Gene interaction in DNA microarray data is decomposed by information geometric measure

Hiroyuki Nakahara¹,*, Shin-ichi Nishimura¹,³, Masato Inoue¹,⁴, Gen Hori² and Shun-ichi Amari¹

¹Lab. for Mathematical Neuroscience, RIKEN Brain Science Institute, Saitama 351-0198, ²Laboratory for Advanced Brain Signal Processing, RIKEN Brain Science Institute, ³Department of Otolaryngology, Faculty of Medicine, University of Tokyo and ⁴Department of Otolaryngology, Graduate School of Medicine, Kyoto University, Japan

Received on May 1, 2002; revised on November 18, 2002; December 10, 2002; accepted on December 18, 2002

ABSTRACT

Motivation: Given the vast amount of gene expression data, it is essential to develop a simple and reliable method of investigating the fine structure of gene interaction. We show how an information geometric measure achieves this.

Results: We introduce an information geometric measure of binary random vectors and show how this measure reveals the fine structure of gene interaction. In particular, we propose an iterative procedure by using this measure (called IPIG). The procedure finds higher-order dependencies which may underlie the interaction between two genes of interest. To demonstrate the method, we investigate the interaction between the two genes of interest in the data from human acute lymphoblastic leukemia cells. The method successfully discovered biologically known findings and also selected other genes as hidden causes that constitute the interaction.

Availability: Softwares are currently not available but are possibly made available in future at http://www.mns.brain.riken.go.jp/~nahara/DNA_pub.html, where all the related information is also linked.

Contact: hiro@brain.riken.go.jp

1 INTRODUCTION

Experiments using DNA microarray chips provide us with a vast amount of information on gene expressions through mRNA transcripts. One of the central challenges is to discover the relationships among gene expression, or gene interactions, hidden in data.

Hierarchical clustering is perhaps the most popular method for this purpose and is shown to be useful in inspecting gene networks (Eisen et al., 1998). This method, however, relies entirely on the second-order interaction (i.e. pairwise interaction). To investigate a gene network, we need to know not only the pairwise but also the third-order and higher-order interactions. For example, we may ask whether one gene may regulate two other genes or not. This is a question of the third-order (and higher) interaction, not of the second order. Recently, graphical models (GMs), including Bayesian networks, have been proven to be useful in inspecting a gene network (Friedman et al., 2000; Pe’er et al., 2001). GMs are a general framework to investigate interaction of random variables (Lauritzen, 1996) and can investigate the higher-order interaction of a gene network, using a graphical structure under the strong definition of conditional independence (Whittaker, 1990).

The present study, based on the information geometry framework (Amari and Nagaoka, 2000), focuses on a specific simple probability model, a probability of a binary random variable vector (Bishop et al., 1975; Akutsu et al., 2000). With this model, we investigate the gene interaction under the weak definition of conditional independence and show that our approach can investigate the finer structure of gene interaction. Specifically, we present an iterative procedure (called IPIG) to decompose a pairwise interaction of the two genes into the elements of higher-order interactions. This procedure is simple and easy to implement, once we fully utilize the properties of the model (Amari, 2001; Nakahara and Amari, 2002a,b), and would have strong merit in microarray data analysis.

The present paper is organized as follows. First, we summarize the properties of the information geometric measure. Second, we propose the iterative procedure, IPIG. Third, we touch upon our preprocessing. Fourth, we show its validity using a microarray dataset of human
acute lymphoblastic leukemia cells. Finally, a discussion follows.

2 METHODS

2.1 Conditional independence

We begin by discussing two definitions, namely weak and strong definitions, of conditional independence. For simplicity, let us consider an example of three variables where two variables \((X_1, X_2)\) are independent conditionally to the other \(X_3\),

\[
P(X_1, X_2|X_3) = P(X_1|X_3)P(X_2|X_3).
\]

The strong definition asserts that the above equation should hold regardless of the values taken by \(X_3\), whereas the weak definition concerns whether the above relation holds or not, given a specific value taken by \(X_3\). When \(X_3\) is binary (i.e. 0 or 1), under the weak definition, we would ask, separately, whether \(P(X_1, X_2|X_3 = 0) = P(X_1|X_3 = 0)P(X_2|X_3 = 0)\) holds and/or \(P(X_1, X_2|X_3 = 1) = P(X_1|X_3 = 1)P(X_2|X_3 = 1)\) holds, whereas we say this conditional independence holds under the strong definition only when both equations hold.

2.2 Information geometric measure

Here, we briefly mention a general log linear model of a binary random variable vector. Any probability distribution of binary random vectors can be exactly log-expanded, so that the log linear `model' refers to this exact log linear expansion, not to any kind of approximated expansion. See Amari (2001); Nakahara and Amari (2002b) for further details (also Bishop et al., 1995).

Let us denote an \(n\)-dimensional binary random variable vector by \(X = (X_1, \ldots, X_n)\). Each \(X_i\) represents one gene and takes the values 0 or 1, indicating that \(X_i\) is not expressed or is fully expressed, respectively, in each microarray experiment. In general, the gene expressions in microarray data are real values (i.e. taking any values in \([0, 1]\) in the above notation). For the moment, however, let us assume that the gene expression, \(X_i\), is binary, for presentation simplicity.\(^1\)

Let \(p = p(x), x = (x_1, \ldots, x_n), x_i = 0, 1\), be its probability. Each \(p(x)\) is given by \(2^n\) probabilities

\[
p_{i_1\cdots i_n} = \text{Prob}\{X_1 = i_1, \ldots, X_n = i_n\},
\]

\(i_k = 0, 1, \quad \text{subject to } \sum_{i_1, \ldots, i_n} p_{i_1\cdots i_n} = 1\)

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

\[
\eta_i = E[x_i], \quad \eta_{ij} = E[x_i x_j], \quad (i < j), \ldots, \eta_{12\cdots n} = E[x_1 \cdots x_n]
\]

which have \(2^n - 1\) components. This coordinate system is called \(\eta\)-coordinates (Amari and Nagaoka, 2000). On the other hand, \(p(x)\) can be exactly log-expanded by

\[
\log p(x) = \sum_{i < j} \theta_{ij} x_i x_j + \sum_{i < j < k} \theta_{ijk} x_i x_j x_k + \ldots + \eta_{12\cdots n} x_1 \cdots x_n - \psi, \quad (1)
\]

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

Given gene expression data, both of the above coordinates can be easily estimated in principle (see Section 2.3). Information geometry assures us that the \(\eta\)- and \(\theta\)-coordinates are dually orthogonal coordinates. This property remarkably simplifies an investigation on dependency of random variables.

In a most general case, a \(n\)-dimensional binary random vector results in \((2^n - 1)\) dimensional coordinates. In microarray data, \(n\) may become \(\mathcal{O}(10^4)\) so that we are unlikely to have enough samples to estimate all coordinates. Any method needs to use some assumptions to overcome the limited number of trials. Hierarchical clustering assumes only the pair-wise interaction. GMs use the strong definition of conditional independence, limit candidates of graph structure, incorporate prior knowledge (i.e. Bayesian) and so on. Similarly, in our approach, we should not use a full model but restrict the model in some ways (Nakahara and Amari, 2002b). The IPIG proposed below is one such approach.

2.3 Conditional independence in information geometric measure

One of the key advantages to use the information geometric measure (e.g. \(\theta\)-coordinates in Equation 1) is that it allows a succinct expression of the weak conditional independence (Theorem 1). Let us divide the indices, \(i = 1, \ldots, n\), into the three mutually exclusive, non-empty subsets, denoted by \((A, B, C)\). Any elements of \(A, B\) etc are indicated by small letters (e.g. \(a \in A\)). Given \(X = (X_1, \ldots, X_n)\), we use the notation \(X_A\), which refers to the set of \(X_i\), whose indices belong to \(A\). We have \(X = X_A \cup X_B \cup X_C\). Let us consider,

\[
P(X_A, X_B|X_C) = P(X_A|X_C)P(X_B|X_C), \quad (2)
\]
where $x_C$ indicates a specific value of $X_C$. Given $x_C$, we divide the indices of $C$ into two terms,

$$C_0 = \{ i ; x_i = 0, i \in C \}, \quad C_1 = C - C_0.$$  \hfill (3)

Each component of the $\theta$-coordinates (Equation (1)) corresponds to the set of indices (e.g. $\theta_{12457}$ corresponds to {1, 2, 4, 5, 7}). Using this correspondence, we now define the subset of $\theta$-coordinates as follows:

$$\Theta(A, B; C_0) = \{ \text{the components whose}$$

$$\text{indices include } a, b \text{ but not } c_0 \}. \hfill (4)$$

Then, we have the following theorem.

**Theorem 1.** Equation (2) $\iff \sum_{\theta \in \Theta(A, B; C_0)} \theta = 0$.

Proof is omitted (see Whittaker, 1990 for the different but related theorem and its proof). The above theorem is given with a simplest case, i.e. Equation 2 provided $X = X_A \cup X_B \cup X_C$, for the presentation simplicity and suffices for the present paper. The generalization is possible. When there are more than two conditioned variables, e.g. $P(X_A, X_B, X_C | X_D) = P(X_A | X_D)P(X_B | X_D)P(X_C | X_D)$, we can prove the similar condition recursively. Also, the condition in case of $X_A \cup X_B \cup X_C \subset X$ can be derived similarly.

### 2.3.2 Four variable case

For the three variable case, the log linear model is given by

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

Any distribution of three binary random variables can be represented by this model. It is easy to compute all coefficients, i.e. easy to estimate them from gene expression data, e.g. $\theta_1 = \log \frac{p(x_1)}{p(\bar{x}_1)}; \theta_{12} = \log \frac{p(x_1 x_2)}{p(\bar{x}_1) p(\bar{x}_2)}$. As an example, let us seek the condition of $P(X_1, X_2 | x_3) = P(X_1 | X_3) P(X_2 | X_3)$. By Theorem 1, we find $\theta_{12} = 0$ and $\theta_{123} = 0$ in the cases of $x_3 = 0$ and $x_3 = 1$, respectively.

Obviously, we have $\theta_{12} = \theta_{123} = 0$ $\iff P(X_1, X_2 | x_3) = P(X_1 | X_3) P(X_2 | X_3)$, i.e. the the strong definition of the conditional independence. GMs exploit this relationship. In contrast, using $\theta$-coordinates with the weak definition, we can dig into a finer structure of gene interaction. $\theta_{12}$ and $\theta_{123}$ have distinct meanings: $\theta_{12}$ indicates the interaction of the two genes, $X_1$ and $X_2$, only when $X_3 = 0$, while $\theta_{12} + \theta_{123}$ indicates the interaction when $X_3 = 1$. Furthermore, $\theta_{123}$ indicates the difference of the interaction between the cases of $X_3 = 0$ and $X_3 = 1$. Hence, if we find $\theta_{12} = 0$ but $\theta_{123} \neq 0$, two genes, $X_1$ and $X_2$, are independent only when the gene $X_3$ is not expressed ($X_3 = 0$), but become dependent when the gene $X_3$ is expressed ($X_3 = 1$). This is useful for understanding the effect of the gene $X_3$ on the other two genes. Finally, we note that their signs can be understood naturally, e.g. the sign of $\theta_{12}$ indicates the positive or negative interaction respectively when $X_3 = 0$.

#### 2.3.2.1 Three variable case

For the three variable case, $\log p(x)$ becomes

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

Let us consider one example,

$$P(X_1, X_2 | x_3, x_4) = P(X_1 | x_3) P(X_2 | x_3, x_4)$$

for all four possible cases, i.e. $(x_3, x_4) = (0, 0), (0, 1), (1, 0), (1, 1)$. By Theorem 1, we get four corresponding conditions, $\theta_{12} = 0, \theta_{12} + \theta_{124} = 0, \theta_{12} + \theta_{123} = 0$ and $\theta_{12} + \theta_{123} + \theta_{124} + \theta_{1234} = 0$ in this order. Comparison of these values estimated from microarray data provides us with valuable information. For example, comparing the last two conditions, $\theta_{124} + \theta_{1234}$ indicates the difference of the interaction of $X_1$ and $X_2$ between two cases of $X_3 = 0$ and $X_3 = 1$, conditional to $X_3 = 1$. We will use these features in an iterative procedure below. There is a hierarchical nature in the $\theta$-coordinates with respect to the number of variables, e.g. the three and four variables models (see Nakahara et al., 2002b).

#### 2.4 Quantifying estimates

In any statistical estimation of parameters, we need to quantify significance of its estimated values, which can be easily done for the information geometric measure. Due to the limited space, we only briefly state the results for the simplest case in the three variable model. See Nakahara et al. (2002b) for a more account, Nakahara and Amari (2002b) for more details (also Bishop et al., 1975; Whittaker, 1990).

We write the mixed coordinates,

$$\xi_k = (\xi_1, \ldots, \xi_7) = (\eta_1, \ldots, \eta_k - 1, \theta_k, \eta_{k+1}, \ldots, \eta_7).$$

Each of these mixed coordinates ($k = 1, \ldots, 7$) is useful in singling out different dependency. We focus on the case of $\xi_7$, which is to single out the triplewise dependency (interaction) i.e. $\theta_7 = \theta_{123}$. Let our estimate $\hat{\xi}_7$ and our null hypothesis $\xi_7^0$. To quantify the difference in $\xi_7$, i.e. $\Delta \xi_7 = \hat{\xi}_7 - \xi_7^0$, KL divergence,

$$D(\xi_7^0; \xi_7) = \sum_x p(x; \xi_7^0) \log \frac{p(x; \xi_7^0)}{p(x; \xi_7)},$$  

is used. We have, asymptotically,

$$2N D(\xi_7^0; \xi_7) \approx I_{NN}(\xi_7^0)(\Delta \xi_7)^2 \equiv \lambda_7$$
where $N$ is the number of samples. $I(\xi^0)$ indicates the Fisher information matrix, with dimension $7 \times 7$, of the mixed coordinates at the point $\xi^0$ in the probability space. By the formulation of the likelihood ratio test, we can even get the $p$-value by $\lambda_7 \sim \chi^2(1)$. The similar procedure is available to single out any subset of $\theta$ i.e. more than two $\theta$ components together (Nakahara and Amari, 2002b).

### 2.5 IPIG; Iterative procedure to inspect two gene interaction

There are various ways to utilize the information geometric measure in microarray data analysis (Nakahara and Amari, 2002b; Nakahara et al., 2002a). Here we focus on the task of investigating an interaction between two genes of interest and of discovering other genes that may influence the interaction. Due to the limited number of samples, it will be, most likely, impractical to start with the full log linear model (Section 2.2).

We hence propose an iterative procedure using the information geometric measure (called IPIG). In the IPIG, given two genes of interest, we incrementally add candidate genes that may regulate its interaction. The IPIG is in the same spirit as the stepwise procedure of variable selection in regression (Draper and Smith, 1998). The selected variables may not necessarily be the best ones, when we consider all possible subsets of variables. Yet the procedure is easy to implement and proceed and may provide reasonably good subsets, which can be then submitted to further biological investigation.

Let us begin by denoting the two genes of interest by $(X_{1s}, X_{2s})$. The number of the remaining genes is $n - 2$ so that they are renumbered as $X_3, \ldots, X_n$ by omitting $X_{1s}$ and $X_{2s}$ from the original set of genes (this renumbering is assumed below in each iteration). In this remaining set of genes, we seek the third gene, $X_{3a}$, which gives the maximum value of $\theta^3_{12k}$ given $X = (X_{1s}, X_{2s}, X_k)$, that is,

$$X_{3a} = \arg \max_{X_k} \theta^3_{12k} \quad (k = 3, 4, \ldots, n).$$

With the three variable model, this $X_{3a}$ gives the maximal change in the interaction of $(X_{1s}, X_{2s})$ when we compare the cases of $X_{3a} = 0$ and $1$ (see Section 2.3.1).

Next we consider the four variable model, $X = (X_{1s}, X_{2s}, X_{3a}, X_k)$. Suppose we are interested in the fourth gene regulation on $(X_{1s}, X_{2s})$ when $X_{3a}$ is expressed ($X_{3a} = 1$). We then search for

$$X_{4s} = \arg \max_{X_k} \theta^4_{12k} + \theta^4_{123k} \quad (k = 4, \ldots, n).$$

This $X_{4s}$ gives the maximal change in the interaction of $(X_{1s}, X_{2s})$ conditional to $X_{3a} = 1$ (see Section 2.3.2). Similarly, given the five variable models, we find $X_{5s} = \arg \max_{X_k} \theta^5_{12k} + \theta^5_{123k} + \theta^5_{124k} + \theta^5_{1234k}$, where $k = 5, \ldots, n$, this $X_{5s}$ gives the maximum change in the interaction conditional to $X_{3a} = X_{4s} = 1$.

### General formula of IPIG

We now provide a general formula of IPIG. Given that $m$ genes (including $X_{1s}, X_{2s}$) are selected, suppose we want to search for the $(m + 1)$th gene that gives maximal change in the interaction of $X_{1s}$ and $X_{2s}$ conditional to a specific value of $x^m = (x_{3s}, \ldots, x_{ms})$, where $x_{is} \in \{0, 1\}$.

To detect this $(m + 1)$th gene (which is the $(m - 1)$th iteration), IPIG is given by

$$X_{m+1s} = \arg \max_{X_k} \sum_{\theta \in \Delta_m} \theta, \quad (k = m + 1, \ldots, n)$$

where we define $\Delta_m$ by

$$\Delta_m = \Theta(\{1s\}, \{2s\}; C_0^{m+}) - \Theta(\{1s\}, \{2s\}; C_0^{m-}). \quad (6)$$

Here, $C_0^{m+}$ and $C_0^{m-}$ denote the set of indices each of which gives $x_i = 1$ among the indices in the specific values of $x^{m+} = (x_{i1} = 1, \ldots, x_{im+} = 1)$ and $x^{m-} = (x_{i1} = 0, \ldots, x_{im-} = 0)$, respectively (modified from Equation (3)), $\Theta(\{1s\}, \{2s\}; C_0^{m-})$ (see Equation (4)) denotes the subset of components in the $\theta$-coordinates under the log linear model of $m + 1$ variables, which is the set of components whose indices include 1* and 2* but do not include any indices belonging to $C_0^{m+}$. $\Theta(\{1s\}, \{2s\}; C_0^{m-})$ is defined similarly. $\Delta_m$ is the set of components that is included in $\Theta(\{1s\}, \{2s\}; C_0^{m-})$ but not in $\Theta(\{1s\}, \{2s\}; C_0^{m+})$.

Notably, the evaluation of $\sum_{\theta \in \Delta_m} \theta$ amounts to evaluating the conditional probability of the three variables (i.e. $X_{1s}, X_{2s}, X_k$), that is, $P(X_{1s}, X_{2s}, X_k|x^m)$. In other words, once we re-parameterize this conditional probability by $\theta' = (\theta_1', \ldots, \theta_{1223}') = (\theta_1', \ldots, \theta_{1232}')$, then, we have $\theta_1' = \sum_{\theta \in \Delta_m} \theta$. Therefore, the quantity evaluated derived in Section 2.5 can be easily done in each iteration. Equation (5) can be also regarded as maximizing the difference between $P(X_{1s}, X_{2s}|x^{m+})$ and $P(X_{1s}, X_{2s}|x^{m-})$. This property can be used by other exploratory procedures (e.g. see Nakahara et al., 2002a). Finally, although we presented IPIG as a ‘strictly’ iterative procedure (i.e. building $x^m$ iteratively), Equation (5) can be performed with any pre-chosen $x^m$. In other words, if there is any prior knowledge (e.g. known regulatory interaction), we can take it into account in choosing $x^m$, which may be called a modified IPIG.
2.6 Preprocessing

There are two steps in our preprocessing. The first step is to discard genes with dubious expression and then normalize the data. Due to the limited space, we skip describing the first step (see Nakahara et al., 2002b).

The second step is to convert the normalized data into the coordinates of the log linear model. We consider the gene expression data, which take real values, as indicating only the relative degree of expression among different experimental conditions. We re-represent this relative degree as bounded in [0, 1], where 0 and 1 indicate zero and full expression, respectively and then represent interaction of genes (with the bounded values) ‘through’ a log linear model of a binary random vector. In other words, we do not regard gene expression by itself as binary.

We provide one approach, applicable to families of linear rank statistics in general (Nakahara et al., 2002a), to convert a real value of each gene into a value bounded in [0, 1]. Here, we adopted a simple one, namely rank order statistics.

We denote the samples for the variable \( X_i \) by \( X_i = (x_{i1}, x_{i2}, \ldots, x_{ij}, \ldots, x_{iN}) \), where each \( x_{ij} \) has a real value, and denote the rank order of \( x_{ij} \) by \( x_{i(j)} \). We then construct the corresponding \( Z^N_i = (z_{i1}, z_{i2}, \ldots, z_{ij}, \ldots, z_{iN}) \), where \( z_{ij} = x_{i(j)}/N \).

Now, each \( Z^N_i \) expresses the relative degree of expression, bounded by [0, 1]. Using these \( Z^N_i \)s, for example, \( p_{100} \) is computed by

\[
p_{100} = \frac{1}{N} \sum_{j} z_{ij} (1 - z_{2j})(1 - z_{3j}).
\]

In a similar manner, all \( p_{ijk} \) (and any \( p_{ij2,\ldots,ln} \)) can be computed, so that it is easy to obtain both \( \theta \)- and \( \eta \)-coordinates after this preprocessing. Thus, we converted real-values of gene expression to numbers bounded in coordinates after this preprocessing. Thus, we converted real-values of gene expression to numbers bounded in coordinates after this preprocessing. Thus, we converted real-values of gene expression to numbers bounded in coordinates after this preprocessing. Thus, we converted real-values of gene expression to numbers bounded in coordinates after this preprocessing. Thus, we converted real-values of gene expression to numbers bounded in coordinates after this preprocessing. Thus, we converted real-values of gene expression to numbers bounded in coordinates after this preprocessing.

3 RESULTS

To demonstrate and validate our proposed method, IPIG, we investigated an Affymetrix microarray dataset of human acute lymphoblastic leukemia (ALL) cells (Yeoh et al., 2002), which contains 327 microarray experiments of 12 558 genes.

We chose to investigate the interaction between the genes XBP-1 and IGHM. XBP-1 is a gene for CREB-like transcription factor, required for plasma cell differentiation, while IGHM is a subunit of immunoglobulin secreted from plasma cell. Therefore, the expression of the two genes can be expected to be somewhat positively correlated. However, it is reported (Reimold et al., 2001) that direct transcriptional control of immunoglobulin by XBP-1 was unlikely. Hence, there is a strong interest in discovering other genes that contribute to the interaction of the two genes.

In fact, there was no evident direct correlation between the two genes (Fig. 1A)\(^\dagger\). In the first iteration of IPIG, we found that the ADPRT gene gives the maximal change\(^\ddagger\) in the interaction between XBP-1 and IGHM,

\[
(X_{1a}, X_{2a}; X_{3a}) = (X_{1b}, IGHM; ADPRT).
\]

To visualize the effect of ADPRT, the data-points in Figure 1A were divided into two groups, with low (<0.5) and high (>0.5) values of ADPRT (Fig. 1B a,b). The sign of \( \theta_{123} \) was negative so that the negative correlation emerged as ADPRT was expressed (Fig. 1B b). When ADPRT was down-regulated, the positive correlation appeared between XBP-1 and IGHM (Fig. 1B a).

IPIG next searched for the fourth gene in two different conditions, namely \( X_{3a} = 0 \) and \( X_{3a} = 1 \). The fourth genes were found to be TM4SF2 and ZFP36L1, conditional to \( X_{3a} = 0 \) and \( X_{3a} = 1 \), respectively.

To show the modulation by TM4SF2 with low values of ADPRT (roughly corresponding to \( X_{3a} = 0 \))\(^\dagger\), the data points in Fig. 1B (a) were divided into groups with low and high values of TM4SF2 (Fig. 1C a,b). The correlation in Figure 1B (a) is modulated by the expression of TM4SF2 so that TM4SF2 expression induces the positive correlation (conditional to low ADPRT expression). Similarly, the modulation by ZFP36L1 is shown in Fig. 1D (a,b), dividing data points in Fig. 1B (b) into groups with low and high values of ZFP36L1. ZFP36L1 expression is facilitatory to ADPRT, strengthening the negative correlation between XBP-1 and IGHM (conditional to high ADPRT expression).

The fifth gene was then searched in two conditions, namely \( X_{3a} = 0, X_{4a} = 1 \), where \( X_{4a} \) = TM4SF2, and \( X_{3a} = 1, X_{4a} = 1 \), where \( X_{4a} \) = ZFP36L1, and then obtained as AF1Q and DFKZp586C1019 (Fig. 1E a and F), respectively. AF1Q expression tends to cause a stronger positive correlation, and DFKZp586C1019 expression a stronger negative correlation. Note that the data points in Figure 1C (b) were largely divided to two groups, concentrated in the upper-right corner (Fig. 1E a) and in Figures 1A and 1C were divided to two groups. In each figure, the correlation is computed only with the points in the figure.

\(^\dagger\) More precisely, ADPRT gene gives the largest \( \lambda \), 4.299, (which yields to \( p = 0.0382 \)).

\(^\ddagger\) The actual calculation to get the \( \theta \)-coordinates is done by using the [0, 1]-bounded values of \( X_{3a} \), i.e. ADPRT. See Section 2.6.

\(^\ddagger\) We emphasize that all data points in Fig. 1A are used in each iteration of IPIG. The visualization in Fig 1, showing only subsets of datapoints in each figure, is only for presentation purpose. In each figure, the correlation is computed only with the points in the figure.
the lower-left corner (Fig. 1E b). This suggests that the apparently strong positive correlation in Figure 1C (b) may be a false one. This kind of observation cannot be made if we inspect only the pair-wise correlation, as hierarchical clustering does.

Figure 2A summarizes the modulation of genes found by IPIG on the interaction between XBP-1 and IGHM. This diagram should not be taken rigidly but can be considered as a guide for further biological examination. Gene data sorted by hierarchical clustering is shown in Fig. 2B. Genes found by IPIG are somewhat close to each other but not next to each other (see figure legend). Thus, IPIG may provide a valuable new information, which would not be found by hierarchical clustering.
Fig. 2. A Scheme drawn from the results in Figure 1. The arrow and flat heads indicate the positive and negative influence onto the gene (and/or the gene interaction), respectively. B Genes are sorted by hierarchical clustering. The positions of genes found by IPIG are also indicated. The gene numbers (counting from the top among 9887) are as follows: AF1Q (5332), DKFZp586C1019 (6037), IGHM (9010), XBP-1 (9070), TM4SF2 (5332), ADPRT (9378), and ZFP36L1 (9384). C Perturbation test. From the original 327 samples, a number of data points is randomly removed and IPIG is used to identify the third gene (i.e. the first iteration of IPIG). This process is repeated a thousand times and we then identified the percentage of the identities of the chosen third gene in the thousand trials (ordinate). The number of data points removed ranges from 1 to 16 (abscissa), where 16 corresponds to 5% of the sample size.

We don’t have a complete biological explanation for the relationship of genes found by IPIG above. But there are several fragmented, relevant pieces of biological information, which are provided elsewhere (Nakahara et al., 2002b) due to the limited space. In short, we consider that the validity of IPIG in microarray data analysis is supported by the biological information.

REMARK. One concern in using IPIG is the stability, i.e. how IPIG behaves if the data is perturbed. The result is shown in Fig. 2C (see the legend for the details of this examination). Only two genes were chosen as the third gene in all conditions. ADPRT, the gene chosen by IPIG with the full sample, was chosen in almost all conditions. TM4SF2 started to occupy 3% with five points removal and became 32.7% with 16 points removal. This indicates that IPIG is somewhat susceptible to perturbation, which is a general tendency of any methods that inspect interaction of random variables. Yet, it also suggests that IPIG works reasonably with a small perturbation in that the ADPRT was dominant, while there was only another gene selected.

4 DISCUSSION
We have shown how the information geometric measure of a binary random vector can be applied to analyze gene interaction. We re-represented real-valued gene interaction by the information geometric measure. This re-representation is simple, one of the strengths of our method. Using the properties of the dual orthogonal coordinates under the weak definition of conditional independence, we can investigate the fine structure of gene interaction. Specifically, we proposed an iterative procedure, called IPIG, to investigate the fine interaction of two genes of interest, examining their higher-order interaction. IPIG is useful in discovering a gene interaction hidden in data. Using dataset of ALL, we demonstrated the validity of IPIG.
We consider it better to provide the discussion on the relation between our approach and two other related methods, namely hierarchical clustering and graphical models. Due to the limited space, however, we only state our view on the relative advantage of our approach; Our approach has advantage for finding and exploring the fine interaction, including higher-order interaction, of a relatively small set of genes (e.g. selecting candidate genes for further biological investigation). See Nakahara et al. (2002b) for our reasoning on this issue.

We also consider it fair to discuss the limitation of IPIG and its future extension. Due to the limited space, we only mention that the result in Remark suggests the practical value of the modified IPIG. See Nakahara et al. (2002b) for a more detailed discussion.

ACKNOWLEDGEMENTS

The authors thank reviewers for their comments. HN thanks J. Read for her comments.

REFERENCES
