細胞株) と、Jurkat 細胞 (急性リンパ性白血病由来細胞株) に作用させた時のタンパク質発現量と生理活性値である。表 1 は実験に使用した化合物とその濃度を示したものであり、ここでの化合物とは食品成分や医薬品といった試薬を表す。また、測定対象としたバイオマーカは Thioredoxin、Survivin、HSP70、XIAP、FADD、TXNRD1、HSP90、MxA、tNOX、NQO1、ERK2、p53、Bcl2 の 13 種類である。

今回推定実験の対象とした生理活性は血管新生抑制活性、細胞増殖抑制活性、抗炎症活性、抗酸化ストレス活性であり、以下にそれらの作用の概要を示す。

細胞増殖抑制活性： がん細胞が増殖するのを抑制する活性。この値が 1.0 より小さいほど活性が強い。

抗炎症活性： 花粉症やアトピー性皮膚炎などのアレルギー性炎症疾患の症状を低減させる活性。この値が 1.0 より小さいほど活性が強い。

抗酸化ストレス： 酸化反応により引き起こされる生体にとって有害な作用である酸化ストレスを抑える活性。この値が 1.0 より大きいほど活性が強い。

血管新生抑制活性： 新たな血管の生長を抑制し、ガンやリウマチ性関節炎などの疾病を抑制することが出来る活性。この値が 1.0 より小さいほど活性が強い。

表 1　モデル構築に使用した化合物名とその濃度

化合物名	濃度 (μM)					
	HepG2 細胞			Jurkat 細胞		
ArachidonicAcid	15	45	100	15	45	100
AtorvastatinCa	3.5	10	35	3.5	10	35
BITC	1.5	5	15	0.5	1.5	5
Capsaicin	10	60	150	1	3	10
ChlorogenicAcid	20	70	200	1	3	10
CLA9C	10	30	100	10	30	100
CLA12C	1	3	10	1	3	10
Curcumin	4	15	40	0.5	1.5	4.0
Cyanidin	40	150	400	5	15	40
Daizein	25	50	150	7	25	70
Delphinidin	15	70	200	5	15	45
EGC	10	30	60	3	10	30
EGCG	7	20	50	2	7	20
FluvastatinNa	7.5	15	50	7.5	25	75
GABA	100	300	1000	100	300	1000
Galangin	8	15	50	8	30	80
Genistein	10	20	60	3	10	30
Glycitein	10	30	100	10	40	120
IFN	100	300	1000	100	300	1000
Kaempferol	6	20	60	6	20	60
LinoleicAcid	20	50	150	5	20	50
LipoicAcid	0.1	0.3	1.0	0.1	0.3	1.0
Lovastatin	5	25	50	2.5	10	25
Pelargonidin	100	250	800	10	30	100
Pravastatin	100	300	1000	25	100	250
Quercetin	5	15	60	5	15	45
Resveratrol	10	30	80	10	30	100
Ribavirin	2	10	30	2	7	20
RosmarinicAcid	5	15	50	5	15	50
Simvastatin	3.5	10	35	3.5	10	35

以上のように、生理活性値は 1.0 を基準とした相対値となっている。

3.2　検証用データ

本研究では、3.1 節で述べた食品成分 30 種類の各々について、3 種類の濃度で HepG2 細胞と Jurkat 細胞に作用させた時のバイオマーカ発現量と生理活性値を用いて、GMDH の推定モデルを構成する。モデルを構築した後、その推定性能を検証するため検証用のデータを用いて実験を行った。検証用のデータは、モデル構築用に使用した工業的に精製された食品成分や医薬品ではなく、実際の食品抽出物を細胞に作用させて測定されている。表 2 は検証用に使われた食品抽出物名とその濃度である。

表2 検証用食品抽出物のその濃度

食品抽出物名	濃度 (μM)		
ニガウリ胎座熱水抽出物	300		
ブルーベリー熱水抽出物	5	10	50
ブルーベリー葉エタノール抽出物	5	10	50
ゴボウ熱水抽出物	300		
ニンジン葉熱水抽出物	400		
カモミール熱水抽出物	200		
ダイコン可食部熱水抽出物	1000		
ダイコン葉熱水抽出物	1000		
グラヴィノール熱水抽出物	40		
ヘベズ果皮熱水抽出物	120	350	
コマツナ熱水抽出物	100	300	
レモンバーム熱水抽出物	100	300	
玉葱葉熱水抽出物	100	300	1000
ローズマリー熱水抽出物	15	50	
大豆九州熱水抽出物	100	300	1000
スペアミント熱水抽出物	30	100	
スイオウ葉熱水抽出物	300		
ステビア熱水抽出物	60	200	
スイートバジル熱水抽出物	120	400	
サトイモ皮熱水抽出物	10	30	80
三番茶熱水抽出物	20	40	

3.3 モデル構築用データの対応付け

3.3.1 測定データの正規化

推定実験に用いるバイオマーカ発現量は、まず初めに基準のタンパク質であるGAPDHの平均発現量で各発現量を除し、さらに各化合物ごとに当該化合物を与えなかった時の平均発現量で除した値を使用する。また、生理活性値についても、各化合物ごとに当該化合物を与えなかった時の生理活性値の平均値によって除した値を使用する。したがって、バイオマーカ発現量、生理活性値ともに、当該化合物を加えなかったときの値を1.0とした相対値となる。また、測定されたタンパク質発現量や生理活性値は非負の実数値であり、最大で2.0~3.0程度の値となる。

3.3.2 モデル構築用データの合成

実際に測定されたデータのみではモデル構築に十分な数とはならないため、実測値に基づきモデル構築用のデータの合成を行った。測定されたデータは平均値を中心としたガウス分布(正規分布)に従っていると仮定し合成を行った。しかし、単純にガウス分布に従ってデータを合成すると、測定値が小さい場合、負の値が含まれてしまうことがある。そこで、本研究ではモデル構築用合成データの平均値に10を加え中心を移動させた。この操作により、合成したときのデータのピークを正の方向に移動させ、負の値の発生を抑えた。

モデル構築用合成データは、(2)式に示したボックスミューラー法により正規乱数になるよう合成した。なお、ボックスミューラー法とは測定値の平均と分散、一様乱数を用いて正規乱数を生成する方法である。

$$U = \mu + \sigma^2(-2\log v_1)^{\frac{1}{2}} \sin 2\pi v_2. \qquad (2)$$

ここで v_1, v_2 は $[0,1]$ の一様乱数であり、μ は平均、σ^2 は分散を表している。モデル構築用データは、すべての化合物とその濃度の組み合わせについて各バイオマーカ発現量及び生理活性値ごとに100,000個合成した。実験に用いるモデル構築用データは、これらの中からランダムに各化合物の各濃度ごとに10個づつ取り出したものを使用した。

3.3.3 単回帰分析によるモデル構築用データの対応付け

バイオマーカ発現量と生理活性値は食品成分と濃度が同じでも独立に測定されている。そのため、推定実験で用いるモデル構築用データを作成するには、バイオマーカ発現量と生理活性値の適切な組み合わせを見つける必要がある。本研究では、単回帰分析により適切な組み合わせを決定した。単回帰分析によって、バイオマーカ発現量と生理活性値の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、$p \leq 0.05$ かつ最小となる組み合わせに基づいてモデル構築用データを作成した。

4 実験条件

4.1 実験に用いたパラメータ

本研究では、予測やモデリング、システム同定などに幅広く応用されているGMDH(Group Method of Data Handling)を用いてバイオマーカ発現量から生理活性値を推定を行う。今回は中間層の数を変更して推定実験を行った。あまりに層の数を増やすと高次項が多数現れ、予測値が不安定になる。よってここでは、中間層が2層と3層のモデルを構築した。またGMDHのパラメータは下記の通りとした。

入力変数： 13(HepG2細胞)、9(Jurkat細胞)

出力変数： 1

表3　HepG2細胞の生理活性値と推定値の平均誤差

生理活性名	GMDHでの平均誤差		ニューラルネットでの平均誤差
	2層モデル	3層モデル	
細胞増殖抑制活性	0.038	0.034	0.023
抗炎症活性	0.134	0.096	0.084
抗酸化ストレス活性	0.372	0.382	0.344
血管新生抑制活性	0.087	0.064	0.076

中間層数2のモデルの順位選択： 以下の通り。

1層目： 上位4個を選択

2層目： 上位2個を選択

中間層数3のモデルの順位選択： 以下の通り。

1層目： 上位9個から8個を選択

2層目： 上位5個から4個を選択

3層目： 上位2個から2個を選択

モデル構築用データ数： 900(化合物)

検証用データ数： 240(食品抽出物)

4.2　実験方法

生理活性の推定手順は以下の通りである。

Step 1： 第3章で述べた手順により作成されたモデル構築用データにより、GMDHを用いて推定モデルを構成する。

Step 2： Step1によって構成された推定モデルに検証用データを入力し、生理活性値の推定を行う。

4.3　実験結果

本研究では中間層の数を2層と3層に変化させ、それぞれでGMDHによりモデル構築を行った。また、GMDHによる生理活性値の推定精度を調べるため、久野ら[2]によるニューラルネットワークを用いた手法との比較を行った。生理活性の実測値とそれぞれの手法における推定値は、(3)式に示した平均二乗誤差 e により評価する。

$$e = \frac{1}{n}\sum_{t=1}^{n}(y_t - y_t')^2. \qquad (3)$$

ここで、n は検証用データの個数、y_t は生理活性の実測値、y_t' は推定値を示している。

GMDHによって構築されたモデル式に検証データを入力した時の平均推定誤差を表3、表4に示した。ニューラルネットワークによる推定はHepG2細胞でのみ行われているため、GMDHとの比較は表3に示した同細胞でのみ行っている。

表4　Jurkat細胞の生理活性値と推定値の平均誤差

生理活性名	GMDHでの平均誤差	
	2層モデル	3層モデル
細胞増殖抑制活性	0.049	0.058
抗炎症活性	0.140	0.166
抗酸化ストレス活性	0.427	0.301

表3から分かるように、HepG2細胞の関しては2層モデル、3層モデルともにあまり違いが見られず、抗酸化ストレス活性のみ2層モデルの方が誤差が小さいという結果になった。一方、表4に示したJurkat細胞では、各層ともあまり違いが見られなかったが、HepG2細胞とは対照的で、抗酸化ストレス活性のみ3層モデルの方が誤差が小さいという結果になった。また、共通して言えることは、どちらの細胞についても、抗酸化ストレス活性の平均推定誤差が大きいということであった。特に、カモミールやローズマリーなどの生理活性値が大きい食品抽出物では、推定誤差が2.0～.0程度と大きくなった。また、GMDHを用いた場合とニューラルネットワークを用いた場合を比較して見てみると、大きな差は出ていないものの全体的にはニューラルネットワークの方が本研究で用いたGMDHよりもやや誤差が小さかった。

4.4　考察

本章では、HepG2細胞で細胞増殖抑制活性、抗炎症活性、抗酸化ストレス活性、血管新生抑制活性の4つの活性についての推定を行い、Jurkat細胞では細胞増殖活性、抗炎症活性、抗酸化ストレス活性の3つの活性についての推定を行った。その結果、どちらの細胞も細胞増殖抑制活性については小さい誤差で推定が可能であった。この活性は比較的生理活性値がまとまっていたため、推定値と実測値との誤差が少なかったのではないかと推察される。一方で抗酸化ストレス活性について

は推定誤差が大きく、この原因として極めて高い生理活性値を示す食品抽出物が含まれていることが挙げられる。また、生理活性値が相対的に大きい、あるいは小さい食品成分についても推定誤差が大きく、これはモデル構築用データにこうした化合物が含まれなかったためと考えられる。

5 おわりに

今日では日本人の３人に１人が「がん」にかかると言われており、人々のがんに対する関心がますます深くなっている。そういった中で、食を通じたがん予防は特段の努力を意識して行うことなく実行できることから注目されており、食品の生理活性を知ることはますます重要になってきている。しかしながら、生理活性の直接の測定や評価には時間や費用がかかり過ぎて必ずしも実用的ではない。

本研究では生理活性を直接測らず、バイオマーカ発現量からモデル式により生理活性を推定する方法を開発することを目的とし、非線形モデルを作ることの出来るGMDHを利用した手法を提案した。推定実験では中間層が２層の場合、３層の場合について評価を行った。その結果、細胞増殖抑制活性については推定誤差が小さく、高い精度での推定が可能であった。しかし、抗酸化ストレス活性が極めて大きいカモミールやローズマリー、あるいは極端に低い抗炎症活性を示すローズマリー、同じく血管新生抑制活性が低いグラヴィノールなどの食品抽出物については誤差が大きいという結果になった。

また本研究では、GMDHの推定精度を確かめるため、久野ら[2]によるニューラルネットワークを用いた手法との比較を行った。各活性ごとにはあまり差が見られなかったが、全体的には本研究で用いたGMDHよりもニューラルネットワークの方が推定誤差がやや小さいという結果になった。

しかし、GMDHと同様にニューラルネットワークでも生理活性値が極端に大きい、あるいは小さい食品についても推定誤差が大きくなっていた。これは、検証用データにモデル構築用データには見られないほどその活性が大きく異なる食品成分が含まれており、モデル構築用合成データの使用によっても対応できなかったためと考えられる。

今後の課題としては、モデル構築用データ数を一層増加させた場合や、新しい活性への対応、またバイオマーカの数や種類などを変化させた場合などの実験が挙げられる。

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] 大澤俊彦, 大東肇, 吉川敏一（編）, “がん予防食品-フードファクターの予防医学への応用”, シーエムシー (1999).

[2] T.KUNO, M.KAMIGUCHI, K.YAMAMORI, I.YOSHIHARA and K.NAGAHAMA, “Development of Physiological Activity Estimation Method of Foods Using Amplitude Extended Neural Networks”, *Proc. The Fourteenth Int'l Symp. on Artificial Life and Robotics*, pp. 658–661 (2009).

[3] A.G.IVAKHNENKO, “Polynomial Theory of Complex Systems”, *IEEE Trans. Systems*, Vol. SMC-1, No. 4, pp. 364–378 (1971).

[4] 池田三郎, 椹木義一, “GMDH(発見的自己組織化法)と複雑な系の同定・予測”, 計測と制御”, Vol. 114, No. 2, pp. 185–195 (1975).

[5] 吉原郁夫佐藤周一, “GA を用いた非線形モデル構築の最適化–GA と GMDH の融合–”, 情処研報（1996-ICS-105）, Vol. 1996, No. 78, pp. 1–6 (1996).