Feature Extraction of Protein Expression Levels Based on Classification of Functional Foods with SOM

Tamon FUKUSHIMA[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)
(fuku@taurus.cs.miyazaki-u.ac.jp)

Abstract: We investigate relations between physiological activities and the protein expression levels of functional foods using Self-Organizing Map (SOM). The input vectors to SOM are composed of the protein expression levels and the physiological activity. A competitive node has two kinds of weights; one is for protein expression levels, the other is for physiological activity. A winner node is decided only by the weights for protein expression levels, and all weights in each node are updated. Each node has artificially generated value of physiological activity. The nodes can be categorized by the above-mentioned physiological activity. The well-trained SOM gives us some suggestions about relations between physiological activities and the protein expression levels.

Keywords: physiological activity, protein expression level, food constituent, functional foods, SOM

I. INTRODUCTION

Recently, health and disease are live issues, and many people pay their attentions to functional foods which are reinforced physiological activities [1]. However, it seems that the physiological activity of food would vary according to seasons and places. Direct measurement of physiological activity is complicated. So a new alternative method for estimating physiological activity is expected.

We focus on protein expression levels which are observed when a constituent in food is given to cells. If we use chemical compounds as these constituents, it is easy to measure quantitative and manageable physiological activities and protein expression levels. By quantitative and manageable measurement, we try to show some relations between physiological activities and protein expression levels. If we can obtain the relation, we can compile knowledge to estimate physiological activities from protein expression levels easily. We reveal relations between physiological activities and the protein expression levels with Self-Organizing Map (SOM).

This paper is consisted of the followings. Section 2 mentions detail of SOM. Section 3 shows the detail of experiment and discusses the result. Section 4 is conclusion.

II. SELF-ORGANIZING MAP

SOM [2] is a kind of neural networks [3] and it can project high-dimensional data on two-dimensional map. The lattice of the grid is either hexagonal or rectangle. SOM is used in widely field in real world such as engineering, economic science, linguistics, etc.

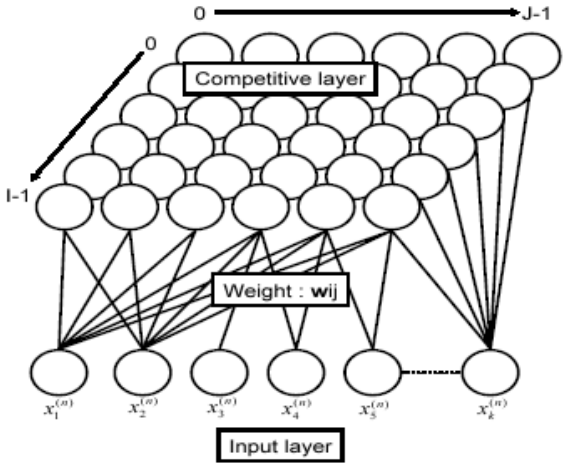

Fig. 1 The structure of SOM

As shown in Fig.1, SOM has two layers; one is the input layer and the other is the competitive layer. A node in the competitive layer is connected to all nodes in the input layer via adjustable weights. An input vector consisting k elements is shown as $x\ (= x_1, x_2, \cdots, x_k)$, and the n-th input vector is denoted as $\boldsymbol{x}^{(n)}$. The position of the node in the competitive layer is indexed by symbol $i\ (= 0, 1, \cdots, I-1)$ and $j\ (= 0, 1, \cdots, J-1)$. The weight vector of each node in the competitive layer is $\boldsymbol{w}^{ij} (= w_1^{ij}, w_2^{ij}, \cdots, w_k^{ij})$.

Fig.2 shows the learning procedure of SOM. At step-1, the weight vectors of nodes in the competitive layer are initialized at random. At step-2, each element of .an input vector is given to corresponding input node. At step-3, a winner node in the competitive layer is selected which has the minimum distance between the

input vector and the weight vector of the node. The above-mentioned distance is defined as Equation (1).

$$D = \sum_k (x_k^n - w_k)^2. \quad (1)$$

At step-4, the weight vector around the winner node is updated according to Equation (2).

$$^{(new)}w^{ij} = {}^{(old)}w^{ij} + \alpha(t)(x^n - {}^{(old)}w^{ij}), \quad (2)$$

where t $(\leq T)$ denotes the number of iterations, $\alpha(t)$ $(0 < \alpha(t) < 1)$ and $\beta(t)$ represent learning rate and a length of rectangle for weight updating area, respectively. $\alpha(t)$ and $\beta(t)$ are defined in Equation (3) and Equation (4), respectively.

$$\alpha(t) = \alpha_{init}(1 - t/T), \quad (3)$$

$$\beta(t) = \beta_{init}(1 - t/T), \quad (4)$$

where α_{init} and β_{init} each are initial values. Equations (3) and Equation (4) imply both the learning rate and the weight updating area reduce as the iteration increase.

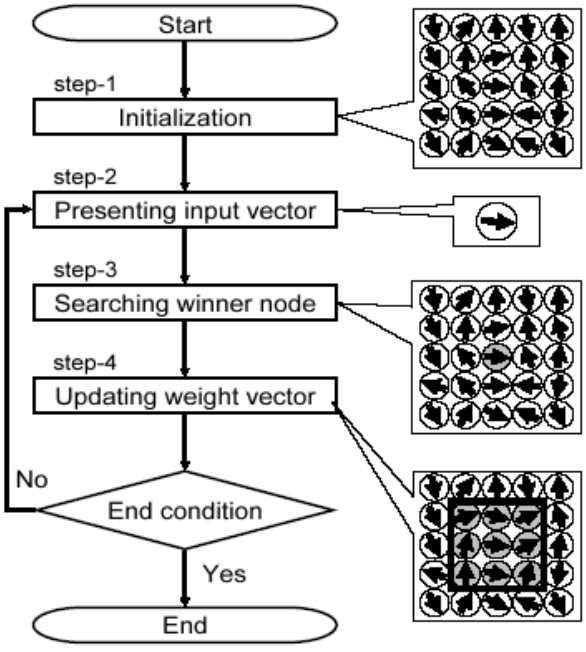

Fig. 2 The learning procedure

III. EXPERIMENTS AND DISCUSSION

1. Structure of the input vector

An input vector consists of fourteen elements; thirteen elements for protein expression levels, and one for physiological activity value. The names of proteins are as follows;

- Thioredoxin
- Survivin
- HSP70
- XIAP
- FADD
- TXNRD1
- HSP90
- MxA
- tNOX
- NQO1
- ERK2
- p53
- Bcl2.

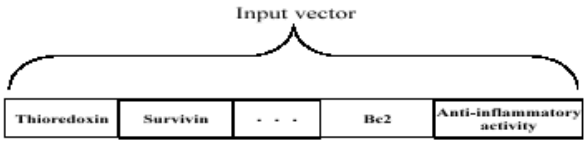

Fig. 3 An example of input vector

We focus on anti-inflammatory activity and anti-oxidative stress activity as physiological activities. An example of input vectors is shown in Fig.3. The constituents to measure their protein expression levels and physiological activities are shown in Table 1.

Table 1 Food Constituents

	concentrations(μM)			ID
LipoicAcid	100	300	1000	A
EGCG	7	20	50	B
Genistein	10	20	60	C
Daizein	25	50	150	D
Glycitein	10	30	100	E
Quercetin	5	15	60	F
Cyanidin	40	150	400	G
Pelargonidin	100	250	800	H
Delphinidin	15	70	200	I
Curcumin	4	15	40	J
GABA	100	300	1000	K
Resveratrol	10	30	80	L
ArachidonicAcid	15	45	100	M
CLA12C	1	3	10	N
CLA9C	10	30	100	O
EGC	10	30	60	P
Kaempferol	6	20	60	Q
IFN	100	300	1000	R
Ribavirin	2	10	30	S
FluvastatinNa	7.5	15	50	T
AtorvastatinCa	3.5	10	35	U
Simvastatin	3.5	10	35	V
Lovastatin	5	25	50	W
Pravastatin	100	300	1000	X
ChlorogenicAcid	20	70	200	Y
Galangin	8	15	50	Z
RosmarinicAcid	5	15	50	a
Capsaicin	10	60	150	b
BITC	1.5	5	15	c
LinoleicAcid	20	50	150	d

2. Learning procedure of SOM

SOM usually uses all the elements of the input

vector and the weight vector to calculate their distance. But we only use thirteen protein expression levels to decide a winner with minimum distance between the input vector and the weight vector. After deciding the winner node, all fourteen elements of the weight vector are updated according to Equation (2). Deciding the winner node based on only protein expression levels lets the SOM organize competitive nodes according to the similarity between protein expression levels. In addition, since our algorithm adjusts all weights including the element for physiological activity, appropriate physiological activity for each competitive node is also generated automatically. By comparing the organized clusters based on the protein expression levels with generated physiological activity values, we try to obtain relations between protein expression levels and physiological activities.

3. Conditions of experiments

The parameters of SOM are as follows;

- Competitive layer size: 60×60
- α_{init}: 0.7
- β_{init}: 24
- Maximum iterations: 1,000
- Number of trials: 5

4. Results and Discussions

A. anti-inflammatory activity

Fig.4 shows a self organized map for anti-inflammatory activity. The dark area means that the nodes have low activity values, and white areas have high activity values. The area A in Fig.4 denotes the nodes where anti-inflammatory activity is the lowest (1st minimum). The area B in Fig.4 denotes the nodes where anti-inflammatory activity is 2nd minimum. As shown in Fig.4, some clusters with similar anti-inflammatory activity are automatically organized.

Fig.5 shows the average value of the weights corresponding to the protein expression levels in the area A and B. As shown Fig.5, expression levels of protein such as Survivin, FADD and HSP90 in the 1st minimum group (area A) are low. Concretely, expression levels of proteins such as Thioredoxin, HSP70 in the 2nd minimum group (area B) are high, and TXNRD1, HSP90 are low. Fig.5 also shows that the expression level of HSP90 suggests characteristic value in both the area A and B. This suggests that the low expression level of HSP90 is a key of high anti-inflammation activity. Although the nodes in both the area A and B have high anti-inflammatory activity, HSP90 is different between them. It means that the different factor works for high anti-inflammation activity in cells. More experiments and discussions will be needed for clarifying this phenomenon.

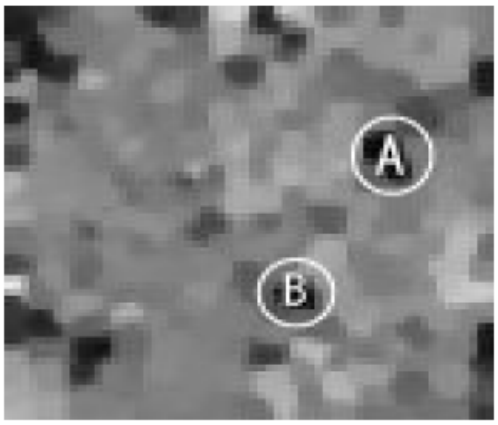

Fig. 4 SOM categorized by anti-inflammatory activity

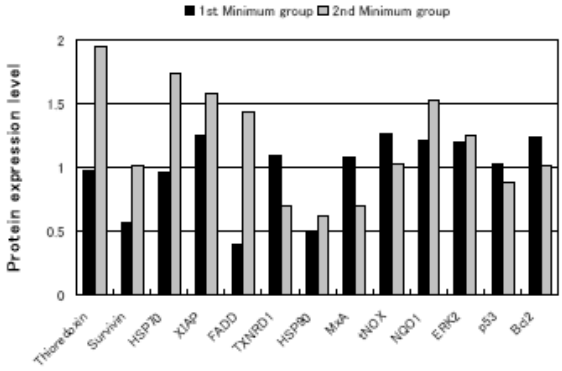

Fig. 5 Protein expression levels in the area A and B in Fig.4.

B. anti-oxidative stress activity

Fig.6 shows a self organized map for anti-oxidative stress activity. The dark area means that the nodes have low activity values and white area have high activity values. The area A denotes group where anti-oxidative stress activity is highest (1st maximum). The area B denotes the group where anti-oxidative stress activity is 2nd maximum. The competitive nodes organized by protein expression levels also compose some clusters based on anti-oxidative stress activity.

Fig.7 shows the average value of weights in the area A and B. As same as anti-inflammatory activity, the average protein expression levels are different from that between nodes in the area A and B. In particular, the proteins, Thioredoxin, HSP70 and TXNRD1 show large difference. It also suggests that the different factor works for high anti-oxidative stress activity in cells. On the other hand, the protein expression level of Survivin is relatively low in both the area A and B. It means that Survivin is probably a key protein to evaluate the anti-oxidative stress activity.

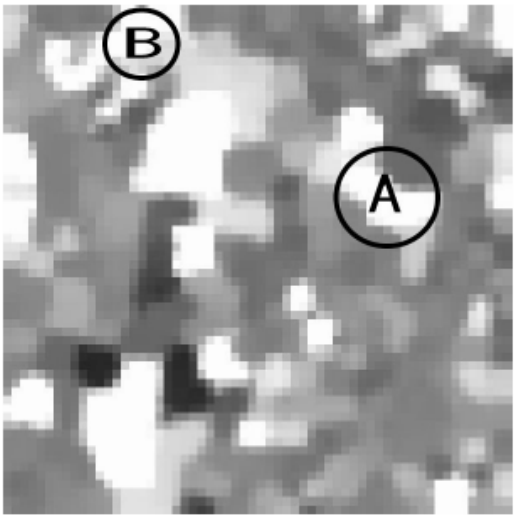

Fig. 6 SOM categorized by anti-oxidative stress activity

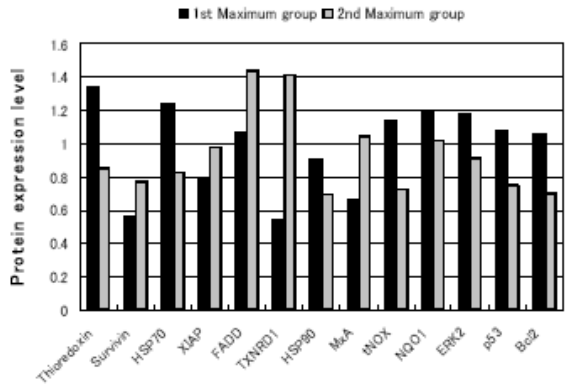

Fig. 7 Protein expression levels in area A and area B in Fig.6.

IV. CONCLUSION

In this paper, we tried to develop a new method to distinguish food if it is functional or not. For that purpose, SOM is employed to clarify relation between physiological activities and the protein expression levels.

Firstly, we compose input vectors by the protein expression levels. The input vector are organized and classified into some classes by SOM. Secondly, we compare the classified results with physiological activity values. Finally we discuss on the characteristic protein expression levels represented as weights in the nodes.

The experiments suggest that there exist two relations between physiological activities and the protein expression levels. One is low expression level of HSP90 is corresponding to low anti-inflammatory activity. The other is low expression level of Survivin to high anti-oxidative stress activity. In addition, our experiments also suggest that different factor will be work even if they cause the same physiological activity.

Future works will be to make clear the relation between protein expression level and physiological activity more concretely.

ACKNOWLEDGEMENT

This study is supported in part by Miyazaki Prefecture Collaboration of Regional Entities for the Advancement of Technological Excellence.

REFERENCES

[1] G.Mazza (1998), Functional Foods. Technomic Pub Co
[2] T.Kohonen (1996), SELF ORGANIZING MAPS (in Japanese). Springer Verlag Tokyo
[3] Y.Yoshitomi (2002), Neural Network (in Japanese). Asakura Publishing Co. Ltd