Estimating Physiological Activities of Functional Foods from Protein Expression Levels Using Bayesian Classifier

Shunichi TOGO[1], Kunihito YAMAMORI[1], Ikuo YOSHIHARA[1], Kiyoko NAGAHAMA[2]

[1] *1-1, Gakuen Kibanadai-nishi, Miyazaki, 889-2192 , Japan, University of Miyazaki*
[2] *16500-2, Higashi Kami-Naka, Sadowara, 880-0303, Japan, Miyazaki Prefectural Industrial Foundation*

(Tel : +81 985 58 7589; Fax : +81 985 58 7589)
(Email address tougou@taurus.cs.miyazaki-u.ac.jp)

Abstract: This paper addresses a novel method to estimate physiological activities of functional foods from protein expression levels using Bayesian classifier. The procedure of classification is as follows. Firstly, making clear relations between physiological activity values of functional foods and protein expression levels using simple regression analysis, secondly constructing Bayesian classifier to classify constituents of foods into three classes; "positive", "neutral" or "negative". Finally, giving the protein expression levels of constituents of foods to be examined to the Bayesian classifier and finding the class which the foods will belong to. Physiological activity of the foods is expressed by the above-mentioned class. We tackle to estimate five physiological activities by using expression levels of thirteen proteins.

Keywords: Functional food, Physiological activities, Protein expression levels, Bayesian classifier

I. INTRODUCTION

Since many people pay attention to keeping their health such as preventing lifestyle-related illnesses in recent years, researchers and companies focus on some kinds of foods for health [1].

Foods have three functions; nutrition, taste and effect on human body. In particular, since the third function of foods is attracting a great deal of attention, many researchers investigate functional foods which are reinforced their physiological activity. The researches need to measure the quantity of constituents of foods, and evaluate the effect of physiological activity of each constituent. However, foods involve many kinds of constituents and it needs many complicate operations to measure and evaluate these constituents. So development of high throughput method to estimate physiological activities is expected. For that purpose, we propose a new method to estimate physiological activities using Bayesian classifier based on protein expression levels when an industrially synthesized constituent of foods is given to lysis of cells. We measure protein expression levels and physiological activities for several kinds of industrially synthesized constituents in advance and the probability densities for all proteins are calculated. Bayesian classifier is constructed to classify these industrially synthesized constituents into appropriate classes according to their physiological activity. When a set of protein expression levels of foods is given to the classifier, it can say the appropriate classes for foods.

II. BAYESIAN CLASSIFIER

1. Bayes' theorem

Bayes' theorem is applicable to classification based on statistical features of the data. Let $C_i\,(i=1,\ldots,K)$ be a set of classes, $\boldsymbol{x}$ be a feature vector of the data that the class is not known [2]. Bayes' theorem is defined by Equation (1),

$$P(C_i \mid \boldsymbol{x}) = \frac{P(\boldsymbol{x} \mid C_i)P(C_i)}{\sum_{i=1}^{K} P(C_i)P(\boldsymbol{x} \mid C_i)}, \tag{1}$$

where $P(C_i \mid \boldsymbol{x})$ and $P(\boldsymbol{x} \mid C_i)$ denote the posterior probability and the probability density, respectively. The class C_I with maximum $P(C_i \mid \boldsymbol{x})$ is the class for $\boldsymbol{x}$ as shown in Equation (2).

$$C_I = \arg\max_i P(C_i \mid \boldsymbol{x}). \tag{2}$$

In this research, a feature vector $\boldsymbol{x}$ consists of expression levels of thirteen proteins and each class C_i corresponds with the effect of physiological activity of foods. We assume the prior probabilities $P(C_i)$ of all classes are the same as shown in Equation (3) because the prior probability is unknown.

$$P(C_i) = 1/K. \tag{3}$$

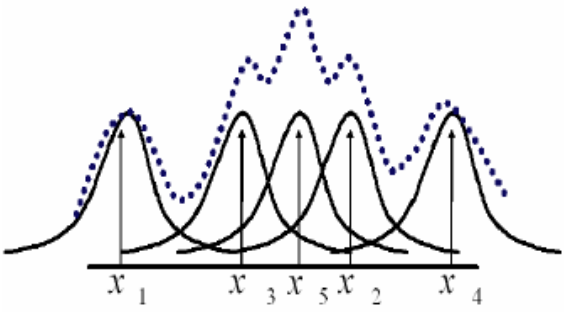

Fig.1 Probability density as sum of kernel functions

The posterior probability $P(C_i \mid \boldsymbol{x})$ is proportional to the probability density $P(\boldsymbol{x} \mid C_i)$, because all the prior probabilities $P(C_i)$ are same for all classes and the denominator $\sum_{i=1}^{K} P(C_i)P(\boldsymbol{x} \mid C_i)$ is common in Equation (1).

2. Parzen Window method

Probability density is seldom known in advance. Therefore we use Parzen window method [3] which can derive the probability density of each class based on sum of kernel functions as shown in Fig.1. Gaussian function in Equation (4) is usually employed for kernel function. Here, μ_d and σ_d^2 denote the average and the variance of the d-th element of the feature vector $\boldsymbol{x}$, respectively.

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right). \qquad (4)$$

When a sample x_k is observed, Gaussian distributions as shown as the solid line in Fig.1 is calculated. The average μ and the variance σ of this Gaussian distribution are x_k and h, respectively. Here h is called as the width of the kernel function and details are described in the following. If five samples $\{x_1, \cdots, x_5\}$ which belong to the same class are observed, the sum of kernel function is shown with dotted line in Fig.1. The sum of kernel function is consistent with the probability density for the class. When $D-$dimensional feature vectors $\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_j, \ldots, \boldsymbol{x}_N\}$ that belong to the same class C_i are given, $P(\boldsymbol{x} \mid C_i)$ is defined by Equation (5),

$$P(\mathbf{x} \mid C_i) = \frac{1}{Nh^D} \sum_{j=1}^{N} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(\mathbf{x}-\mathbf{x}_j)^2}{2h^2} \right), \qquad (5)$$

Classification function g_i for the class C_i is defined by Equation (6) derived from Equation (5) and Equation (1).

$$g_i = P(C_i \mid x) = \frac{P(C_i)}{P(x)} \frac{1}{Nh^D} \sum_{j=1}^{N} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(\mathbf{x}-\mathbf{x}_j)^2}{2h^2} \right). \qquad (6)$$

Equation (6) says that the classification depends on the width of kernel function h [4]. So it is necessary to find appropriate h to achieve accurate classification.

We optimize parameter h as follows; a test sample is selected and Bayesian classifier with various value of h is constructed by using the others. The classifier that can classify the test sample into the correct class with the highest posterior probability is selected. These are repeated as changing the test sample.

III. EXPERIMENT

1. Data for experiments

A. Constituents and Concentrations

We use thirty kinds of constituents which have three concentrations as shown in Table 1. The columns of Table 1 show concentrations for each constituent. The central column of the concentration is so-called *IC50*. The constituents are poured over HepG2 cells and protein expression levels and physiological activities are measured.

B. Protein Expression Levels

We measured protein expression levels for each constituent in Table 1 six times. So a data-set involves 540($=30\times3\times6$) experimental results. We also measure the expression levels of following proteins; Thioredoxin, Survivin, HSP70, XIAP, FADD, TXNRD1, HSP90, MxA, tNOX, NQO1, ERK2, p53 and Bcl2.

The protein expression levels are normalized in two aspects; one is the expression level of protein GAPDH, and the other is the average expression levels without the constituent.

C. Physiological Activities

We select five physiological activities for estimation as follows;

- Antiviral activity
  Activity to suppress viral cancer
- Anti-proliferative activity

Activity to suppress proliferation of cancer

- Anti-angiogenic activity
  Activity to suppress angiogenesis around cancer
- Antioxidant activity
  Activity to suppress oxidization stress
- Anti-inflammatory activity
  Activity to suppress allergic inflammation

The physiological activities are divided into three classes; "positive", "neutral" and "negative". "Positive" means the effect of physiological activities is good for human, "negative" also means the effect of physiological activities is bad for human, and "neutral" also means there is little effect of physiological activities for human.

Table 1 Constituents of food

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

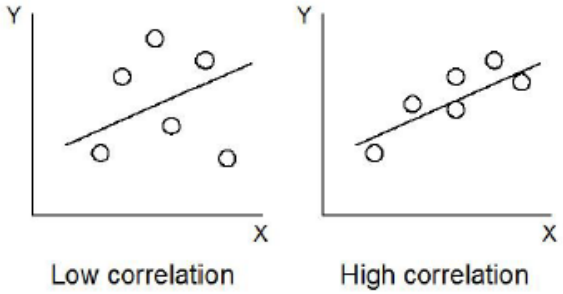

Fig.2 Correspondence between the protein expression levels and the physiological activities

D. Appropriate correspondence between the protein expression levels and the physiological activities

Since the protein expression levels and the physiological activities are measured separately, it is needed to find an appropriate correspondence between the protein expression levels and the physiological activities even if they are observed by the same constituent and concentration. For this correspondence, we use simple regression analysis, whose concept is illustrated in Fig.2. The X-axis and the Y-axis denote the protein expression levels and the physiological activities, respectively.

We select a combination of the protein expression levels and the physiological activities with the minimum p-value among all the combinations because the smallest p-value has the highest degree of correspondence.

2. Procedure of Experiment

The procedure to estimate physiological activities of foods consists of 3 steps as shown in Fig.3.

step-1. We select one constituent for validation, and the other for model building. We make the probability density function from model-building constituents using Parzen Window method and verify whether one constituent for validation is classified into the proper class or not.

step-2. We repeat the step-1 with various width of kernel function and decide a width of kernel function so as to classify the validation data into proper class with the highest probability.

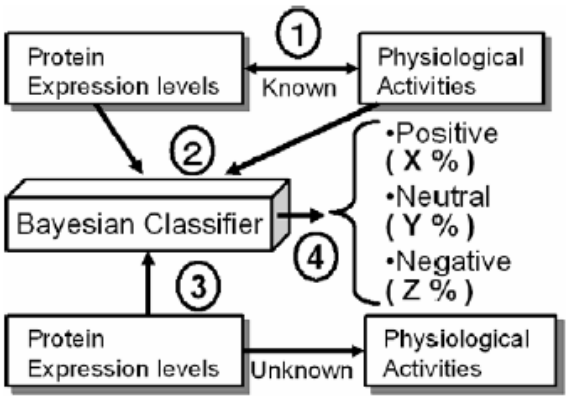

Fig.3 Procedure of experiment

step-3. We classify four kinds of constituents of foods preparing for validation into classes based on the probability density found at the step-2.

IV. CONCLUSION

Many people pay attention to keeping their health by taking some kinds of foods. Because foods involve many kinds of constituents, a new method to estimate their physiological activities of foods is expected.

In this paper, we proposed a novel method to estimate physiological activities of functional foods from the protein expression levels using Bayesian classifier. Since the protein expression levels and the physiological activities are measured separately, we employ simple regression analysis for finding an appropriate correspondence between the protein expression levels and the physiological activities. We also use industrially synthesized constituent of foods because the measurements of physiological activities for these synthesized constituent are relatively easier than that of actual constituent of foods.

Future work is to go into particulars to evaluate our proposed method using constituents of foods.

ACKNOWLEDGEMENT

A part of this research is supported by JST Grant-In-Aid of Miyazaki Prefecture Collaboration of Regional Entities for the Advancement of Technological Excellence.

REFERENCES

[1] G.Mazza (1998), Functional Foods. Technomic Pub Co
[2] Hiroshi Watanabe (2007), Introduction to Bayesian statistics (in Japanese). Fukumura Shuppan Inc.
[3] Christopher M. Bishop (1995), Neural Networks for Pattern Recognition. OXFORD University 49-57
[4] Hideki Asoh, Koji Tsuda, Noboru Murata, et al. (2005), Pattern recognition and statistics of learning (in Japanese). Iwanami Shoten, Publishers 1-43