Finding an Area with Close Phenotype Values to Predict Proteins That Control Phenotypes

Takatoshi Fujiki[†], Empei Gaku[†], and Takuya Yoshihiro[‡]

[†]Graduate School of Systems Engineering, Wakayama University, Japan
[‡]Faculty of Systems Engineering, Wakayama University, Japan
{s101044, s121010, tac}@sys.wakayama-u.ac.jp

Abstract - Because phenotypes of living creatures are expressed reflecting on interactions among genes and proteins, relations among phenotypes and proteins (or genes) have been regarded as a key issue to be clarified to understand the system of creatures. In this paper, we try to find the relation among two proteins A, B, and a phenotype P, where there is a group of samples *G*, whose expression levels of A and B are both close to one another, and they always have close values of P. In this paper, we propose a method to extract a pair of proteins that effect on a target phenotype, from a dataset that consists of protein expression profiles and phenotype values.

Keywords: Proteomic Analysis, Two-Dimensional Electrophoresis, Phenotype, Expression Profile, Data Mining

1 INTRODUCTION

After the entire human DNA sequence was made public, many post-genome researches started to investigate the systems of living creatures. Proteome analysis is a field of such a post-genome research. The proteome analysis is a research field to analyze comprehensively the entire protein sets, in which functions and interactions of proteins that maintain living creatures are actively investigated.

As a method in proteome analyses, there is a technique called 2D electrophoresis [1]. The 2D electrophoresis enables us to measure expression levels of thousands of proteins in a biological tissue simultaneously. From the protein expression profiles obtained by the technique, we can clarify the functions and the interactions of proteins.

In many researches, major goal of researchers is to identify proteins that effect on a certain phenotype. For this purpose, a method for discovering the relationship between one protein and one phenotype is often used. One of the most basic methods is to calculate the correlation coefficient between protein expression levels of a protein and values of a phenotype item. Relationship between two items can be revealed by a relatively simple statistical method. However, the correlation coefficient evaluates only the liner relationship between two items. In contrast, Qu, et al. proposed a method to discover the nonlinear relationship between a gene and a phenotype using orthogonal polynomials [2].

On the other hand, there are a few researches that try to discover relationships in which more than one proteins effect on one phenotype. Zhang, et al. studied the interaction among a triplet of genes by comparing the correlation coefficients of genes A and B between two cases where another gene C expresses and does not express [3]. As another method, Inoue, et al. developed an algorithm to predict interactions among three proteins A, B and C based on correlation coefficient [4], and Fujiki, et al. developed an algorithm to predict interactions among three proteins A, B and C based on conditional probability [5]. If we regard C as a phenotype, those methods can be used to investigate the relationship between proteins and phenotypes.

In this paper, we propose a new method to detect interactions from different approaches. Specifically, we try to find the relation among two proteins A, B, and a phenotype P, where there is a group of samples *G*, whose expression levels of A and B are close to one another, and they always have close values of P. We evaluate the proposed method by applying the proposed method to the real data set.

Note that, to the best of our knowledge, this study is the first study that tries to find a set of two proteins that effect on a phenotype by finding a group of samples *G* whose expression levels of proteins A and B are close to one another that also have close values of a phenotype P.

The remainder of this paper is organized as follows. In Section 2, we describe the relation among two proteins and a phenotype assumed in this paper. In Section 3, we describe the proposed algorithm in detail. In Section 4, we evaluate our method by applying it to a real protein expression profile and a data set of phenotype. Finally, in Section 5, we conclude our study.

2 THE RELATIONSHIP BETWEEN PROTEINS AND PHENOTYPE WE SUPPOSE

2.1 Phenotype of Creature

Phenotype is a character that a creature has. For example, phenotype is an individual's traits, such as a size of body, a color, a pattern, etc. It is generally said that phenotype is largely determined by genes, but also considerably depends on growth environment of individuals. Many researches try to investigate the system of creatures that determines phenotypes. Such kind of researches are especially valuable when they target on several economically important phenotypes. For example, beef marbling scores and carcass weight of Wagyu beef have direct impact on the economical price of beef.

2.2 Protein Expression Profile

Protein expression levels are the amount of each proteins included in a biological sample. The protein expression levels

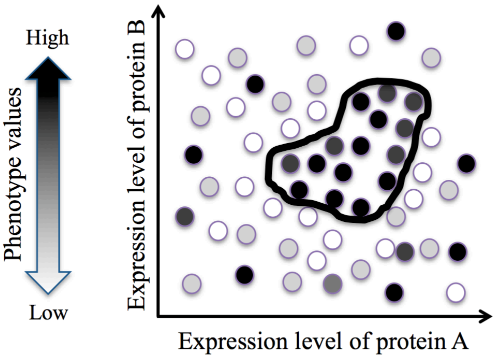

Figure 1: Example of Area That We Want to Demand.

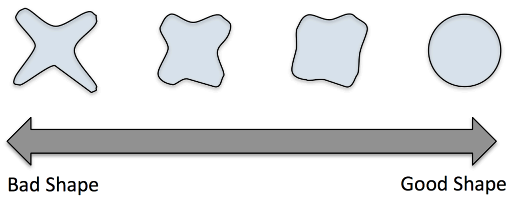

Figure 2: Good or Bad Shape of Area.

are typically measured by the 2D electrophoresis method [1]. This method is used for protein expression measurement widely.

The 2D electrophoresis is a method to separate proteins with 2-dimensions through two steps of electrophoresis. Generally, proteins are separated with isoelectric point in the first dimension, then they further are separated by molecular weight in the second dimension. Typically, the number of proteins included in a profile ranges from several hundred to thousands.

The expression profiles are the data that consists of expression levels of proteins included in a biological sample. The expression profiles are obtained by 2 steps. First, we obtain a 2D electrophoresis image through the 2D electrophoresis experiment. Second, we measure the areas of the islands revealed by the first step using image processing techniques.

2.3 Relationship of Two Proteins and A Phenotype We Suppose

We suppose two proteins that effect on a phenotype. In this paper, we try to find the relation among two proteins A, B, and a phenotype P, where there is a group of samples G, whose expression levels of A and B have close values a and b with each other, and they always have close value p of P.

Figure 1 shows an example of this relationship. We consider a 2-dimensional plane that has two axes of expression levels of proteins A and B. Each sample is plotted in this plane, and the deepness of the color of the samples represents phenotype values (i.e., samples with deep color represent high phenotype values and those with light color represent low phenotype values). Here, if there are no relationship among those two proteins and the phenotype, the distribution of the color of the samples would be uniform, i.e., the samples with various colors are plotted uniformly. In contrast, if some relationships exist, it is thought that the distribution would not be uniform. In this paper, as shown in Fig. 1, we extract the area in which all the samples have close phenotype values. We consider that the existence of such areas indicates the relationship between proteins and phenotype. Namely, by extracting such areas, it is possible to estimate the combination of two proteins and the expression levels that control a phenotype.

Table 1: Data Format of Protein Expression Profiles.

Sample ID	Protein ID			
	A	B	C	⋯
1	0.000582	0.000107	0.000541	⋯
2	0.000563	0.000111	0.000458	⋯
3	0.000495	0.000126	0.000333	⋯
⋯	⋯	⋯	⋯	⋯

Table 2: Data Format of Phenotype Data Set.

Sample ID	Phenotype			
	Beef Marbling Standard	Carcass Weight	Rib-eye Area	⋯
1	4	422.7	44	⋯
2	9	470.7	53	⋯
3	7	433.5	50	⋯
⋯	⋯	⋯	⋯	⋯

3 EXTRACTION METHOD OF AREA WITH CLOSE PHENOTYPE VALUES

3.1 Format of Input Data

We use two sets of input data in the proposed method. One is a protein expression profile and the other is a set of phenotype data. We assume that the protein expression profile is obtained from the 2D electrophoresis experiment. The expression profile consists of the expression levels of each protein contained in each biological sample. We let $i(1 \leq i \leq I)$ be a sample, and let $j(1 \leq j \leq J)$ be a protein. Then, the expression level e_{ij} of a protein j included in a sample i is a real value. We show an example of the expression profile in Table 1.

We assume that a phenotype data set is represented by a table. Then the phenotype data set consists of the real values that represent the degree of phenotype (hereafter, we call them the *phenotype values*). We let $p(1 \leq p \leq P)$ be a phenotype, and the phenotype value p_i of a phenotype p included in a sample i is a real value. We show an example of the phenotype data set in Table 2. This example shows a case of brand cattle, in which we have BMS (Beef Marbling Standard), carcass weight, rib-eye area, etc. as phenotypes.

3.2 Areas That We Wish to Extract

In this paper, we extract a pair of proteins A and B that effect on a target phenotype p, by finding an area in which

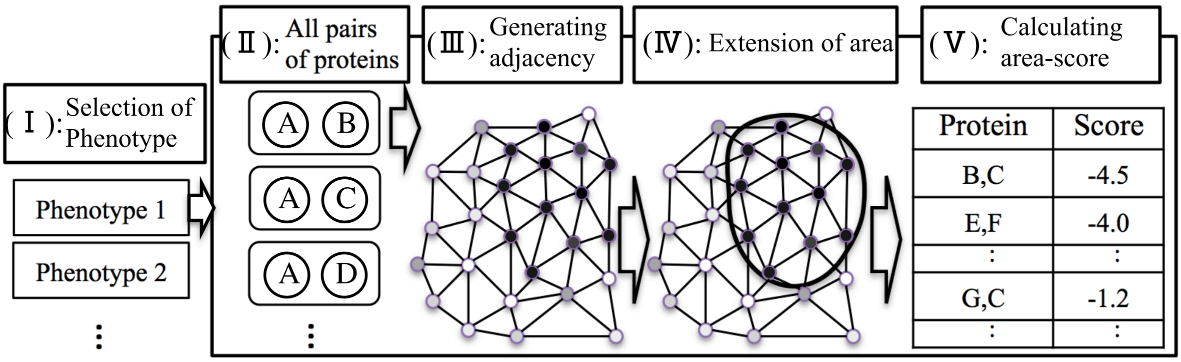
(Ⅰ): Selection of Phenotype
Phenotype 1
Phenotype 2
(Ⅱ): All pairs of proteins
A B
A C
A D
(Ⅲ): Generating adjacency
(Ⅳ): Extension of area
(Ⅴ): Calculating area-score
Protein
Score
B,C
-4.5
E,F
-4.0
G,C
-1.2

Figure 3: Overview of Our Proposed Method.

there is a set of samples whose expression levels of A and B are close to one another that also have close values of p on the 2-dimensional plane. In this section, we describe the criteria that the area should satisfy.

We consider two criteria with which we evaluate areas. Two criteria are on the phenotype values, and on the shape of the area, respectively. First, we describe the criterion on the phenotype values included in the area. It is required that the variance of the phenotype values included in the area is significantly smaller than those of all samples. Namely, it means that the samples that take a narrow range of phenotype values are included in it.

Next, we describe the criterion on the shape of the area. The criterion on shape is that the shape does not have big unevenness on a boundary line, i.e., the shape is “not warped” or “distorted”. Namely, in this paper, we regard the circle as the best shape, whereas we regard the “warped” shape as bad shape. (See Fig. 2). Without this criterion, i.e., if an area is allowed to be any shape, we can extract areas of any distorted shape by choosing arbitrary samples freely that have close phenotype values. By limiting shape of areas, we can evaluate the area properly based on the sample distribution.

3.3 Overview of Proposed Method

We designed an algorithm to find the area that holds the criterion we described in Section 3.2. Note that the problem we treat is a combinatorial optimization problem whose search space is exponentially large so that we can hardly find the optimal solution. Thus, we designed our algorithm as a greedy one that explores areas from a small one by expanding it gradually with the best samples that forms the best areas at that time. We describe the overview of the proposed method as follows. (See Fig. 3 in parallel.)

(a) We select one phenotype to analyze (Fig. 3(I)).

(b) We compose all the possible pairs of proteins for the phenotype selected in step (a) (Fig. 3(II)).

(c) We generate an adjacency graph from the samples on a 2-dimensional plane whose two axes are the expression levels of proteins A and B (Fig. 3(III)). We generate the adjacency graph as the Delaunay graph. We will give a short explanation of the Delaunay graph in the following Sections 3.4.

(d) We repeat extending the area using the graph that is generated in step (c) (Fig. 3(IV)). We start with the area that includes one arbitrarily sample (we call this sample *starting sample*). Then, we repeat extending the area with the most suitable samples until it comes to contain all samples. We perform this process from every starting sample. We describe this extending process in the following Section 3.5.

(e) We calculate the variance of the phenotype values for all the areas throughout the extending process i.e., we calculate the variance every time after extending the area with one sample. Then, for each individual area, we calculate its z-value (we call it the *area-score*) that indicates the statistical probability that the value of the variance occurs (Fig. 3(V)). We extract the areas whose area-score is greater than the threshold. We describe about the calculating area-score in the following Section 3.6.

3.4 Step(c): Generating the Adjacency Graph from Samples

In this Section, we explain the algorithm to generate the adjacency graph from the samples on the 2-dimensional plane.

First, we generate a Voronoi diagram [6] on the 2-dimensional plane. A Voronoi diagram (Fig. 4) is a diagram obtained by dividing space into a number of areas. The boundary lines (dotted lines) between samples are composed of perpendicular bisectors between two samples. The plane is divided into areas (called Voronoi area) corresponding to each sample by the boundary line.

By connecting every pair of samples corresponding to two adjacent Voronoi areas, the Delaunay diagram (Fig. 5) that represents the adjacency among samples is generated. Then, we let $N(i)$ be the sample set adjacent to sample i.

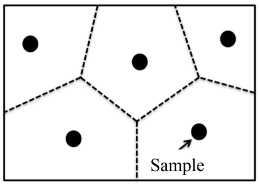

Figure 4: Volonoi Diagram.

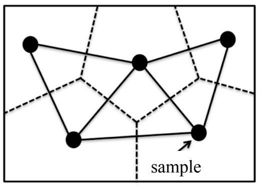

Figure 5: Delaunay Diagram.

3.5 Step(d): Extension of Areas

3.5.1 Overview of Extension Algorithm

We describe an algorithm to extend the area we wish to extract. We show the overview of the process as follows.

(1) An initial area consisting of one sample is determined by selecting a starting sample arbitrarily.

(2) We select a set of *extension candidate samples* from the samples that are adjacent to the current area so as not to make the shape of the extended area distorted.

(3) We select an *extension sample* from the set of *extension candidate samples* and extend the area by adding this extension sample.

(4) If the area does not include all samples, we return to (2).

Our algorithm searches for good areas through the process that expands an area by adding samples one by one greedily. Thus, the result largely depends on selection of starting samples. So, as for (1), the strategy to select starting samples should be determined according to the practical requirements. For instance, if users are interested in retrieving areas in which high-value samples are collected, it is recommended to start with high-value samples. Similarly, users may benefit from starting with low-value or middle-value samples for some cases. The strategy should be determined according to the situation. As for (2) and (3), details are described in the following sections 3.5.2 and 3.5.3, respectively.

3.5.2 Method to Select Extension Candidate Samples

In this Section, we explain the method to select the set of *extension candidate samples* mentioned in Section 3.5.1 (2). We let C be a set of extension candidate samples, and let D be the current area. C is a set of samples that satisfy conditions (i) and (ii) among the samples that is adjacent to D. We prevent extensions from creating donut-shape by setting these conditions as follows.

(i) Candidate sample must be adjacent to more than one samples that are included in D.

(ii) Samples on the boundary of D adjacent to the candidate sample must be continuous on the boundary.

We explain that the area does not become donut-shape using an example. In the area shown in Fig. 6, the samples that are surrounded by a black square are the samples that satisfies condition (i). Among them, the X-marked sample does not satisfy the condition (ii) because the three samples on the boundary line of D adjacent to this X-marked sample are not continuous on the boundary line. If we add this X-marked sample to D, the area that is extended becomes the donut-shape.

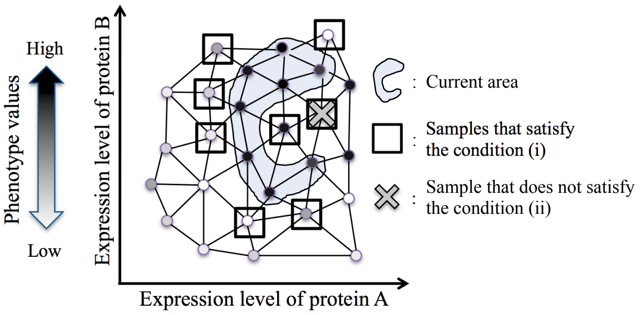
High
Phenotype values
Low
Expression level of protein B
Expression level of protein A
: Current area
: Samples that satisfy the condition (i)
: Sample that does not satisfy the condition (ii)

Figure 6: Conditions of Extension Candidate Sample.

3.5.3 Method to Select Extension Sample

We explain the algorithm to select a *extension sample* mentioned in Section 3.5.1. An extension sample is the sample that is the most desirable to be added to D from a set of extension candidate sample. As described in Section 3.2, the sample that is the most desirable is the sample that satisfies the two criteria on the phenotype values and on the shape of the area, respectively.

We describe the steps to select the extension sample from the extension candidate sample set C. First, we calculate the *shape-cost* $T(x)$ for every *extension candidate sample* $x \in C$. $T(x)$ evaluates the shape of the area that is created by adding the sample x to the current area D. Now, we let D_x be the area that is created by extending D with x. The less $T(x)$ is, the more the shape of D_x is distorted.

Next, we calculate the *phenotype-cost* $Z(x)$ to evaluate the phenotype value of the samples included in D_x. If $Z(x)$ is small, the phenotype value of x take a value close to the samples in D, and x would not increase the variance of the phenotype values of samples included in D. Therefore, we select the extension sample x such that $Z(x)$ is the smallest in C, and satisfy $T(x) \geq T_{thresh}$, where T_{thresh} is the threshold to the $T(x)$. Here, we consider the shape of the area that satisfies $T(x) \geq T_{thresh}$ to ensure the area has a good shape. If there is no x to satisfy $T(x) \geq T_{thresh}$ we regard x such that $Z(x)$ is the smallest in C as the extension sample.

We explain the method to calculate the shape-cost $T(x)$. We calculate $T(x)$ based on the ratio between the boundary length of D_x and the area of D_x. In general, if the area is the same, the boundary length is shorter when the shape is close to circle. We calculate $T(x)$ using this property. First, we let L_x and S_x be the length of the boundary line and the area of D_x, respectively. Here, the radius r_x of the circle whose circumference is just L_x is written as $r_x = \frac{L_x}{2\pi}$. Similarly, the radius r'_x of the circle whose area is just S_x is written as $r'_x = \sqrt{\frac{S_x}{\pi}}$. Finally, We define $T(x)$ as the ratio of r_x and r'_x

as follows:

$$T(x) = \frac{r'_x}{r_x} = \frac{2\sqrt{\pi S_x}}{L_x}.$$

Note that, $T(x)$ takes a value between 0 and 1, and the larger $T(x)$ is, the closer the shape is to a circle.

The phenotype-cost $Z(x)$ evaluates the variance of phenotype values included in D_x, which is created by extending D with x. We calculate $Z(x)$ as the z-value, which is known as a kind of statistic. $Z(x)$ is defined as $Z(x) = \left|\frac{p_x - \mu_x}{\sigma_x}\right|$, where p_x is the phenotype value of x, and μ_x and σ_x are the average and the standard deviation of the phenotype values of samples in D_x. Note that, $Z(x)$ represents the amount of difference between the phenotype value p_x and the average μ_x of phenotype values in D_x, which is measured as the number of the unit value σ_x. If the absolute value of $Z(x)$ is small, it means that p_x is close to the phenotype values of samples in D_x. Namely, if we add such p_x to D, the variance of D would not be increased.

3.6 Step(e) Calculating Area Score

In this Section, we explain the retrieval of the areas we wish to extract.

In this paper, we wish to extract the area with small variance of the phenotype values of the samples in the area. However, in general, if the number of samples in the current area is low, the variance is small. Therefore, in this paper, in order to retrieve the good area under variation of the number of samples, we calculate the variance of every area throughout our process to extend areas, and aggregate the variance of all the areas. Then, we judge whether each area is the one we wish to extract, by calculating the *area-score* of each area as a relative "position" of its variance in the distribution of variances of all areas that includes the same number of samples.

Now, we let $(i_1, i_2)\ \ (1 \leq i_1 < i_2 \leq J)$ be the pair of proteins, and let $n\ \ (1 \leq n \leq I)$ be the number of samples in the area. We suppose the extending process of the area with a starting sample m on the plane whose axes are two proteins (i_1, i_2). Here, we define the area where number of samples in the area is n as $D^{(n)}_{m,(i_1,i_2)}$. Note that, $D^{(n)}_{m,(i_1,i_2)}$ is determined uniquely by n, m and (i_1, i_2). Now, we let p_i be the phenotype value of sample i ($i \in D^{(n)}_{m,(i_1,i_2)}$), and $E[D^{(n)}_{m,(i_1,i_2)}]$ and $V[D^{(n)}_{m,(i_1,i_2)}]$ be the average and the variance of the phenotype values of samples in $D^{(n)}_{m,(i_1,i_2)}$.

We explain the method to calculate the *area-score*. First, we calculate $E[D^{(n)}_{m,(i_1,i_2)}]$ corresponding to combination of n, m and (i_1, i_2) as follows:

$$E[D^{(n)}_{m,(i_1,i_2)}] = \frac{1}{n} \sum_{i \in D^{(n)}_{m,(i_1,i_2)}} p_i.$$

Similarly, we calculate $V[D^{(n)}_{m,(i_1,i_2)}]$ as follows:

$$V[D^{(n)}_{m,(i_1,i_2)}] = \frac{1}{n-1} \sum_{i \in D^{(n)}_{m,(i_1,i_2)}} (p_i - E[D^{(n)}_{m,(i_1,i_2)}])^2.$$

Next, we calculate the average μ_n and the standard deviation σ_n of $V[D^{(n)}_{m,(i_1,i_2)}]$ with all areas whose number of samples in the area is n as follows:

$$\begin{aligned} \mu_n &= \frac{1}{|M| \times J(J-1)/2} \sum_{m \in M} \sum_{1 \leq i_1 < i_2 \leq J} V[D^{(n)}_{m,(i_1,i_2)}], \\ \sigma_n &= \sqrt{\frac{\sum_{m \in M} \sum_{1 \leq i_1 < i_2 \leq J} (V[D^{(n)}_{m,(i_1,i_2)}] - \mu_n)^2}{|M| \times \frac{J(J-1)}{2} - 1}}. \end{aligned}$$

Finally, we calculate the z-value for the variance $V[D^{(n)}_{m,(i_1,i_2)}]$ of each area using μ_n and σ_n as the area-score $R^{(n)}_{m,(i_1,i_2)}$. The area-score $R^{(n)}_{m,(i_1,i_2)}$ is defined as follows:

$$R^{(n)}_{m,(i_1,i_2)} = \frac{V[D^{(n)}_{m,(i_1,i_2)}] - \mu_n}{\sigma_n}.$$

If $R^{(n)}_{m,(i_1,i_2)}$ is small, it means that the area rarely appears statistically. Therefore, we expect the area $D^{(n)}_{m,(i_1,i_2)}$ whose are-score $R^{(n)}_{m,(i_1,i_2)}$ is small enough for the output of the proposed method. For such areas D, we suppose there would be an interaction among two proteins i_1, i_2, and the phenotype.

4 EVALUATION AND DISCUSSION

4.1 Evaluation Method

We evaluate the proposed method by applying it to real protein expression profiles and a phenotype data set obtained by the author's collaborative work in Wakayama [7]. The protein expression profiles that we use in our evaluation are obtained by a 2D electrophoresis-based experiment [8].

A measurement error occurs in the measurement of the protein expression levels. Therefore, we performed 2D electrophoresis twice for each sample to confirm the accuracy of each electrophoresis experiment. From the result of the duplicated measurement, we removed the values considered to be low reliability from expression profiles. Specifically, we measured two expression values for each pair of a protein and a sample. If the larger expression level is larger than 1.3 times the value of the smaller expression level, we consider the expression level for the protein and the sample to be a null value as they are not reliable. Otherwise, the average of the two expression levels is used for each sample-protein pair. As a result, the expression profiles used for our evaluation consist of 90 samples and 47 proteins. In addition, the expression profiles are standardized in advance so that the average and the standard deviation of the expression levels with each sample are 0 and 1, respectively.

We performed an evaluation using "Carcass weight" as an important phenotype among many items included in the phenotype data set of beef cattle. As a pre-processing, we also standardized the phenotype data.

In order to evaluate the performance of the proposed method, we implemented a *simple method* to extend areas to be compared with the proposed method. The simple method is the method that replaces the extension algorithm explained in Sections 3.4 and 3.5. The simple method adds a sample that is close in the Euclidean distance to the start sample m to the current area $D^{(n)}_{m,(i_1,i_2)}$. Consequently, the shape of the area $D^{(n)}_{m,(i_1,i_2)}$ that is obtained by the simple method is nearly a circle centered on the start sample m. Thus, the simple algorithm is equivalent to the algorithm that retrieves the best circular areas in the plane.

We evaluate the performance of the proposed method by comparing it with the simple method by calculating the variance $V[D^{(n)}_{m,(i_1,i_2)}]$ and the average $E[D^{(n)}_{m,(i_1,i_2)}]$.

Here, we describe the parameters in the evaluation experiment. We determined the threshold of the shape-cost $T(x)$ as $T_{thresh} = 0.7$ through a careful preliminary experiments to find a balancing point under the trade-off between the shape-cost and the area-score, and we set the number of samples in the area between 20 and 40 in order to ensure the reliability of the variance of the phenotype values in $D^{(n)}_{m,(i_1,i_2)}$. In addition, as the starting sample m, we use the sample whose phenotype value is within the bottom 10% among all samples. As actual requirements, because it is expected to extract the areas whose samples have low phenotype values, we confirm that the proposed method extracts the area whose phenotype value is low.

4.2 Result and Discussion

Tables 3 and 4 show the results of the ranking of top 10 combinations of proteins with respect to the area-scores. Table 3 is the result of the case where we applied the proposed method to the expression profiles and the phenotype data. On the other hand, Table 4 is the result of the simple method. These tables include the columns of protein ID of proteins A and B, the number of samples in the area, the area-score, $V[D^{(n)}_{m,(i_1,i_2)}]$ and $E[D^{(n)}_{m,(i_1,i_2)}]$. Note that, in Table 3 and Table 4, we leave only the best area out of the same protein pairs.

These results show that both $V[D^{(n)}_{m,(i_1,i_2)}]$ and $E[D^{(n)}_{m,(i_1,i_2)}]$ in the proposed method are smaller than those in the simple method. It was found from the result that the proposed method could extract areas better than the simple method. In order to confirm it in detail, Fig. 7 shows the scatter plots of the ranking of the top 50 areas extracted by the proposed method and the simple method. The vertical axis represents $E[D^{(n)}_{m,(i_1,i_2)}]$ and the horizontal axis represents $V[D^{(n)}_{m,(i_1,i_2)}]$. As is apparent from Fig. 7, both $E[D^{(n)}_{m,(i_1,i_2)}]$ and $V[D^{(n)}_{m,(i_1,i_2)}]$ extracted by the proposed method is found to be lower values than those of the simple method. From these results, we confirmed that the phenotype values of the areas extracted by the proposed method are lower than those extracted by the simple method, and the samples included in the area have close phenotype value each other. In other words, it can be said that the proposed method can extract "good area," compared with the simple method.

Next, we confirm whether the shape of the area extracted by the proposed method is "good shape" or not. As a typical example of the extracted areas, we show the shape of the rank-1 area in Fig. 8.

Table 3: Ranking of Areas with Proposed Method.

Ranking	Protein A	Protein B	Number of samples	Area score	Variance in area	Average in area	Shape score
1	3899	4491	39	-2.7545	0.2510	-0.5539	0.7003
2	5639	5735	31	-2.5615	0.1862	-0.6034	0.7012
3	3648	4491	38	-2.4033	0.3012	-0.4405	0.7057
4	828	5733	36	-2.3852	0.2832	-0.3981	0.7002
5	3648	5727	40	-2.3596	0.3283	-0.3444	0.7010
6	3899	3598	30	-2.3408	0.2153	-0.5549	0.7175
7	4491	5727	29	-2.3014	0.2090	-0.7281	0.7058
8	5636	5654	38	-2.3002	0.3193	-0.4944	0.7001
9	3648	5726	38	-2.2910	0.3209	-0.4495	0.7276
10	4491	5730	40	-2.2879	0.3406	-0.3662	0.7060

Table 4: Ranking of Areas with Simple Method.

Ranking	Protein A	Protein B	Number of samples	Area score	Variance in area	Average in area
1	3648	4491	31	-2.9546	0.3939	-0.3227
2	4491	5657	40	-2.8999	0.5544	-0.2364
3	4491	5688	40	-2.8186	0.5688	-0.2780
4	4491	5686	39	-2.8077	0.5571	-0.2671
5	4491	5721	26	-2.8066	0.3203	-0.4939
6	828	5660	38	-2.8003	0.5436	-0.1725
7	4491	5724	39	-2.6507	0.5856	-0.3966
8	4491	5734	36	-2.6493	0.5437	-0.2875
9	828	4991	40	-2.6394	0.6005	-0.2203
10	5637	5644	25	-2.6194	0.3477	-0.4936

Figure 8 shows the scatter diagram of the rank-1 area in Table 3. The horizontal axis and the vertical axis represent the standardized expression levels of protein A and B, respectively. The shape-cost of the area is 0.7003, which is the value close to threshold $T_{thresh} = 0.7$. We found that this area is close to a circular shape to same extent and is allowable as an area. That is, the shape of this area extracted by the proposed method is "good shape."

Then, we see whether this area is a "good area" or not by examining the phenotype value of the samples in the rank-1 area in Table 3. Figure 9 shows the histogram of the phenotype values included in the area, and a histogram of the phenotype values of all samples. The vertical axis represents the number of samples and the horizontal axis represents the carcass weight. Since the carcass weight has been standardized, the average of the carcass weight of all samples is 0, and the variance is 1. The phenotype values in the extended area are distributed in a relatively narrow range between -1.5 and 0.5, and the distribution is unimodal. We find that $V[D^{(n)}_{m,(i_1,i_2)}] = 0.2510$ is considerably lower than the whole variance 1. Moreover, the $E[D^{(n)}_{m,(i_1,i_2)}] = -0.5539$ is sufficiently smaller than the whole average 0.

From the above reasons, we found that the area extracted by the proposed method is the area that we want to find because both $V[D^{(n)}_{m,(i_1,i_2)}]$ and $V[D^{(n)}_{m,(i_1,i_2)}]$ are small enough.

One of the essential future tasks is to explore how to utilize the proposed method in practice. We do notice that the approach of three-way interactions (i.e., interactions among three proteins, or two proteins and a phenotype) generally

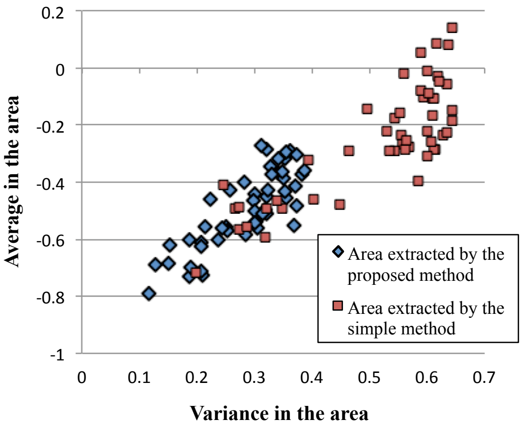

Figure 7: Distribution of Areas Extracted by Proposed Method and Simple Extension Method.

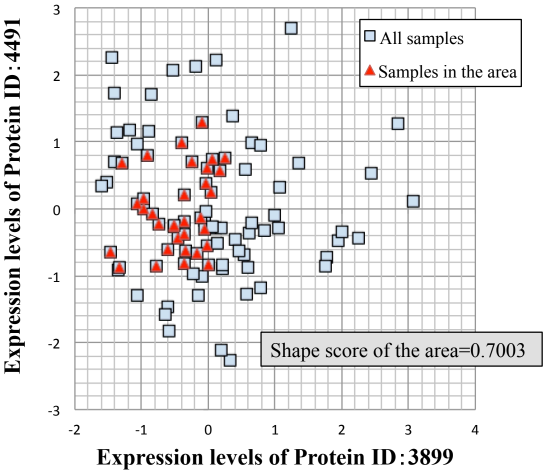

Figure 8: Distribution of Areas of Rank 1 in Table 3.

involves a difficulty in clarifying the performance of methods because specific physical interactions or phenomena are rarely connected directly to them. Thus, to reveal the practical capability of the analysis on three-way interactions, we require to design the processes in which these analytical methods effectively work. This is, in fact, a challenge that requires a considerable deal. For example, we can try to connect some physical interactions or phenomena to our analysis to clarify the direct meaning of our analysis. Also, we can try to show that our analytical result can support to explore biomarkers that control a target phenotype, or accelerate to find proteins included in some pathways or related to some biological functionality. Anyway, these are generally a part of important future work for the approach of three-way interaction analysis.

5 CONCLUSION

In this paper, we proposed a method to find areas with close phenotype values to predict proteins that control phenotypes. By extracting areas including samples with close phenotype values, which rarely occur statistically, it is possible to estimate the relationship among two proteins and a phenotype.

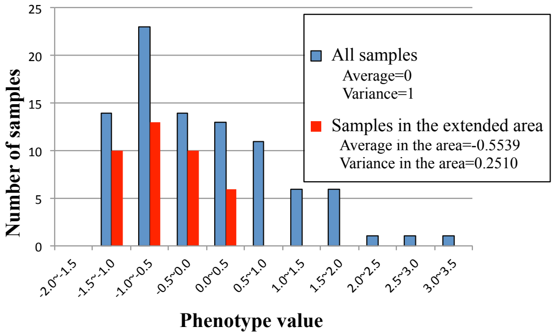

Figure 9: Histogram of Areas of Rank 1 in Table 3.

We performed the evaluation experiment using real data set obtained by the author's collaborative work in Wakayama [7]. In order to evaluate the performance of the proposed method, we implemented a simple method to be compared with the proposed method. As a result, we found that the proposed method extracted the area better than the simple method. That is, the proposed method is able to extract the area that the variance of the phenotype values in the area is small.

ACKNOWLEDGEMENT

This work was partly supported by "the Program for Promotion of Basic and Applied Researches for Innovations in Bio-oriented Industry" of NARO (National Agriculture and Food Research Organization), and "the Program for Promotion of Stockbreeding" of JRA (Japan Racing Association).

REFERENCES

[1] A. Malcolm Campbell, Laurie J. Heyer, Discovering Genomics, Proteomics and Bioinformatics, Benjamin Cummings (2006).

[2] Y Qu, S Xu, "Quantitative trait associated microarray gene expression data analysis," Molecular Biology and Evolution, Vol. 23, No. 8, pp. 1558-1573 (2006).

[3] J. Zhang, Y. Ji, L. Zhang, "Extracting Three-way Gene Interactions from Microarray Data," Bioinformatics, Vol. 23, No. 21, pp. 2903–2909 (2007).

[4] E. Inoue, S. Murakami, T. Fujiki, T. Yoshihiro, A. Takemoto, H. Ikegami, K. Matsumoto, and M. Nakagawa, "Predicting Three-way Interactions of Proteins from Expression Profiles Based on Correlation Coefficient," IPSJ Transactions on Bioinformatics, Vol. 5, pp. 34–43 (2012).

[5] T. Fujiki, E. Inoue, T. Yoshihiro, M. Nakagawa, "Prediction of Combinatorial Protein-Protein Interaction from Expression Data Based on Conditional Probability," In: Protein-Protein Interactions - Computational and Experimental Tools, InTech Web Press, pp. 131–146 (2012).

[6] M. de Berg, O. Cheong, M. van Kreveld, M. Overmars, Computational Geometry: Algorithms and Applications 3rd ed, Springer (2008).

[7] Collaboration of Regional Entities for the Advancement of Technological Excellence in Wakayama, http://www.yarukiouendan.jp/techno/kessyu/

[8] K. Nagai, T. Yoshihiro, E. Inoue, H. Ikegami, Y. Sono, H. Kawaji, N. Kobayashi, T. Matsuhashi, T. Otani, K Morimoto, M. Nakagawa, A. Iritani and K. Matsumoto, "Developing an Integrated Database System for the Large-scale Proteomic Analysis of Japanese Black Cattle," Animal Science Journal, Vol. 79, No. 4 (2008) (in Japanese).

(Received November 22, 2014)
(Revised May 22, 2015)

Takatoshi Fujiki recieved his B.E. and M.E. degrees from Wakayama University in 2010 and 2012, respectively. He is currently a doctoral course student in Wakayama University. He is interested in data mining, machine learning, and bioinformatics. He is a student member of IPSJ.

Empei Gaku received his B.E. and M.E. degrees from Wakayama University in 2011 and 2013, respectively. He is currently working with Azbil Corporation.

Takuya Yoshihiro received his B.E., M.I. and Ph.D. degree from Kyoto University in 1998, 2000 and 2003, respectively. He was an Assistant Professor in Wakayama University in 2003-2009, and from 2009 he is an Associate Professor in Wakayama University. He is interested in computer networks, graph theory, bioinformatics, medical systems, and so on. He is a member of IEEE, IEICE and IPSJ.