Log linear model and gene networks in DNA microarray data

Hiroyuki Nakahara[1]*, Shin-ichi Nishimura[1,3], Masato Inoue[1,4],
Gen Hori[2], Shun-ichi Amari[1]

1 Lab. for Mathematical Neuroscience, RIKEN Brain Science Institute, Saitama 351-0198,Japan

2 Laboratory for Advanced Brain Signal Processing, RIKEN Brain Science Institute

3 Department of Otolaryngology, Faculty of Medicine, University of Tokyo

4 Department of Otolaryngology, Graduate School of Medicine, Kyoto University

{hiro@, nishi@mns.,minoue@,hori@bsp.,amari@}brain.riken.go.jp

Keywords: Information geometry, graphical log-linear model, DNA microarray data, Wiskott-Aldrich Syndrome Protein (WASP)

April 29, 2002

Abstract

Motivation: Given a vast information of gene expression data, it is critical to develop a simple and reliable method to investigate a fine structure of gene interaction. We show how an information geometric measure help us this investigation.

Results: We introduce an information geometric measure of the log linear model of binary random vector, making connection to graphical models. By representing gene interaction through this model, we can investigate a fine structure of gene interaction. By using specific examples, we provide a detailed description of the method. To validate the method, we also demonstrate that the method can successfully discover biologically-known findings, with respect to Wiskott-Aldrich Syndrome Protein, using a microarray dataset of human tumor cells.

Contact: hiro@brain.riken.go.jp

1 INTRODUCTION

Experiment using DNA microarray chips provides us with a vast amount of information on gene expressions through mRNA transcripts simultaneously. One of central challenges is to discover relationship of gene expression data, or gene network, hidden in data. To infer such a gene network, there have been strong interest in applying graphical models.

Graphical models, including Bayesian networks, are a general framework in statistics and computer science to investigate interaction of random variables (Pearl, 1988; Lauritzen, 1996). We have seen, for example, that hierarchical clustering, a popular method in field, is useful in

*corresponding author; tel: +81-48-467-9663; fax +81-48-467-9693; hiro@brain.riken.go.jp inspecting gene networks (Eisen et al., 1998). Graphical models are in principle more powerful than hierarchical clustering in that the former can treat a finer structure of interactions among variables than the latter because the latter only use the second-order (pairwise) interaction among variables. To investigate a gene network, we need to know not only pairwise but also the third-order (triplewise) and higher-order interactions. As a simplest situation, we may ask whether one gene may co-regulate two other genes or not and this is a question of the third-order (and higher) interaction, not of the second-order. While theories of graphical models have been developed in past decades, also with ongoing current progress (Pearl, 1988; Whittaker, 1990; Lauritzen, 1996), analyses by graphical models have recently proven to be very useful in analyzing gene expression data (Friedman et al., 2000; Pe'er et al., 2001).

The present study somewhat follows their footsteps in gene expression analysis, however, by using different perspective from the information geometry framework (Amari and Nagaoka, 2000) and focusing on a specific simple probability model(Amari, 2001; Nakahara and Amari, 2002; Nakahara and Amari, submitted), namely a log linear probability model of a binary random variable vector (Bishop et al., 1975). Interestingly, this model corresponds, roughly, to graphical log-linear model, called in the graphical model framework (Whittaker, 1990), although the former contains the latter because the latter has more constraints (Whittaker, 1990). Using the information geometry perspective, we can fully utilize the properties of the model and show that this model has a significant merit in analyzing gene expression data.

The present paper is organized as follows. First, we briefly make a connection of an information geometric measure of the log-linear model with graphical models in a general case. Second, due to limited space, we limit ourselves to discuss the measure in the case of three and four variables so that we can provide detailed explanation. While our discussion and demonstration in the present paper mainly treat those cases, we emphasize that the scope of this approach can be easily extended, as already demonstrated in analysis of multiunit neural spike data (Nakahara and Amari, submitted). It is very simple to re-represent microarray data by the information geometric measure, which is one of the merits of our method. The re-representation of the measure already helps us look into gene interaction. We can further quantify the measured values in a simple manner, too. Third, we touch upon our preprocessing before treating real data and show how gene interaction of real-value in microarray data can be represented *through* the log linear model of a binary random vector. Fourth, we show its validity using a microarray dataset of human tumor cells (Khan et al., 2001) and investigate a function of Wiskott-Aldrich Syndrome Protein (WASP) on other genes in phagocytosis. Our method reveals a function of WASP on other genes. Finally, discussion follows.

2 METHODS

2.1 Preliminaries

Here, we make a brief remark on graphical models relevant for the present study, using undirected graphs. A key concept of graphical models is Markov independence, or conditional independence. Graphical models try to estimate dependency of random variables, utilizing graph relation (i.e. nodes and edges), where nodes and edges represent random variables and their dependency, respectively. To measure the dependency, graphical models utilize conditional independence: Each variable X_i is independent from other variables, once the values of its 'parents', which are random variables directly connected to X_i by edges, are fixed. To illustrate this point, let us divide a random variable vector $X^n = (X_1, ..., X_n)$ into three terms, $(X_1, \mathrm{pa}(X_1), \mathrm{npa}(X_1))$, where $\mathrm{pa}(X_1)$ is parents of X_1 and $\mathrm{npa}(X_1)$, which is variables other than X_i and $\mathrm{pa}(X_i)$. We then have conditional independence,

$$P(X_1|X_2, ..., X_n) = P(X_1|\mathrm{pa}(X_1), \mathrm{npa}(X_1)) = P(X_1|\mathrm{pa}(X_1)).$$

When two variables, say X_1 and X_2, are independent, the two variables do not have an edge between them. A simplest example of three variable case is given in Fig 1 A, where two variables (X_1, X_2) are independent conditionally to the other X_3, so that we have

$$P(X_1, X_2|X_3) = P(X_1|X_3)P(X_2|X_3).$$

Thus, edges indicate dependency between random variables (i.e. nodes), based on conditional independence. We show below how conditional independence is translated into information geometric measures, also indicating the merits of using this measure.

In graphical models, in general, there is a process of 'estimating graph': We usually start with assuming a graph structure, then use data to validate the structure, that is, add and/or delete nodes and edges (variables and dependency) and finally obtain estimated graphs. We show below how this validation can be performed in use of the information geometric measure.

2.2 Log linear model and information geometric measure

Here, we briefly mention a most general log linear model of a binary random variable vector. See (Amari, 2001; Nakahara and Amari, submitted) and also (Bishop et al., 1975) for details.

After preprocessing (see Section 2.6), we represent variability of gene expression 'through' the probability distribution of a binary random vector variable. For a moment, let us assume that we obtain a n-dimensional binary random variable vector, $X = (X_1, \cdots, X_n)$; Each X_i represents each gene and becomes 0 or 1, which can be considered to indicate that X_i is not expressed or fully expressed, respectively, in each microarray experiment[1].

Any probability distribution of binary random vector can be exactly expanded by log linear model. Let $p = p(\boldsymbol{x}), \boldsymbol{x} = (x_1, \cdots, x_n), x_i = 0, 1$, be its probability. Each $p(\boldsymbol{x})$ is given by 2^n probabilities

$$p_{i_1 \cdots i_n} = \mathrm{Prob}\left\{X_1 = i_1, \cdots, X_n = i_n\right\}, \;\; i_k = 0, 1, \;\; \text{subject to} \sum_{i_1, \cdots, i_n} p_{i_1 \cdots i_n} = 1$$

and hence, the set of all the probability distributions $\{p(\boldsymbol{x})\}$ forms a $(2^n - 1)$-dimensional manifold $\boldsymbol{S}_n$. One coordinate system of $\boldsymbol{S}_n$ is given by the expectation parameters,

$$\eta_i \;\; = \;\; E\left[x_i\right], \;\; (i = 1, ..n) \;\; \eta_{ij} = E\left[x_i x_j\right], \;\; (i < j), ..., \;\; \eta_{12\cdots n} = E\left[x_i \cdots x_n\right]$$

which have $2^n - 1$ components. This coordinate system is called η-coordinates in $\boldsymbol{S}_n$ (Amari and Nagaoka, 2000). On the other hand, $p(\boldsymbol{x})$ can be exactly expanded by

$$\log p(\boldsymbol{x}) = \sum \theta_i x_i + \sum_{i<j} \theta_{ij} x_i x_j + \sum_{i<j<k} \theta_{ijk} x_i x_j x_k + \cdots + \theta_{1 \cdots n} x_1 \cdots x_n - \psi,$$

[1] In our analysis, we do not actually take a gene expression as binary, as discussed in Section 2.6, 3 later. For presentation simplicity, however, we describe a gene expression as binary till Section 2.6.

where the indices of θ_{ijk}, etc. satisfy $i < j < k$, etc and ψ is a normalization term, corresponding to $-\log p(x_1 = x_2 =, ... = x_n = 0)$. All θ_{ijk}, etc., together have $2^n - 1$ components, forming another coordinate system, called θ-coordinates (Amari and Nagaoka, 2000).

Given gene expression data, both of the above coordinates can be easily estimated in principle. Information geometry assure us that the η-coordinates and θ-coordinates are dually orthogonal coordinates. The properties of the dually orthogonal coordinates remarkably simplifies investigation on dependency of random variables. While details of a general case can be found in (Amari, 2001; Nakahara and Amari, submitted), we show the merits of the dual coordinates, using specific examples below, in the present paper.

In a most general case, a n-dimensional binary random vector results in $2^n - 1$ dimensional coordinates. In microarry data, n may become $\mathcal{O}(10^4)$ so that we would never have enough samples to estimate all coordinates. In practice, hence, we should not use an above full model as but restrict our model, based on domain knowledge and/or by some other approaches (Nakahara and Amari, submitted). Any choice of graph structure, or any dependency, in graphical model framework corresponds to one of restricted models (Whittaker, 1990). This is another reason why we chose to use some specific examples for illustration below.

2.3 Conditional independence in information geometric measure: Three variable case

In the next few sections, we illustrate use of the information geometric measure and its relation to graphical models in case of three and four variables. Remember that a generalization of these examples is possible, which we hope would be self-evident in each example. For three variable case, the log linear model is given by

$$\log p(\boldsymbol{x}) = \sum \theta_i x_i + \sum \theta_{ij} x_i x_j + \theta_{123} x_1 x_2 x_3 - \psi. \tag{1}$$

It is easy to compute all coefficients, i.e. easy to estimate them from gene expression data:

$$\begin{aligned} &\theta_1 = \log \frac{p_{100}}{p_{000}}, \quad \theta_2 = \log \frac{p_{010}}{p_{000}}, \quad \theta_3 = \log \frac{p_{001}}{p_{000}}, \theta_{12} = \log \frac{p_{110} p_{000}}{p_{100} p_{010}}, \quad \theta_{23} = \log \frac{p_{011} p_{000}}{p_{010} p_{001}}, \\ &\theta_{13} = \log \frac{p_{101} p_{000}}{p_{100} p_{001}}, \theta_{123} = \log \frac{p_{111} p_{100} p_{010} p_{001}}{p_{110} p_{101} p_{011} p_{000}}, \quad \psi = -\log p_{000}. \end{aligned} \tag{2}$$

Any distribution of the three binary random variable, or any type of dependency of them, can be represented by this θ-coordinates, which we denote by

$$\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3; \theta_4, \theta_5, \theta_6; \theta_7) = (\theta_1, \theta_2, \theta_3; \theta_{12}, \theta_{13}, \theta_{23}; \theta_{123}).$$

Let us go back to an example shown in Fig 1 A, where the three variables (now binary ones) have a relation $P(X_1, X_2|X_3) = P(X_1|X_3)P(X_2|X_3)$. This relation can be represented in terms of θ-coordinates as follows.

Theorem 1.

$$\begin{aligned} P(X_1, X_2|X_3 = 0) = P(X_1|X_3 = 0)P(X_2|X_3 = 0) &\iff \theta_{12} = 0 \\ P(X_1, X_2|X_3 = 1) = P(X_1|X_3 = 1)P(X_2|X_3 = 1) &\iff \theta_{12} + \theta_{123} = 0 \end{aligned}$$

Obviously, by this theorem, we realize

$$P(X_1, X_2|X_3) = P(X_1|X_3)P(X_1|X_3) \iff \theta_{12} = \theta_{123} = 0. \tag{3}$$

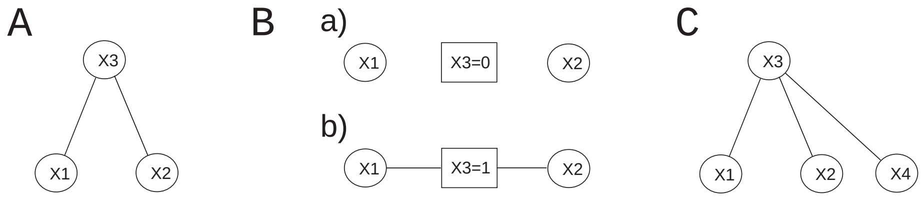

Figure 1: **A** Example of three variables **B** When $\theta_{12} = \theta_{123} = 0$ a. $X_3 = 0$ b. $X_3 = 1$. A rectangle is to indicate that a value of the random variable is fixed (as for X_3), while a circle is to indicate a random variable. **C** Example of four variables

Hence, $\theta_{12} = \theta_{123} = 0$ is equivalent to the relation shown in Fig 1 A[2].

Theorem 1, however, tells us more than Eq 3 does. Theorem 1 indicates that by inspecting the measurement of $\theta_{12}, \theta_{123}$ from gene expression data, we can directly infer a gene interaction. There is a distinctive meaning of θ_{12} and θ_{123}. Recall that $X_3 = 0$ indicates that the gene X_3 was not expressed, whereas $X_3 = 1$ indicates that the gene X_3 was expressed (also see Preprocessing section). Hence, if we find $\theta_{12} = \theta_{123} = 0$ in data, we see that two genes, X_1 and X_2, are independent (conditionally to X_3), regardless of whether $X_3 = 0$ or $= 1$. On the other hand, for example, if we find $\theta_{12} = 0$ but $\theta_{123} \neq 0$, two genes, X_1 and X_2, are (conditionally to X_3) independent only when the gene X_3 is not expressed (Fig 1 B a) but, more importantly in gene data analysis, that X_1 and X_2 become dependent when the gene X_3 is expressed (Fig 1 B b). This is very useful information, for example, to see an effect of the gene X_3 on the other two genes. Furthermore, the sign of θ_{123}, positive or negative, indicates that the gene X_3 induces positive or negative, respectively, interaction between genes X_1 and X_2. Note that analysis of this type (i.e., $\theta_{12} = 0, \theta_{123} \neq 0$) is not usually treated in graphical models because such a type violates correspondence of graph relation with conditional independence in graphical model framework (Whittaker, 1990).

2.4 Quantifying estimates: three variable case

Next, we address how to quantify our estimates of θ-coordinates. For example, is $\theta_{123} = 0.001$ significantly different from zero or not? How about $\theta_{123} = 0.1$ or 10? This question corresponds to 'validation of a graph structure' in graphical models and is an important question in the information geometric measure as well. In statistical estimation of parameters for a probability model of multiple random variables, in general, we have to take care of dependency between random variables, because their dependency may lead to correlated estimation errors, for example, which is why graphical models need to 'propagate beliefs' (Lauritzen, 1996). With such a care, we need to quantify significance of its estimated values.

Dual orthogonality of θ- and η-coordinates allows us to treat these issues in a fairly simple manner. Let us first write η-coordinates (whose definition is similar to those in Preliminaries section) in this three variable case as follows:

$$\boldsymbol{\eta} = (\eta_1, \eta_2, \eta_3; \eta_4, \eta_5, \eta_6; \eta_7) = (\eta_1, \eta_2, \eta_3; \eta_{12}, \eta_{13}, \eta_{23}; \eta_{123}).$$

Note that estimation of these coefficients is very easy. Estimation errors of the dual coordinates are orthogonal to each other, in other words, we can estimate any subset of θ-coordinates

[2] We assume $\theta_{13} \neq 0, \theta_{23} \neq 0$ so that there are edges between X_1 and X_3, between X_2 and X_3 in Fig 1 A

independent from the complementary subset of η-coordinates. Second, quantification of estimated values can be easily done via the Fisher information matrix of the mixed coordinates and due to the dual orthogonality, this Fisher information matrix has a very simple structure, resulting in easy procedure of quantification (Nakahara and Amari, submitted).

To use the dual orthogonality of the θ- and η-coordinates, we now introduce the mixed coordinates as follows:

$$\boldsymbol{\zeta}_k = (\zeta_1, ..., \zeta_7) = (\eta_1, ..., \eta_{k-1}, \theta_k, \eta_{k+1} ..., \eta_7)$$

Each of these mixed coordinates ($k = 1, ..., 7$) is useful to single out different dependency. Here, we only explain the case of $\boldsymbol{\zeta}_7$ in detail (also see the end of this section). $\boldsymbol{\zeta}_7$ is useful to single out the triplewise dependency/interaction i.e., $\theta_7 = \theta_{123}$, regardless of the first-order and the second-order modulation of variables, which are represented by $(\eta_1, ...\eta_6)$.

Suppose our estimate of expression data is given by $\hat{\boldsymbol{\zeta}}_7 = (\hat{\eta}_1.,,.\hat{\eta}_6, \hat{\theta}_7)$ and suppose we like to investigate whether θ_7 is different from zero. Then, our null hypothesis, denoted by $\boldsymbol{\zeta}_7^0$, is given by $\boldsymbol{\zeta}_7^0 = (\hat{\eta}_1.,,.\hat{\eta}_6, \theta_7^0)$, where $\theta_7^0 = 0$. Difference in ζ_7, i.e., $\hat{\theta}_{123} - \theta_{123}^0 = \hat{\theta}_7$, by itself, does not attain a nature of distance. This is because the metric varies at each probability distribution $p(\boldsymbol{x}; \boldsymbol{\zeta}_7)$ (Amari and Nagaoka, 2000). We can, however, quantify the discrepancy between $p(\boldsymbol{x}; \boldsymbol{\zeta}_7^0)$ (null hypothesis) and $p(\boldsymbol{x}; \hat{\boldsymbol{\zeta}}_2)$ (our estimate) by KL divergence,

$$D(\boldsymbol{\zeta}_7^0; \hat{\boldsymbol{\zeta}}_2) = \sum_{\boldsymbol{x}} p(\boldsymbol{x}; \boldsymbol{\zeta}_7^0) \log \frac{p(\boldsymbol{x}; \boldsymbol{\zeta}_7^0)}{p(\boldsymbol{x}; \hat{\boldsymbol{\zeta}})}.$$

The KL divergence works as a quasi-distance. We have, asymptotically[3],

$$2ND(\boldsymbol{\zeta}_7^0; \hat{\boldsymbol{\zeta}}_2) \approx I_{77}(\boldsymbol{\zeta}_7^0)(\hat{\theta}_{123} - \theta_{123}^0)^2 = I_{77}(\boldsymbol{\zeta}_7^0)(\hat{\theta}_{123})^2 = I_{77}(\boldsymbol{\zeta}_7^0)(\hat{\theta}_7)^2 \equiv \lambda_7$$

where N is the number of samples. The right hand side is a quadratic approximation of the left hand side. $I(\boldsymbol{\zeta}_7^0)$ indicates the Fisher information matrix, with dimension 7 × 7, of the mixed coordinates at the point $\boldsymbol{\zeta}_7^0$ in the probability space. Generally speaking, with this quadratic approximation, we need to use a general Riemannian metric, i.e., of the form $\sum_{i,j=1}^7 g_{ij}\zeta_i\zeta_j$. In our case, however, because of dual orthogonality, we only need to use a single component, $I_{77}(\boldsymbol{\zeta}_7^0)$ in $I(\boldsymbol{\zeta}_7^0) = (I_{ij}(\boldsymbol{\zeta}_7^0))$ $(i, j = 1, ..., 7)$ and furthermore, have a simple formula of $I_{77}(\boldsymbol{\zeta}_7^0)$ (Nakahara and Amari, submitted). Hence, we can easily quantify the estimation of θ_{123}. We can then follow the formulation of the likelihood ratio test and even get p-value by $\lambda_7 \sim \chi^2(1)$, where $\chi^2(k)$ denotes χ^2 probability distribution with a degree of freedom, k.

In a similar manner, we can inspect any of θ_k ($k = 1, ..., 7$) by using $\boldsymbol{\zeta}_k$. Furthermore, when we like to inspect two θ parameters together, say θ_4 and θ_7 together[4], we can formulate the mixed coordinates such as $\boldsymbol{\zeta}_{47} = (\eta_1, \eta_2, \eta_3, \theta_4, \eta_5, \eta_6, \theta_7)$. Then, we can again use the dual orthogonal property to single out the two term together independent from other terms. Similarly, we can single out any subset of $\boldsymbol{\theta}$, so that we can easily inspect gene network structure and single out interesting dependency.

[3] Below, we similarly use notation λ_k, corresponding to $\boldsymbol{\zeta}_k$ (e.g. Fig 3)

[4] For example, this may be interesting in inspecting $\theta_4 + \theta_7 = \theta_{12} + \theta_{123} = 0$ in relation to Theorem 1

2.5 Conditional independence: Four variable case

Here, we briefly illustrate a case of four variables, hoping to provide a sense of a general scope of our approach. The log linear model now becomes

$$\log p(\boldsymbol{x}) = \sum \theta_i x_i + \sum \theta_{ij} x_i x_j + \sum \theta_{ijk} x_i x_j x_k + \theta_{1234} x_1 x_2 x_3 x_4 - \psi. \tag{4}$$

We use an example shown in Fig 1 C. First we have

$$\begin{aligned} & P(X_1, X_2, X_4 | X_3) = P(X_1|X_3)P(X_2|X_3)P(X_4|X_3) \\ \iff & P(X_1, X_2 | X_3, X_4) = P(X_1, X_2|X_3) \ \text{ and } \ P(X_1, X_2|X_3) = P(X_1|X_3)P(X_2|X_3) \end{aligned}$$

We can relate the above quantities to θ-coordinates by inspecting each term in the above right hand side identity. Note the following relation:

$$\begin{aligned} P(X_1, X_2|X_3 = 0, X_4) = P(X_1, X_2|X_3 = 0) \ &\iff \ \theta_{14} = \theta_{24} = \theta_{124} = 0 \\ P(X_1, X_2|X_3 = 1, X_4) = P(X_1, X_2|X_3 = 1) \ &\iff \ \theta_{14} + \theta_{134} = \theta_{24} + \theta_{234} = \theta_{124} + \theta_{1234} = 0 \end{aligned}$$

Once these conditions are satisfied, we have, similarly to Theorem 1:
Theorem 2.

$$\begin{aligned} & P(X_1, X_2, X_4|X_3 = 0) = P(X_1|X_3 = 0)P(X_2|X_3 = 0)P(X_4|X_3 = 0) \\ \iff & \quad \theta_{14} = \theta_{24} = \theta_{124} = 0, \ \ \theta_{12} = 0 \\ & P(X_1, X_2 X_4|X_3 = 1) = P(X_1|X_3 = 1)P(X_2|X_3 = 1)P(X_4|X_3 = 1) \\ \iff & \quad \theta_{14} + \theta_{134} = \theta_{24} + \theta_{234} = \theta_{124} + \theta_{1234} = 0, \ \ \theta_{12} + \theta_{123} = 0 \end{aligned}$$

The above relation tells us, in terms of θ-coordinates, more than the relation of $P(X_1, X_2, X_4|X_3) = P(X_1|X_3)P(X_2|X_3)P(X_4|X_3)$, similarly to discussion in Section 2.3. Thus, we can again infer details of gene network by inspecting θ-coordinates estimated from data. We may infer recursively a larger size of gene network in a similar manner, although it is certainly constrained with the limitation of number of samples in general. We also only mention that quantification of estimated value can be dome in a similar manner to Section 2.4.

Let us remark on the relation between three and four variable cases, using example of $\theta^3_{123} = \log \frac{p_{111}p_{100}p_{010}p_{001}}{p_{110}p_{101}p_{011}p_{000}}$, which is of the three variable (Eq 2), and of θ^4_{123} and θ^4_{1234}, which are from the four variable case (Eq 4). We can easily compute, for example, $\theta^4_{123} = \log \frac{p_{1110}p_{1000}p_{0100}p_{0010}}{p_{1100}p_{1010}p_{0110}p_{0000}}$. Notably, we have $\theta^3_{123} \neq \theta^4_{123}$. This is because the three variable model ignores the fourth variable influence on the three variables. Next, by inspection, we get a relation $\theta^4_{1234} + \theta^4_{123} = \theta^4_{123|4}$, where we defined $\theta^4_{123|4} = \log \frac{p_{1111}p_{1001}p_{0101}p_{0011}}{p_{1101}p_{1011}p_{0111}p_{0001}}$. By comparing θ^3_{123} and $\theta^4_{123|4}$, we see that the two terms together, $\theta^4_{1234} + \theta^4_{123}$, indicates the third-order interaction of (X_1, X_2, X_3) conditional to $X_4 = 1$, whereas θ^3_{123} indicates this interaction regardless of the value of X_4.

2.6 Preprocessing

Here we discuss our preprocessing procedure, which takes two steps. First, each sample data is normalized by multiplying because each experiment has been affected by differences in various factors (e.g. different amounts of initial mRNA) and these factors are considered to have, roughly, proportional effects. After this normalization, we discard the genes whose expression of all samples are very low, i.e., less than a certain threshold, which corresponds to presumption that such genes have no expression in all samples and hence, are of no interest. In fact, this step is already taken in the data used in the present study.

Provided the above procedure of obtaining data by itself, we consider the gene expression data as indicating only the relative degree of expression, using real values, compared with a process of an original cell (or original biological tissue). It is then more visible to assign 0 to the null expression and 1 to the full degree of expression to quantify the relative degree of expression. This is the second step: We re-represent the degree of gene expression as in $[0, 1]$ and then represent interaction of genes 'through' a log linear model of a binary random vector. In other words, *we do not regard gene expression by itself as binary.* We only use the log linear model as a means to represent interaction of genes. This point is important because, for example, some researches are concerned with whether binary Boolean network could be enough to fully represent gene interaction (Akutsu et al., 2000).

How then can we convert a real value of each gene into a value bounded in $[0, 1]$? There can be several ways of this quantification. One way is to use order statistics. The other, used in the present study, is to normalize those relative degree of expression. Now let us denote values of gene expression of i-th gene by $(x_{i1}, x_{i2}, ..., x_{iN})$, where each x_{ij} has a real value and N is the number of samples. For this purpose, we obtain the mean and variance of $(x_{i1}, x_{i2}, ..., x_{iN})$, denoted by μ_i and σ_i^2, and convert x_i to z_i by $z_i = \frac{1}{\sqrt{2\pi\sigma_i^2}} \int_{-\infty}^{x_i} \exp\{-(t-\mu_i)^2/2\sigma_i^2\}dt$. Now each of $(z_{i1}, z_{i2}, ..., z_{iN})$ expresses a relative degree, bounded by $[0, 1]$. Using these z_is, for example, p_{100} is computed by

$$p_{100} = \frac{1}{N}\sum_{j}^{N} z_{1j}(1 - z_{2j})(1 - z_{3j}).$$

In a similar manner, all p_{ijk} (and any $p_{i_1 i_2 \ldots i_n}$) can be computed, so that it is easy to obtain both θ- and η-coordinates after this preprocessing.

3 RESULTS

The present study investigated a microarray dataset of human tumor cells (Khan et al., 2001), available from http://www.nhgri.nih.gov/DIR/Microarray/Supplement/. The data set[5] contains 88 microarray experiments of 2,308 genes. We focus on analyzing the gene network associated with Wiskott-Aldrich Syndrome Protein (WASP). Wiskott-Aldrich syndrome, characterized by thrombocytopenia, eczema and immunodeficiency, results from mutation in WAS gene. The product of WAS gene, i.e. WASP, plays a key role in phagocytosis (see below).

We first took all possible combination of three genes, one of which is fixed as WASP and we denote each set of three genes by $(X1, X2, X3) = (X, Y, \text{WAS})$. After preprocessing described above, we estimated their interaction in terms of θ-coordinates and hence, got seven-dimensional vector $\boldsymbol{\theta}$ for each set of three genes. Here we mostly inspect θ_{12} and θ_{123}.[6] These two θ components could reveal the effect of WAS expression on other two genes (Fig 1 B).

Figure 2 shows the results of estimation. To help grasp nature of WAS (Fig 2 A), we also plotted the values by taking X_3 = COL5A1 gene from the same data set (Fig 2 B). COL5A1

[5] Gene names were abbreviated according to UniGene clusters in the following description.

[6] Indeed, it is necessary to investigate all components to fully determine the gene structure of each set of genes but due to limited space, we cannot discuss all of them. Still, we make some brief comments on other components when we discuss specific examples of the set later.

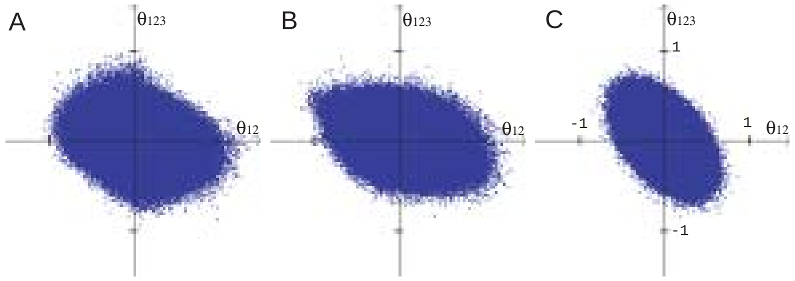

Figure 2: Distribution of $(\theta_{12}, \theta_{123})$. **A** Human tumor cell data (Khan 2001 et al), X_3 =WAS. **B** X_3 =COL5A1. **C** Independently-random data.

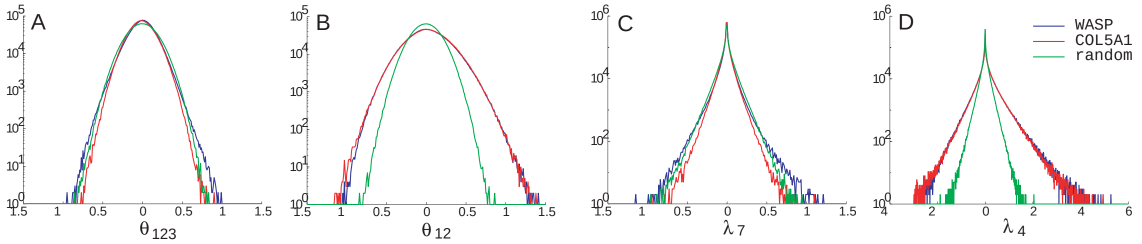

Figure 3: Histograms of the number of genes according to **A** θ_{123} **B** θ_{12} **C** λ_7 **D** λ_4 (bin: 0.01)

gene is biologically less-interacting because it is for structural protein (collagen fiber), not for regulatory protein. Furthermore, we also generated random data, using independent Gaussian distributions, and plotted the estimated values (Fig 2 C).

It is clear, even by simply looking at the shape of distributions, that both of WAS and COL5A1 (Fig 2 AB) have interaction different from a completely independent one (Fig 2 C), while WAS still has interaction slightly different from COL5A1 (Fig 2 B). This observation becomes more clear when we plot θ_4 (θ_{12}) and θ_7 (θ_{123}) separately in histogram (Fig 3 AB) and their quantified values, λ_4 and λ_7 separately (Fig 3 CD). WAS and COL5A1 have almost the same second-order interaction (as a distribution), yet different from a random one (Fig 3 BD). WAS has the third-order interaction different from both COL5A1 and a random one, while these two have almost the same third-order interaction (Fig 3 AC).

To demonstrate the utility of the measure, we chose to investigate the set of genes that has $\theta_{12} \approx 0$ but a large θ_{123} in the above distribution (see Section 2.3); First, we collected the set of genes that has λ_4 values less than 0.00098, corresponding to 2.5 percent range around $\theta_4 = 0$ under $\chi^2(1)$. There remained (roughly) 1.7×10^5 sets of genes among original (roughly) 26.6×10^5 sets of genes. Then, we looked into the value of λ_7 in this pool of 1.7×10^5 sets.

The set of genes with the largest λ_7 is shown in Fig 4 A, where the value of other two genes, TLOC1 and EST, are plotted in scatter-gram with the WAS values indicated by color. Note that we are now discussing real values of three genes (bounded in $[0, 1]$ by our preprocessing). As the value of WAS increases from 0 to 1, the correlation (COR)[7] between the other two genes emerges through a positive θ_{123}. We may also note in Fig 4 A that as the WAS value changes, the each mean of the two genes seems to change (e.g, the EST values with small

[7] As a reference, when we compute the COR by separating scattered points with a threshold of 0.5 WAS value to two groups, the COR of the points with < 0.5 WAS value and those with ≥ 0.5 was -0.03 and 0.796, respectively. The COR values in the following text and figures are calculated similarly.

WAS values tend to be small). This is because changes in the WAS value affect the marginal distribution of each gene through the terms such as θ_{13}, θ_{23} etc. We can easily quantify such an effect as well (not shown here).

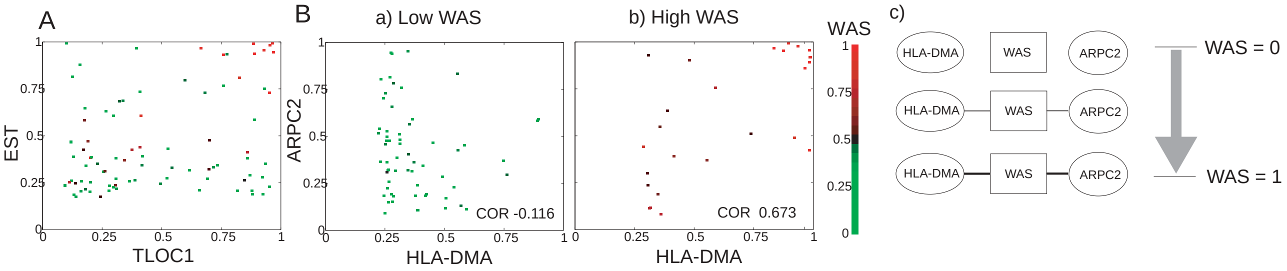

Figure 4: **A, B** Scatter-gram of the expression values of two genes with WAS values colored. In B, with WAS values < 0.5 (a) and ≥ 0.5 (b) **B c** Scheme for the COR driven by WAS.

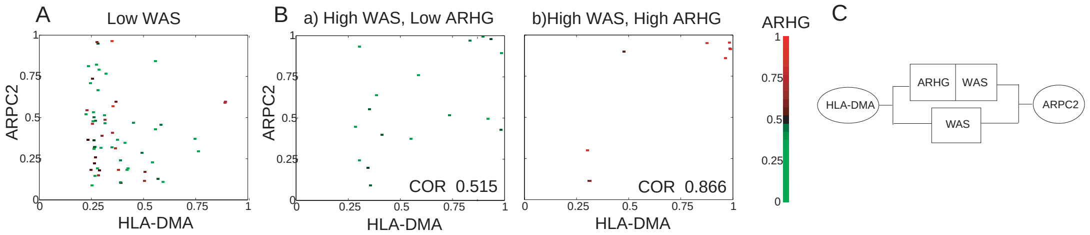

Figure 5: Example by the four variable model. **A** WAS < 0.5 **B** WAS ≥ 0.5 (a) ARHG < 0.5 (b) ARHG ≥ 0.5 **C** Scheme for the COR driven by ARHG and WAS.

There is the set of genes (Fig 4 B), among the top five largest λ_7, that reflects a biologically-known relation, which we consider as supportive for the method's validity. Let us describe the relation shortly: WASP plays an important role in phagocytosis (Higgs and Pollard, 2000; May and Machesky, 2001). WASP is activated by receptors on the cell surface that detect things to take up. Activated WASP interacts with Arp 2/3 complex, which assembles actin filaments so that plasma membrane invagination occurs. Then, phagosome, fused with lysosomes, forms phagolysosomes in which the engulfed material is killed and digested. Digested peptides are bound with HLA-DM protein and then transported to the cell surface for presentation to T cells (Ramachandra et al., 1999). The fact that single mutation in WAS gene causes Wiskott-Aldrich syndrome indicates that WASP is indispensable in this whole process.

Our method indicates that the COR between ARPC2 (Arp 2/3 complex) and HLA-DMA emerges as WASP becomes expressed (Fig 4 B). This corresponds to the above description in that WASP is indispensable in phagocytosis, in which ARPC2 and HLA-DMA participate. Furthermore, let us observe in Fig 4 Ba that the scattered points tend to align vertically around small HLA-DMA values. In other words, given small WAS values, we observe that (1) the ARPC2 values are rather scattered and (2) that the HLA-DMA values tend to concentrate in small values. This (1) may correspond to the known fact that Arp 2/3 complex can be activated by other proteins than WASP (May, 2001). To the author's knowledge, there is no biological evidence for a direct relation between WASP and HLA-DMA. With (2), yet, our analysis suggest that WASP may have a tight relation to HLA-DMA expression, which is further supported by the observation that the HLA-DMA values tends to be large, when the WASP value is large (Fig 4 Bb).

We then investigate the gene triplet (X_1, X_2, X_3) =(HLA-DMA, ARPC2, WAS), using the four variable model: We investigated the set of gene, in each of which we fixed X_1, X_2, X_3 as above and assigned a gene to X_4 from the rest of genes. Based on Theorem 2, we chose to investigate the set of genes that satisfies $\theta_{14} = \theta_{24} = \theta_{124} = \theta_{12} = 0$ and $\theta_{134} = \theta_{234} = 0$. After collecting the set of genes that satisfies the above, we sorted all the sets in order of $\theta_{123} + \theta_{1234}$. The set with the tenth largest of $\theta_{123} + \theta_{1234}$ is given by X_4 =ARHG[8]. Members of Rho-subfamily proteins are known to activate WASP and ARHG is one of them. Hence, we should expect that the effect of WASP on HLA-DMA and ARPC2 would be enhanced by presence of ARHG, because the more ARHG, the more likely WASP is activated and because WASP can be activated without ARHG, e.g. by other members of Rho-family.

We observed the corresponding relation by the information geometric measure (Fig 5). When the WAS values are small, the COR between ARPC2 and HLA-DMA is rather weak and is not affected by the degree of ARHG expression[9] (Fig 5 A). On the other hand, when the WAS values are large, the COR exists (Fig 4 Bb) and furthermore, the COR is stronger when ARHG is expressed (Fig 5 Bb) than when ARHG is not expressed (Fig 5 Ba).

4 DISCUSSION

We have shown how the information geometric (IG) measure of the log linear model of a binary random vector can be applied to analyze gene networks, using the three and four variable cases and also making correspondence to graphical model framework. We re-represented gene interaction, where gene expressions are given by real values, 'through' this log linear model in terms of θ-coordinates, i.e., the IG measure. This re-representation by the IG measure is very simple, which is one of the strengths of our method. We indicated that each term of the IG measure can have a distinct meaning and showed, using such a distinct meaning, that we can investigate a fine structure of gene interaction, some of which would not be treated in graphical model framework. A method of quantifying the estimated IG measure is also presented. Using a dataset of human tumor cells, we demonstrated the validity of IG measure to investigate gene network in relation to WAS gene. The IG measure successfully discovered biologically known findings, indicating its validity. With its simplicity of procedure and its flexibility of investigating a fine structure, we consider that the IG measure is useful in analyzing microarray expression data, for example, discovering a gene interaction hidden in data and selecting candidate genes for further biological investigation.

Let us briefly comment on limitation and future studies of the present study. First, while a generalization of the three and four variable cases treated here is assured theoretically and also demonstrated in neural spike firing data (Amari, 2001; Nakahara and Amari, submitted), we should try to examine the method on expression data, using a larger size of the model. Second, one of nice properties of graphical model is a systematic treatment of directed and undirected graphs, which makes graphical models user-friendly. Estimated directed graphs help users interpret the estimated dependency, sometime causality of variables in relation to

[8] Description of quantified selection (using λ_k) is omitted here. Also, CDC42, the most famous protein that activates WASP, was not included in the dataset.

[9] The COR values with two groups of small and large ARHG values (i.e. < 0.5 and ≥ 0.5 values) are -0.28 and 0.09, respectively.

biological knowledge. While the IG measure is excellent in investigating a finer structure of gene interaction, so far, it is difficult to interpret estimated IG measure as directed graphs, or causality. This issue remains for future.

References

T. Akutsu, S. Miyano, and S. Kuhara. 2000. Inferring qualitative relations in genetic networks and metabolic pathways. *Bioinformatics*, 16(8):727–734.

S. Amari and H. Nagaoka. 2000. *Methods of Information Geometry.* AMS and Oxford University Press.

S. Amari. 2001. Information geometry on hierarchical decomposition of stochastic interactions. *IEEE Transaction on Information Theory*, pages 1701–1711.

Y. M. M Bishop, S. E. Fienberg, and P. W. Holland. 1975. *Discrete Multivariate Analysis.* MIT Press, Cambridge, USA.

M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. 1998. Cluster analysis and display of genome-wide expression patterns. *PNAS*, 95(25):14863–14868.

N. Friedman, M. Linial, I. Nachman, and D. Pe'er. 2000. Using bayesian networks to analyze expression data. *J Comput Biol*, 7(3-4):601–620.

H. N. Higgs and T. D. Pollard. 2000. Activation by cdc42 and pip2 of wiskott-aldrich syndrome protein (wasp) stimulates actin nucleation by arp2/3 complex. *J Cell Biol*, 150(6):1311–1320.

J. Khan, J. S. Wei, M. Ringner, L. H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. R. Antonescu, C. Peterson, and P. S. Meltzer. 2001. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat Med*, 7(6):673–679.

S. L. Lauritzen. 1996. *Graphical Models.* Oxford Science Publications, Oxford.

R. C. May and L. M. Machesky. 2001. Phagocytosis and the actin cytoskeleton. *J Cell Sci*, 114(6):1061–1077.

R. C. May. 2001. The arp2/3 complex: a central regulator of the actin cytoskeleton. *Cell Mol Life Sci*, 58(11):1607–1626.

H. Nakahara and S. Amari. 2002. Information-geometric decomposition in spike analysis. In T. G. et al. Dietterich, editor, *NIPS*, volume 14, page in press. MIT Press: Cambridge.

H. Nakahara and S. Amari. submitted. Information geometric measure for neural spikes.

J. Pearl. 1988. *Probablistic reasoning in intelligent systems.* Morgan Kaufmann, San Mateo, CA.

D. Pe'er, A. Regev, G. Elidan, and N. Friedman. 2001. Inferring subnetworks from perturbed expression profiles. *Bioinformatics*, 17 Suppl 1:S215–S224.

L. Ramachandra, E. Noss, W. H. Boom, and C. V. Harding. 1999. Phagocytic processing of antigens for presentation by class ii major histocompatibility complex molecules. *Cell Microbiol*, 1(3):205–214.

J. Whittaker. 1990. *Graphical models in applied multivariate statistics.* John wiley & sons, Chichester, England.