IPSJ Transactions on Bioinformatics Vol.5 34–43 (June 2012)

[DOI: 10.2197/ipsjtbio.5.34]

Original Paper

Predicting Three-way Interactions of Proteins from Expression Profiles Based on Correlation Coefficient

Etsuko Inoue[1] Sho Murakami[2,†1] Takatoshi Fujiki[2] Takuya Yoshihiro[1,a)]
Atsushi Takemoto[3] Haruka Ikegami[3] Kazuya Matsumoto[3] Masaru Nakagawa[1]

Received: November 26, 2011, Accepted: March 12, 2012, Released: June 29, 2012

Abstract: In this study, we propose a new method to predict three-way interactions among proteins based on correlation coefficient of protein expression profiles. Although three-way interactions have not been studied well, this kind of interactions are important to understand the system of life. Previous studies reported the three-way interactions that based on switching mechanisms, in which a property or an expression level of a protein switches the mechanism of interactions between other two proteins. In this paper, we proposed a new method to predict three-way interactions based on the model in which *A* and *B* work together to effect on the expression level of *C*. We present the algorithm to predict the combinations of three proteins that have the three-way interaction, and evaluate it using our real proteome data.

Keywords: protein protein interaction, expression profile, three-way interaction

1. Introduction

Interactions among proteins have been regarded as a key issue to understanding the systems of living creatures, because they consist of vast assortment of proteins and their bodies are maintained by the complex interactions among these proteins. Although there is considerable knowledge about the interactions among proteins, it is still not enough to construct a global image of biological activities.

Many studies have been conducted to investigate pairwise interactions between two genes or two proteins. In case of genes, correlation coefficients of the expression levels in microarray expression profiles are often used for this purpose. As for pairwise protein protein interactions (PPIs), many methods have been proposed because a variety of data are available to predict direct interactions of proteins. The most direct approach to tackle PPIs is to identify their evidence of PPIs through *in vitro* or *in vivo* experiments, such as the yeast two-hybrid [1] or tandem affinity purification methods [2]. Pairwise interactions can also be predicted using public databases. Several studies use sequence data such as the method based on conservation of gene neighborhoods [3], the Rosetta Stone method [4], [5], and the sequence-based co-evolution method [6]. Many advanced methods are proposed [7], [8], [9] that utilize public data such as 3D-structures, domains, motifs, pathways, and phylogenetic profiles. These methods and their results are available on the Web [10], [11], [12], [13], [14].

To infer more complex interactions, studies to identify interaction networks from expression data exist, such as the Boolean network model [15], [16] and the Bayesian network model [17]. Note that in many cases, these models treat gene interaction networks, but it is surely possible to treat protein networks. These studies infer a network that representing causal relationships among proteins, including the interactions among more than two proteins. However, the inferred networks include both two-way and more than three-way interactions so that the combinatorial effects that emerge only when the related proteins gather cannot be retrieved separately. Note that this property also appears in the multiple linear regression analysis, which is one of the basic statistical analyses to retrieve the relation among more than three variables. Another drawback of the Boolean and Bayesian network model is that, to infer reliable interaction networks, these methods require large samples of expression data.

Only a few studies have been conducted so far on three-way interactions. Zhang, et al. studied the interaction among a triplet of genes by comparing the correlation coefficients of genes *A* and *B* in two cases, when another gene *C* expresses and when it does not express [18]. Kayano, et al. used expression profiles and genotype data to detect the switching of the correlation sign, i.e., positive and negative correlations, occurred according to the genotype [19]. Their three-way interactions are pure three-way interactions separated from the two-way interaction effect, but they are quite limited because they detect the interaction of two genes based on an interaction that is switched by another binary state property.

In this paper, we present another method to infer three-way in-

[1] Faculty of Systems Engineering, Wakayama University, Wakayama 640–8510, Japan
[2] Graduate School of Systems Engineering, Wakayama University, Wakayama 640–8510, Japan
[3] Graduate School of Biology-Oriented Science and Technology, Kinki University, Kinokawa, Wakayama 649–6493, Japan
[†1] Presently with NTT Data Corporation
[a)] tac@sys.wakayama-u.ac.jp

© 2012 Information Processing Society of Japan

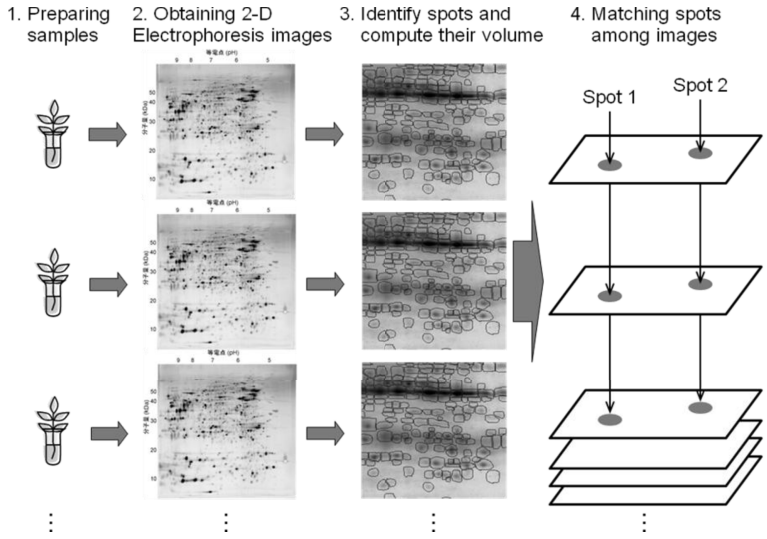


Fig. 1 Process of 2D electrophoresis to obtain expression profiles.

teractions among proteins from expression profiles. Our method is based on the PPI model in which a pair of proteins A and B work together to effect the expression level of C, and the amount of the effect is proportional to the number of A-B pairs that works together with C.

The remainder of this paper is organized as follows. In Section 2, we describe the protein interaction model used in our method and present the basic idea to retrieve the combinatorial effect among the three proteins. In Section 3, we describe the statistical operations to estimate the size of the combinatorial effect. In Section 4, we evaluate our method by applying it to real protein expression profiles, and finally, in Section 5 we conclude our study.

2. Estimating Total Interaction Effect among Three Proteins

2.1 Expression Profile

An expression profile is the data that consists of expression levels e_{ij} of proteins $i \in I$ included in a biological sample $j \in J$, where I is the set of proteins and J is the set of samples. Because we also refer to proteins as A, B, C, and so on, the expression level of protein A in sample j is denoted by e_{Aj}. Expression profiles are frequently used in biological analysis since several high-throughput experiments to obtain expression profiles became popular. For proteins, experiments such as 2D-electrophoresis, protein chips, and mass spectrometry based methods are available. Typically, the number of proteins included in a profile range from several hundreds to thousands. Note that in this study, we apply our method to protein expression profiles because our interaction model supposes a relationship among proteins. However, it is possible to apply our proposed method to gene expression profiles. For genes, the microarray technique is the most popular method of obtaining expression profiles, where thousands to tens of thousands of genes are treated simultaneously. The number of samples is usually several tens, and at most hundreds.

For example, **Fig. 1** illustrates the process of a 2D electrophoresis based experiment [20] from which we obtained the expression profiles used in the evaluation part of this paper. First, we obtain a 2D electrophoresis image from each target sample through biological experimental processes. Second, we identify areas where proteins are separated using image processing software, and we compute the expression level of each spot. Third, we match the spots of the same protein in the images. Finally, we normalize the values of expression levels using a normalization method as a preprocessing step to the data mining that follows. As a result, we obtain a set of protein expression values for each protein in each sample, called expression profiles, as shown in **Fig. 2**.

Sample ID	Protein ID				
	A	B	C	D	…
1	0.003144	0.001562	0.001363	0.000572	…
2	0.005048	0.002316	0.001558	0.000781	…
3	0.00364	0.001842	0.00157	0.000656	…
4	0.005834	0.002258	0.001733	0.000837	…
5	0.005237	0.002325	0.001858	0.000876	…
6	0.001622	0.003075	0.002357	0.000505	…
:	:	:	:	:	

Fig. 2 Input data format.

2.2 Basic Strategy to Predict Interactions

The PPI model that we propose in this study is shown in **Fig. 3**. Three proteins, A, B, and C, would be committed in this model. Proteins of A and B each directly or indirectly interact with C, but if both A and B are expressed together, they have a significantly larger effect on C. In this study, we call the two-way effect the *sole effect*, i.e., the effect of protein A on C and B on C. The three-way effect on C that emerges only when two proteins A and B express together is called as the *combinatorial effect*. Then, we call the composition of the two sole effects and the combinatorial effect as the *total effect*. From our expression dataset, we aim to retrieve the combinatorial effect of A and B, which is not seen if A

© 2012 Information Processing Society of Japan

and B could exist independently. To measure this combinatorial effect, we first estimated the total effect of A and B on C, and then we subtracted the two sole effects of A and B from the total effect.

Our algorithm to estimate the combinatorial effect level is based on the correlation coefficient. The outline of our algorithm is shown in **Fig. 4**. For a triplet of proteins, A, B and C, the following four steps are used. 1) First we compute the two sole effect levels. The sole effect level between two proteins is simply computed as the correlation coefficient between them. We denote the sole effect level from A to C by α, and that from B to C by β, respectively. That is, $\alpha = \mathrm{cor}(A, C)$ and $\beta = \mathrm{cor}(B, C)$, where the function cor denotes the correlation coefficient. 2) Second, we estimate the total effect level t using the algorithm described in Section 2.3. 3) Third, we perform a statistical simulation to compute the total effect level under the two assumptions that the two sole effect levels are α and β, respectively, and no combinatorial effect exists among the triplet. Note that as α and β, we use the sole effect levels obtained from the target triplet A, B, and C in the real data. Through a sufficient number of repetitions, the simulation generates the distribution S. The detail of the simulation is explained in Section 3.1. Because S represents the distribution of total effect levels under no combinatorial effect, the location of t on S shows how rare the computed total effect level t is, and it directly indicates the combinatorial effect level. 4) Fourth, we measure the probability of the value t occurring with respect to S, as the statistical z value. The z value is defined as $z = \frac{(t-u)}{\sigma}$, where μ is the average and σ is the standard deviation of S. This z value is the estimated strength of the combinatorial effect of the target triplet A, B, and C; if z is high, then the combinatorial effect level among them is also high.

To complete our algorithm, in Section 2.3, we show the algorithm to estimate the total effect level t. In Section 3.1 we provide the detailed algorithm of the statistical simulation.

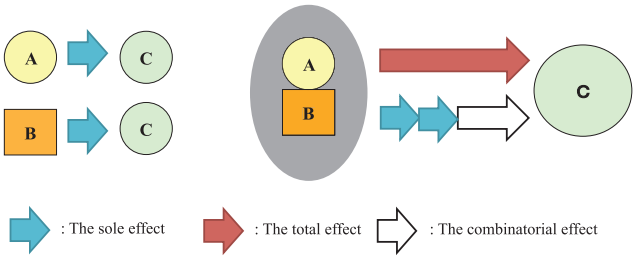


Fig. 3 Interaction model for three proteins A, B, and C.

2.3 Estimating the Total Effect Level t

We estimate the total effect level t by means of the correlation coefficient. According to our protein interaction model, the number of A and B working units indicates the total effect level. If we can assume that the same amount (in expression level) of A and B forms a working unit, we need to consider only $min(e_{Aj}, e_{Bj})$ for the number of working units in sample j. However, this is not correct because the expression level per molecule is different among proteins. Thus, we should find the optimum ratio between the expression levels of A and B, i.e., the point at which they have the largest effect on C. **Figure 5** illustrates this problem. In Fig. 5 (a), the number of working units is not correctly expected because the optimum ratio of A and B is not achieved. As a result, $min(e_{Aj}, e_{Bj})$ and C do not result in a high correlation. However, if the ratio of A and B is optimal, the correlation value is good as shown in Fig. 5 (b), and the number of working units will fit for C, resulting in a high correlation coefficient.

If there is no interaction among A, B, and C, then the correlation coefficient of the working units and C would not be high. To compute the optimum ratio of A and B, we attempt to compute every possible ratio of A and B and choose the best one, i.e., we choose the ratio that provides the highest correlation coefficient.

Now we describe our data mining process to find all possible ratios of A and B. See **Fig. 6** for the range of the ratio. For a ratio k, we define the number of working units in each sample j as $M_{kA,B} = \{e_j^{(min)} | e_j^{(min)} = min(ke_{Aj}, e_{Bj}), j \in J\}$. Then, we compute the correlation coefficient of $M_{kA,B}$ and C for several possible k, where k is a real number. To show an important property, let k_{min} be the minimum ratio of the expression values of A and B among every sample, i.e., $k_{min} = min_{j \in J}(e_{Bj}/e_{Aj})$. Then, note that, if $k \leq k_{min}$, the correlation coefficient between $M_{kA,B}$ and C takes

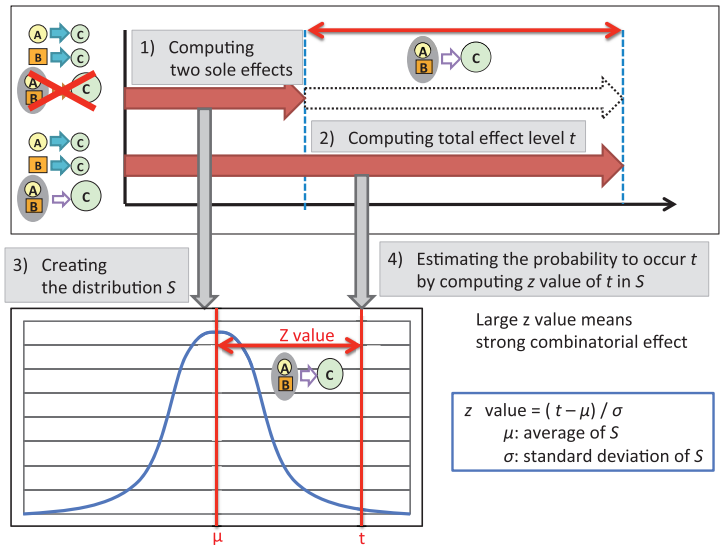


Fig. 4 Outline of algorithm.

© 2012 Information Processing Society of Japan

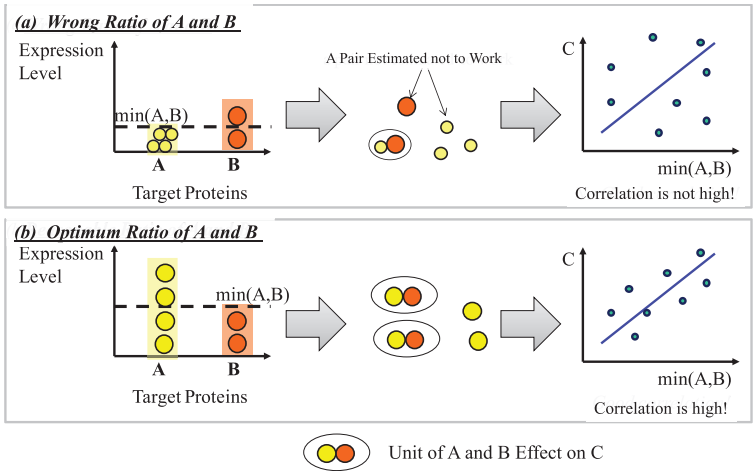


Fig. 5 Optimum ratio between proteins A and B.

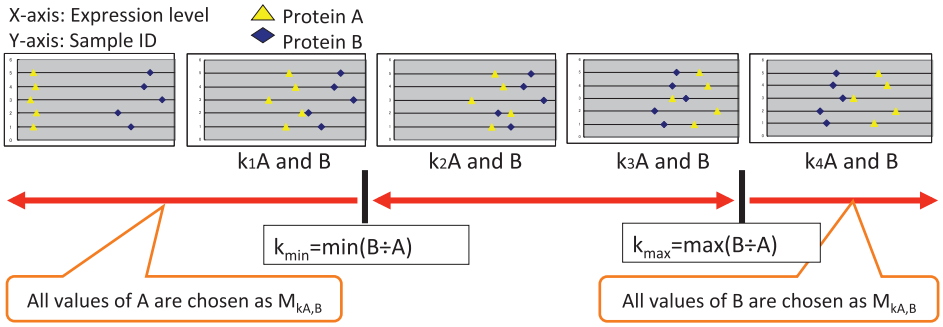


Fig. 6 Range of the ratio to be considered.

the same value because the values of $M_{kA,B}$ are always chosen from A in every sample. Similarly, let k_{max} be the maximum ratio, i.e., $k_{max} = max_{j \in J}(e_{Bj}/e_{Aj})$, then the correlation coefficient always takes the same value if $k \geq k_{max}$. This indicates that we should use values of k between k_{min} and k_{max}. To examine the values of possible correlation coefficients between $M_{kA,B}$ and C, we try every possible value of $k = e_{Aj}/e_{Bj} (j \in J)$ and take the maximum correlation coefficient. If $|J|$ is too large, we can uniformly skip several values of k to reduce the computational load.

In summary, for a protein expression data set including $|I|$ proteins and $|J|$ samples, we compute the correlation coefficients between $M_{k_jA,B}$ and C for every distinct $k_j = e_{Bj}/e_{Aj} (j \in J)$, and find the minimum value. This is the total effect level among A, B, and C denoted by t. Formally,

$$\begin{aligned} \text{Total Effect Level } t &= \max_j \left\{ \mathrm{cor}\left(M_{k_jAB}, C \right) \right\} \\ &= \max_j \left\{ \mathrm{cor}\left(\min\left(k_jA, B \right), C \right) \right\} \\ &= \max_j \left\{ \mathrm{cor}\left(\min\left(\frac{e_{Bj}}{e_{Aj}} A, B \right), C \right) \right\}. \end{aligned}$$

3. Estimating Combinatorial Effects Using Statistical Distribution

3.1 Computing the Distribution S: Total Interaction Level without Combinatorial Effects

In this section, we present the algorithm and the statistical model to compute the distribution S. for a particular combination of proteins A, B, and C, where S is assumed to be the statistical distribution of the total effect levels. This is under the assumption that there is no combinatorial effect among A, B, and C. In other words, we assume only the two sole effects α and β over A, B, and C, and do not consider any other effect among them.

Note that in our simulation, we use the normal distribution for A, B, and C as the most general distribution. Furthermore, as shown in Section 4, a considerable number of proteins follow the normal distribution in the protein expression profiles used in our evaluation.

To meet the above constraints, we first generate the artificial distribution of A, B, and C by generating expression values as random variables following the normal distribution with a common average and a standard deviation. That is, $\mu_A = \mu_B = \mu_C$ and $\sigma_A = \sigma_B = \sigma_C$, where we let μ_A and σ_A be the averages and the standard deviations of protein A, respectively. We discuss the validity of this condition in Section 3.2. In addition, because of the constraint of sole effects, the distributions should hold $\mathrm{cor}(A, C) = \alpha$ and $\mathrm{cor}(B, C) = \beta$. To make the correlation coefficients of A-C α, we repeat the exchange of two expression values of A (i.e., we exchange the expression values of two samples) as long as the correlation coefficient of A-C approaches α. The same step is repeated for B until the correlation coefficient of B-C reaches β.

In this manner, we obtain the random normal distribution of A, B, and C where $\mathrm{cor}(A, C) = \alpha$ and $\mathrm{cor}(B, C) = \beta$. By applying the algorithm presented in Section 2.3 to these artificial distributions, we obtain the total effect level under the assumption of the no combinatorial effect. With a sufficient number of repetitions

© 2012 Information Processing Society of Japan

Correlation Coefficient of B−C (β)	Correlation Coefficient of A−C (α)																			
	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95	1.00
0.05	0.069 0.031	0.101 0.032	0.143 0.032	0.188 0.032	0.233 0.032	0.280 0.032	0.326 0.031	0.373 0.032	0.418 0.031	0.467 0.031	0.514 0.030	0.562 0.030	0.609 0.029	0.657 0.029	0.705 0.028	0.753 0.027	0.800 0.026	0.850 0.024	0.898 0.023	0.955 0.017
0.10	— —	0.122 0.034	0.156 0.034	0.197 0.033	0.240 0.033	0.286 0.032	0.331 0.031	0.379 0.032	0.423 0.030	0.471 0.031	0.518 0.030	0.566 0.030	0.612 0.029	0.660 0.028	0.707 0.027	0.755 0.027	0.802 0.026	0.851 0.024	0.899 0.023	0.955 0.017
0.15	— —	— —	0.178 0.035	0.211 0.035	0.249 0.034	0.294 0.034	0.337 0.032	0.384 0.032	0.428 0.031	0.475 0.031	0.521 0.030	0.569 0.030	0.616 0.029	0.663 0.028	0.710 0.027	0.757 0.026	0.804 0.026	0.852 0.024	0.899 0.023	0.954 0.018
0.20	— —	— —	— —	0.233 0.036	0.263 0.036	0.303 0.035	0.344 0.033	0.390 0.033	0.433 0.031	0.480 0.031	0.525 0.030	0.573 0.030	0.618 0.029	0.665 0.028	0.712 0.027	0.759 0.026	0.804 0.026	0.852 0.024	0.898 0.023	0.953 0.018
0.25	— —	— —	— —	— —	0.285 0.037	0.318 0.037	0.354 0.035	0.397 0.034	0.439 0.032	0.485 0.032	0.529 0.030	0.576 0.030	0.622 0.029	0.668 0.028	0.714 0.027	0.761 0.026	0.806 0.026	0.852 0.024	0.898 0.023	0.951 0.019
0.30	— —	— —	— —	— —	— —	0.341 0.037	0.371 0.036	0.409 0.035	0.447 0.033	0.492 0.032	0.535 0.030	0.581 0.030	0.626 0.028	0.672 0.028	0.717 0.027	0.763 0.026	0.807 0.025	0.853 0.024	0.898 0.023	0.950 0.019
0.35	— —	— —	— —	— —	— —	— —	0.392 0.036	0.423 0.036	0.457 0.033	0.499 0.033	0.541 0.031	0.586 0.030	0.630 0.028	0.675 0.028	0.720 0.027	0.765 0.026	0.809 0.025	0.854 0.024	0.898 0.023	0.948 0.020
0.40	— —	— —	— —	— —	— —	— —	— —	0.446 0.035	0.474 0.034	0.511 0.033	0.550 0.031	0.593 0.030	0.636 0.028	0.681 0.027	0.725 0.026	0.770 0.025	0.812 0.025	0.857 0.023	0.899 0.022	0.948 0.020
0.45	— —	— —	— —	— —	— —	— —	— —	— —	0.493 0.033	0.524 0.033	0.559 0.031	0.600 0.030	0.641 0.028	0.685 0.027	0.728 0.025	0.772 0.025	0.814 0.024	0.858 0.023	0.899 0.022	0.946 0.020
0.50	— —	— —	— —	— —	— —	— —	— —	— —	— —	0.544 0.032	0.573 0.030	0.609 0.030	0.648 0.028	0.689 0.027	0.732 0.026	0.775 0.025	0.816 0.024	0.858 0.024	0.897 0.024	0.942 0.023
0.55	— —	— —	— —	— —	— —	— —	— —	— —	— —	— —	0.593 0.029	0.622 0.029	0.657 0.027	0.696 0.026	0.737 0.025	0.779 0.024	0.819 0.024	0.859 0.023	0.897 0.024	0.939 0.025
0.60	— —	— —	— —	— —	— —	— —	— —	— —	— —	— —	— —	0.642 0.028	0.669 0.027	0.704 0.026	0.743 0.025	0.784 0.024	0.822 0.024	0.861 0.024	0.897 0.025	0.935 0.026
0.65	— —	— —	— —	— —	— —	— —	— —	— —	— —	— —	— —	— —	0.688 0.025	0.716 0.024	0.751 0.024	0.790 0.023	0.827 0.023	0.865 0.023	0.898 0.025	0.933 0.028
0.70	— —	— —	— —	— —	— —	— —	— —	— —	— —	— —	— —	— —	— —	0.733 0.023	0.760 0.022	0.796 0.022	0.832 0.022	0.867 0.022	0.898 0.024	0.927 0.028
0.75	— —	— —	— —	— —	— —	— —	— —	— —	— —	— —	— —	— —	— —	— —	0.778 0.020	0.806 0.020	0.838 0.021	0.871 0.022	0.898 0.025	0.923 0.027
0.80	— —	— —	— —	— —	— —	— —	— —	— —	— —	— —	— —	— —	— —	— —	— —	0.822 0.018	0.846 0.019	0.876 0.020	0.900 0.021	0.924 0.020
0.85	— —	— —	— —	— —	— —	— —	— —	— —	— —	— —	— —	— —	— —	— —	— —	— —	0.860 0.017	0.884 0.018	0.907 0.018	0.933 0.015
0.90	— —	— —	— —	— —	— —	— —	— —	— —	— —	— —	— —	— —	— —	— —	— —	— —	— —	0.900 0.016	0.922 0.015	0.949 0.011
0.95	— —	— —	— —	— —	— —	— —	— —	— —	— —	— —	— —	— —	— —	— —	— —	— —	— —	— —	0.942 0.012	0.969 0.009
1.00	— —	— —	— —	— —	— —	— —	— —	— —	— —	— —	— —	— —	— —	— —	— —	— —	— —	— —	— —	0.996 0.002

Fig. 7 Distribution table of S: Precomputed through computer simulation for a pair of α and β. The upper value is the average, and the lower value is the standard deviation.

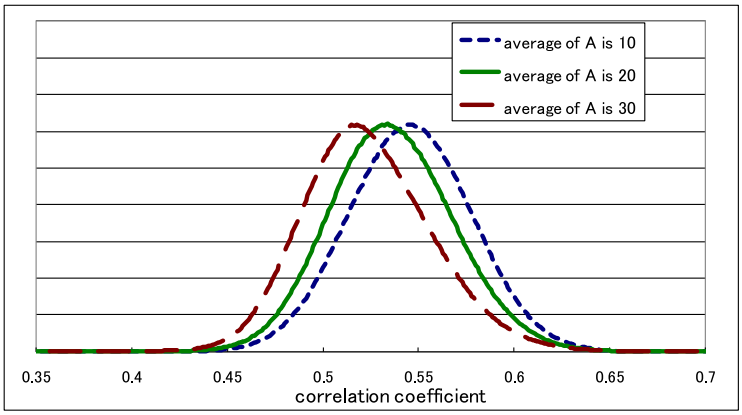


Fig. 8 Distribution S with average variation.

of this process, i.e., distribution generation of A, B, and C, and total effect level computation, we finally obtain the distribution S, which represents the probability of the total effect levels under the assumption of the no combinatorial effect.

Computing the distribution S for every combination of proteins, however, requires considerable computational run time. To reduce the computational time, we prepare, in advance of the computation of total effect levels, the distribution table with the average and the standard deviation of the distribution S for each of the values of α and β. In this study we computed the table with the interval of α and β 0.05, as shown in **Fig. 7**. From the table, we use the values of α and β nearest to the value of the given triplet as the approximate value.

3.2 Discussions of the Distribution S

Now we describe the distribution S varies when the averages and the standard deviations of A, B and C vary, and we conclude that using the common average and the standard deviation in computing the distribution S is appropriate.

We first note that μ_C and σ_C have no effect on the distribution S, because the correlation coefficient is the same even when we add or multiply a constant to all the expression values of C. Therefore, we concentrate on the averages and the standard deviations of A and B.

Without loss of generality, we can fix μ_B and σ_B and vary μ_A and σ_A. **Figure 8** shows the distribution S where σ_A and σ_B are fixed at 1, μ_B at 10, and μ_A varies between 10 and 30. This result is obtained through the computation described in Section 3.1 where α and β are both 0.4, the number of trials is 10,000,000 times. This result clearly shows that the correlation coefficient between $M_{kA,B}$ and C takes a lower value as the difference between μ_A and μ_B increases, and it takes the highest value when $\mu_A = \mu_B$.

Regarding the variation of σ_A, in our method, we select the best correlation coefficient between $M_{kA,B}$ and C among several possible ratios k. This indicates that if σ_A and σ_B differ, such as in the case where $\sigma_B = p\sigma_A$, then the average $\mu_B = p\mu_A$ has the same total effect level as in the case where $\mu_A = \mu_B$ and $\sigma_A = \sigma_B$.

© 2012 Information Processing Society of Japan

(Note that μ_A and σ_A are both multiplied by p when all the expression levels are multiplied by p.) This indicates that the case of $\mu_A = \mu_B$ and $\sigma_A = \sigma_B$ takes the maximum value of total effect levels.

The above discussion shows that the precomputed distribution table, shown in Fig. 7, gives the largest estimated values of S for each α and β. Therefore, in our method, the combinatorial effect cannot be overestimated, i.e., it is always estimated at less than or equal to the true value.

Note that in the simulation, we can use any value of $\mu_A = \mu_B = \mu_C$ and $\sigma_A = \sigma_B = \sigma_C$, because the obtained z values are independent of these values as long as $\mu_A = \mu_B = \mu_C$ and $\sigma_A = \sigma_B = \sigma_C$.

However, the distribution S does not necessarily follow a normal distribution, although the curve of S is similar to the normal distribution curve. However, this does not violate the validity of our algorithm, because the z-value is generally used for a single-peak mountain shape distribution, even if it is not exactly the normal distribution.

4. Evaluation

4.1 Experiment

We evaluate the proposed method by applying it to real protein expression profiles obtained by a 2D electrophoresis-based experiment [20]. We implemented the proposed method in the C++ language. The input data set includes 195 samples and 879 proteins, and the data is processed by global normalization [21] in advance.

Because our method uses the normal distribution for expression levels of A, B, and C, we first confirm whether the expression data follows the normal distribution. For each protein, we omit values that depart from the average by more than 2.5 times the standard deviation as outliers. We apply the Jarque-Bera test [22] to judge whether the expression of each protein follows the normal distribution. The result shows that 454 out of the 879 proteins follow a normal distribution with the significance level of 5%. In the following evaluation, we use these 454 proteins.

To maintain the reliability of the results, we performed several manipulations over the expression data. First, we omitted outliers of expression levels by ignoring values that are greater than 2.5 times the standard deviation from the average. Second, we omitted the combination of A, B, and C if the number of non-null expression values is less than 80% of all the expression values of A, B, and C. Finally, for the scale k that gives the best correlation coefficient between $M_{kA,B}$ and C, if more than 70% of the values in $M_{kA,B}$ are chosen from either kA or B, we exclude the combination of A, B, and C.

4.2 Results

The histogram of the retrieved combinations with a z value of more than 7 is shown in **Fig. 9**. **Figure 10** is the histogram expanded from Fig. 9. **Figure 11** shows the histogram of the case where we assume that there is no combinatorial effect, which is calculated by accumulating the normal distribution trials for the same number of combinations as in the input data. From the comparison of these histograms, the real data has a significantly larger

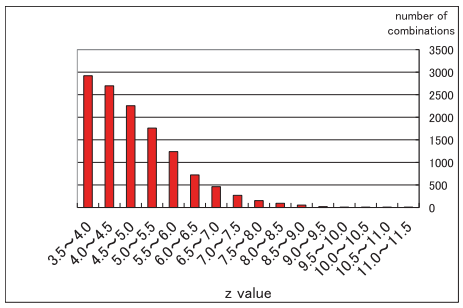


Fig. 9 Histogram of retrieved combinations.

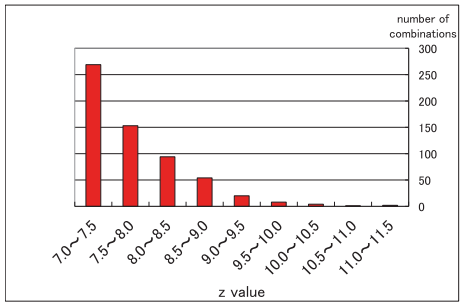


Fig. 10 Histogram of retrieved combinations (expanded).

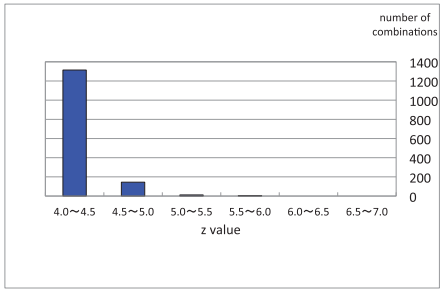


Fig. 11 Histogram of normal distribution.

z value than in the case that assumes no combinatorial effects. These results infer that the real data input includes a significant number of combinations that have the combinatorial effect.

Figure 12 shows the scatter plots of the best-score combination as a typical example. The vertical axis represents the expression level of C, and the horizontal axis represents the expression levels of A, B, and $M_{kA,B}$. In this case, the correlation coefficient of A-C is 0.0449609, B-C is 0.0233452, and $M_{kA,B}$-C is 0.450916. The correlation coefficients of A-C and B-C are quite low, but the value is significantly high for $M_{kA,B}$-C, which results in a very high z value of 12.35. We confirmed that similar to the values in this example, the majority of high z value combinations do not have large correlation coefficients of $M_{kA,B}$-C.

4.3 Effect of Outliers

It is well known that outliers significantly affect correlation coefficients. Because our method is based on correlation coefficients, the results are also significantly affected by outliers. In this section, we show the effect of outliers and the necessity of normality filtering as a preprocess for our algorithm.

We first show a typical example of the outlier effect. **Figure 13** shows the three distributions of a protein combination, i.e., (a)$M_{kA,B} - C$, (b)$A - C$ and (c)$B - C$ distributions of a protein

© 2012 Information Processing Society of Japan

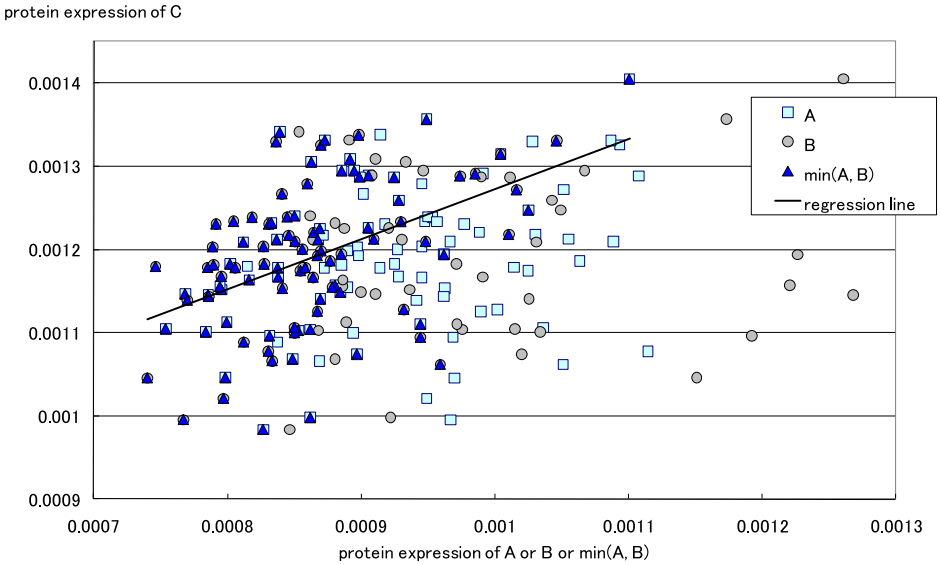
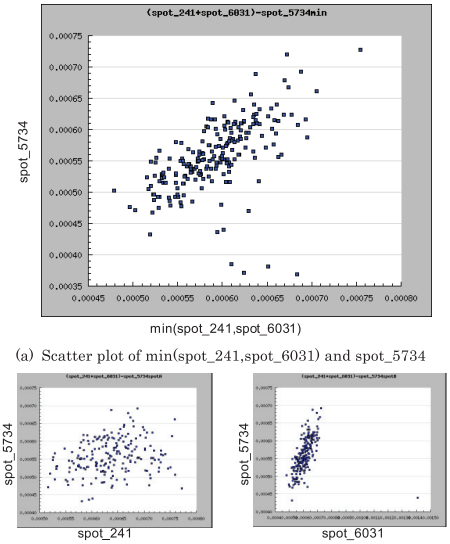


Fig. 12 Distribution of top rank combination.

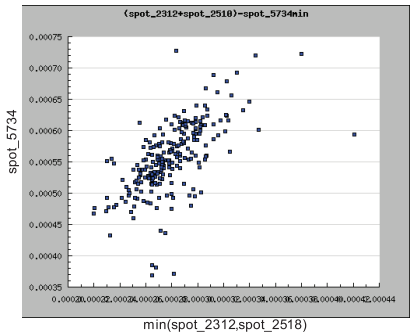


(a) Scatter plot of min(spot_241,spot_6031) and spot_5734

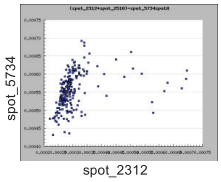
(b) spot_241 and spot_5734

(c) spot_6031 and spot_5734

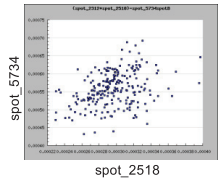
Fig. 13 Type-1 outlier.



(a) Scatter plot of min(spot_2312, spot_2518) and spot_5734



(b) spot_2312 and spot_5734



(c) spot_2518 and spot_5734

Fig. 14 Type-2 outlier.

combination A, B and C. Figure 13 (c) shows a significant outlier that reduces $\beta(= \mathrm{cor}(B, C))$. With this incorrect value of β, the combinatorial effect among them is estimated to be significantly larger than the true value. This type of "false positive" can be excluded by, for example, removing outlier samples that deviate from the average by 2.5σ as we did in the evaluation.

Furthermore, we found another type of a false-positive as illustrated in **Fig. 14**. In the (b)A-C distribution, many samples are assembled on the leftside and several outlier-like samples are sparsely plotted on the rightside. Because these values are not considered to be in $M_{kA,B}$, the combinatorial effect is not appropriately estimated. That is, this type of false positives occur due to their abnormal distributions. In fact, the number of this type of false positives is large; the ratio of false positive distributions are shown in **Fig. 15**. This Figure shows the histogram of z values without applying the normality test filter. Here, we judged whether each high z value combination was a false positive. Although the judgment is done subjectively, it is apparent that significant ratios of the high z value combinations have poor distributions and are judged as false positives. Thus, this type of false positives should be excluded.

For this purpose, in our algorithm, we applied filtering with the normality test. (As described in Section 4.1, we selected 454 out of 879 proteins using the Jarque-Bera test.) As a result, the filtered results shown in Figs. 9 and 10 include very few false positive distributions. From the above results, we concluded that the normality test filter has a significant effect in excluding false positive distributions.

We further note that limiting the range of k is also a method of excluding outlier effects, although the effect is much less than that of the normal distribution test.

4.4 Validation

As a result of applying our proposed method to a real protein expression dataset obtained by a 2D-electrophoresis proteomic

© 2012 Information Processing Society of Japan

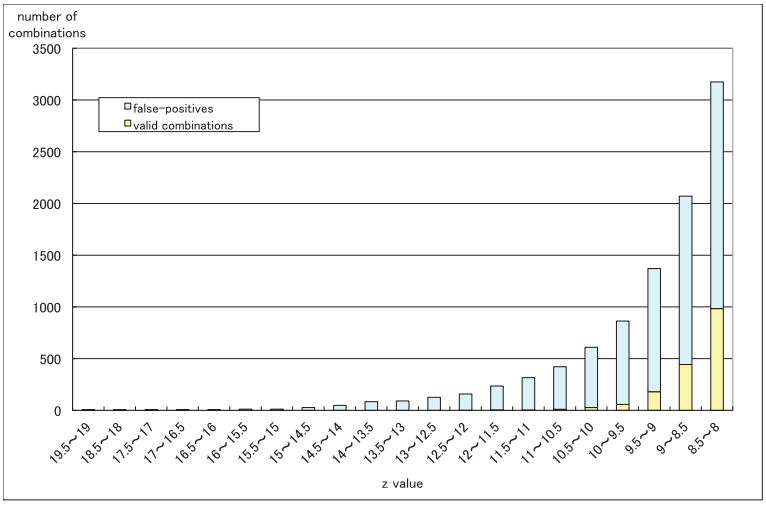


Fig. 15 Histogram without normality test filter.

Table 1 Combination predicting an interaction network as a combinatorial effects.

Z-value	Combination of proteins (A,B, and C proteins)	Correlation coefficient		
		$M_{kA,B}$-C	A-C	B-C
7.669534	A: carbonic anhydrase 2 (CA2) B: vimentin (VIM) C: heat shock 60 kDa protein 1 (HSPD1)	0.306103	0.0189431	0.0439853

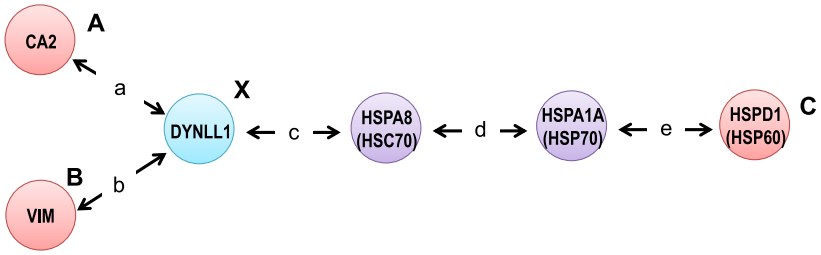


Fig. 16 Building of an interaction network with a combination of proteins (*A*, *B*, and *C*) retrieved from our proposed methods using the base of four available public repositories for PPIs. Carbonic anhydrase 2 (CA2) as protein *A* and vimentin (VIM) as protein *B* have an independent and direct interaction with dynein light chain 8 (DYNLL1) as protein *X*. In addition, DYNLL1 indirectly interacts with heat shock 60 kDa protein 1 (HSPD1) as protein *C* on intervening with heat shock cognate 71 kDa protein (HSPA8) and heat shock protein 70 (HSPA1A). Experimental confirmation methods of PPI were mass spectrometry with (a, b, c, and e)or without co-immunoprecipitation (d).

analysis and cutting off more than 7.0 of z values, we obtained 107 combinations of all three known proteins, which were estimated to show the combinatorial effects. To predict the retrieved combinatorial effects among the three proteins, we validated the interaction network of these proteins by using four available public PPI repositories: the Biomolecular Object Network Databank (BOND) [11], the IntAct database [12], the Molecular INTeractions (MINT) database [13], and the Human Protein Reference Database (HPRD) [14]. First, we assumed that proteins *A* and *B* directly associated with each other and a complex of proteins *A* and *B* directly or indirectly interacted with protein *C*. In this case, in all the databases, there were no data sets that directly detected interaction relationships between proteins *A* and *B*. Hence, we could not find features of the combinatorial effects. Next, we hypothesized that each of the proteins *A* and *B* directly interacts with protein *X* as the fourth protein but *A* and *B* have no direct interaction each other, and protein *X* directly or indirectly interacts with protein *C*. In this situation, we discovered one combination with Dynein light chain 8 (DYNLL1) as protein *X* as a candidate, in which carbonic anhydrase 2 (CA2) as protein *A*, vimentin (VIM) as protein *B*, and heat shock 60 kDa protein 1 (HSPD1) as protein *C* were included as shown as **Table 1**. **Figure 16** shows an interaction network built with these proteins retrieved from our proposed method along with the PPI repositories. The predicted interaction network comprises a total of five proteins, in which PPIs were identified using the yeast two-hybrid system and mass spectrometry with co-immunoprecipitation. Published literature reveals that the identified interaction networks may be involved in the apoptotic pathway [24], [25], [26]. Thus, this result suggests that the retrieved combinatorial effect derived from applying our proposed method to a real protein expression data set can predict a network topology. Furthermore, our proposed method tends to

© 2012 Information Processing Society of Japan

be able to help for deducing an interaction network of proteins that cannot predict in the observed biology.

4.5 Discussion

In this study, we demonstrated an example of predicting an interaction network by applying our proposed method to a proteomic dataset. One of the difficulties in evaluating our method is that we could not know whether other combinations of three known proteins are false positives, because the interactions recorded in all public databases represent only part of the primary literature. From the same reason, it is currently difficult for us to expect known typical three-way interactions to be retrieved with our method. Nevertheless, further investigation with more data sources is expected to confirm the accuracy of our proposed method.

5. Conclusion

In this paper, we proposed a method to retrieve three-way interactions among three proteins by using correlation coefficients. Our method estimates the combinatorial effect level by subtracting two sole effects *A*-*C* and *B*-*C* from the total effect. Because our method uses correlation coefficients, we can predict three-way interactions by using a smaller number of samples campared with Bayesian or Boolean networks.

We applied the proposed method into a real protein-expression data set [20], From the result, we inferred that several hundreds of combinations have the three-way interaction. Note that it is currently difficult to precisely confirm the accuracy of our result because various types of indirect interactions are possible among proteins, and only some of interactions are currently reported in the literature. However, by identifying a combination of three proteins having a combinatorial interaction, we showed the validity of the proposed method in helping to explore protein interactions.

Note that the cells contain various types of protein interaction networks: binary interactions, pathways, complex, and network topology [23]. Analysis of protein interaction networks can uncover unforeseen biological functions of known proteins. Therefore, predicting PPIs by proposing computational models is important for understanding cellular roles of proteins in the cell.

In future, to increase the accuracy and the validity of the proposed method, we plan to identify more combinations of proteins in which three-way interactions are identified.

Acknowledgments This work was partly supported by the Program for Promotion of Basic and Applied Researches for Innovations in Bio-oriented Industry.

References

[1] Fields, S. and Song, O: A novel genetic system to detect protein-protein interactions, *Nature*, Vol.340, pp.245–246 (1989).
[2] Rigaut, G., Shevchenko, A., Rutz, B., Wilm, M., Mann, M. and Séraphin, B.: A generic protein purification method for protein complex characterization and proteome exploration, *Nature Biotechnology*, Vol.17, pp.1030–1032 (1999).
[3] Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G.D. and Maltsev, N.: The use of gene clusters to infer functional coupling, *Proc. Natl. Acad. Sci. USA*, Vol.96, No.6, pp.2896–2901 (1999).
[4] Enright, A.J., Iliopoulos, I., Kyrpides, N.C. and Ouzounis, CA.: Protein interaction maps for complete genomes based on gene fusion events, *Nature*, Vol.402, No.6757, pp.86–90 (1999).
[5] Marcotte, E.M., Pellegrini, M., Ng, H.L., Rice, D.W., Yeates, T.O. and Eisenberg, D.: Detecting protein function and protein-protein interactions from genome sequences, *Science*, Vol.285, No.5428, pp.751–753 (1999).
[6] Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D. and Yeates, T.O.: Assigning protein functions by comparative genome analysis: protein phylogenetic profiles, *Proc. Natl. Acad. Sci. U.S.A*, Vol.96, No.8, pp.4285–4288 (1999).
[7] Pazos, F. and Valencia, A.: In silico two-hybrid system for the selection of physically interacting protein pairs, *Proteins: Structure, Function and Genetics*, Vol.47, No.2, pp.219–227 (2002).
[8] Comeau, S.R., Gatchell, D.W., Vajda, S. and Camacho, C.J.: ClusPro: A Fully Automated Algorithm for Protein-protein Docking, *Nucleic Acids Research*, Vol.32 (Web server issue), pp.W96–99 (2004).
[9] Jothi, R. and Przytycka, T.M.: Computational approaches to predict protein-protein and domain-domain interactions, *Bioinformatics Algorithms: Techniques and Applications*, Mondoiu, I.I. and Zelikovsky, A. (Eds.), pp.465–492, Wiley Press, ISBN 978-047-0097-73-1 (2008).
[10] Tuncbag, N., Kar, G., Keskin, O., Gursoy, A. and Nussinov, R.: A Survey of Available Tools and Web Servers for Analysis of Protein-Protein Interactions and Interfaces, *Briefings in Bioinformatics*, Vol.10, No.3, pp.217–232 (2009).
[11] The Biomolecular Object Network Databank (BOND), available from ⟨http://bond.unleashedinformatics.com/⟩.
[12] the IntAct Database, available from ⟨http://www.ebi.ac.uk/intact/main.xhtml⟩.
[13] The Molecular InTeractions (MINT), available from ⟨http://mint.bio.uniroma2.it/mint/Welcome.do⟩.
[14] The Human Protein Reference Database (HPRD), available from ⟨http://hprd.org/⟩.
[15] Liang, S., Fuhrman, S. and Somogyi, R.: REVEAL, a General Reverse Engineering Algorithm for Inference of Genetic Network Architectures, *Proc. Pacific Symosium on Biocomputing '98*, pp.18–29 (1998).
[16] Shmulevich, I., Dougherty, E.R., Kim, S. and Zhang, W.: Probabilistic Boolean Networks: A Rule-based Uncertainty Model for Gene Regulatory Networks, *Bioinformatics*, Vol.18, No.2, pp.261–274 (2002).
[17] Friedman, N., Linial, M., Nachman, I. and Pe'er, D.: Using Bayesian Networks to Analyze Expression Data, *Journal of Computational Biology*, Vol.7, No.3/4, pp.601–620 (2000).
[18] Zhang, J., Ji, Y. and Zhang, L.: Extracting Three-way Gene Interactions from Microarray Data, *Bioinformatics*, Vol.23, No.21, pp.2903–2909 (2007).
[19] Kayano, M., Takigawa, I., Shiga, M. and Tsuda, K.: Efficiently Finding Genome-wide Three-way Gene Interactions from Transcript- and Genotype-data, *Bioinformatics*, Vol.25, No.21, pp.2735–2743 (2009).
[20] Nagai, K., Yoshihiro, T., Inoue, E., Ikegami, H., Sono, Y., Kawaji, H., Kobayashi, N., Matsuhashi, T., Ohtani, T., Morimoto, K., Nakagawa, M., Iritani, A. and Matsumoto, K.: Developing an Integrated Database System for the Large-scale Proteomic Analysis of Japanese Black Cattle, *Animal Science Journal*, Vol.79, No.4. (in Japanese) (2008).
[21] Lu, C.: Improving the Scaling Normalization for High-density Oligonucleotide GeneChip Expression microarrays, *BMC Bioinformatics*, Vol.5, p.103 (2004).
[22] Jarque, M. and Bera, A.K.: A Test for Normality of Observations and Regression Residuals, *International Statistics Review*, Vol.55, No.2, pp.163–172 (1987).
[23] Ghavidel, A., Cagney, G., and Emili, A.: A Skeleton of the Human Protein Interactome, *Cell*, Vol.122, No.6, pp.830–832 (2005).
[24] Navarro-Lérida, I., Martínez Moreno, M., Roncal, F., Gavilanes, F., Albar, JP. and Rodríguez-Crespo, I.: Proteomic identification of brain proteins that interact with dynein light chain LC8, *Proteomics*, Vol.4, pp.339–346 (2004).
[25] Deighton, RF., Kerr, LE., Short, DM., Allerhand, M., Whittle, IR. and McCulloch, J.: Network generation enhances interpretation of protemic data from induced apoptosis, *Proteomics*, Vol.10, pp.1307–1315 (2010).
[26] Alard, JE., Dueymes, M., Mageed, RA., Saraux, A., Youinou, P. and Jamin, C.: Mitochondrial heat shock protein (HSP) 70 synergizes with HSP 60 in transducing endothelial cell apoptosis induced by anti-HSP60 autoantibody, *FASEB J*, Vol.23, No.8, pp.2772–2779 (2009).

© 2012 Information Processing Society of Japan

Etsuko Inoue received her B.E., M.E. and Ph.D. degrees from Wakayama University in 2002, 2004 and 2007, respectively. She is an Assistant Professor in Wakayama University from 2007. She is interested in database systems, web applications, data visualization, and so on. She is a member of IPSJ.

Sho Murakami received his B.E. and M.E. degrees from Wakayama University in 2008 and 2010, respectively. He is currently working with NTT Data Corporation.

Takatoshi Fujiki received 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.

Takuya Yoshihiro received his B.E., M.I. and Ph.D. degrees from Kyoto University in 1998, 2000 and 2003, respectively. He was an assistant professor in Wakayama University from 2003 to 2009. He has been an associate professor in Wakayama University from 2009. He is currently interested in the graph theory, distributed algorithms, computer networks, medial applications, and bioinformatics, and so on. He is a member of IEEE, IEICE, and IPSJ.

Atsushi Takemoto received his B.E. from Kinki University in 2010. He is currently a master course student in Kinki University. He is interested in proteome analysis and bioinformatics.

Haruka Ikegami received her B.E. from Kinki University in 2004. She was a researcher in Kinki University from 2004 to 2012. She is interested in agriculture proteomics, bioinformatics, mass spectrometry, and so on.

Kazuya Matsumoto received his B.A. degree from Utsunomiya University in 1984, and then M.S. and Ph.D. (Doctor of Agriculture) degrees from Kyoto University in 1986 and 1989, respectively. He was a staff scientist in NT. Science and Tosoh Co. from 1989 to 1995, and a visiting scientist in The Institute of Medical Science, The University of Tokyo from 1995 to 1998. On 1998, he moved as an Assistant Professor to Kinki University. He is a Professor in Kinki University. One of his major research themes is applications of bioinformatics analysis for animal science.

Masaru Nakagawa received his B.E, M.E, and Ph.D. degrees from Osaka University in 1970, 1972, and 1990, respectively. He was a researcher in NTT laboratory from 1972 to 1994, and from 1994 he was a Professor in Kinki University. He is a Professor in Wakayama University from 1994. He is interested in Database Design and Computer System Engineering. He is a member of IPSJ, JSAI and JSIK.

(Communicated by *Kiyoko Kinoshita*)

© 2012 Information Processing Society of Japan