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

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, gene interaction, DNA microarray data, acute lymphoblastic leukemia (ALL)

November 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. We show how this measure can be used to reveal the fine structure of gene interaction. In particular, we propose an iterative procedure by using the information geometric 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 genes, XBP-1 and IGHM, using data from human acute lymphoblastic leukemia cells. The method successfully discovered biologically-known findings and also selected other genes as hidden causes to constitute the interaction.

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 infer the gene interaction, while having the advantage of the relatively low computational cost. To investigate a gene network, we need to know not only pairwise but also the third-order and higher-order interactions. As a simple 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,

[*] corresponding author; tel: +81-48-467-9663; fax +81-48-467-9693; hiro@brain.riken.go.jp

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 in statistics and computer science to investigate interaction of random variables (Pearl, 1988; 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 and its log linear expansion (Bishop et al., 1975; Akutsu et al., 2000). With this model, we investigate the gene interaction under the *weak* definition of conditional independence and thereby 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; Nakahara and Amari, 2002b). This procedure 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, to utilize the measure to investigate gene interaction. Third, we touch upon our preprocessing by which we convert the real-valued gene expressions in microarray data to the information geometric measure. Fourth, we show its validity using a microarray dataset of human acute lymphoblastic leukemia cells (Yeoh et al., 2002) and investigate an interaction of two genes, XBP-1 and IGHM[1]. 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. Then we write

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

The strong definition of conditional independence 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. That is, when X_3 is binary, taking 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 holds.

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. Thus, we emphasize that when we say the log linear 'model', it refers to this exact log linear expansion, not to any kind of approximated expansion. See (Amari, 2001; Nakahara and Amari, 2002b) and also (Bishop et al., 1975) for further details.

Let us denote an n-dimensional binary random variable vector by $X = (X_1, \cdots, X_n)$. Each X_i represents one gene and takes the

[1] Gene names below are abbreviated according to UniGene standards

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, till Sec 2.6 [2].

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,

$$\begin{aligned} \eta_i &= E\left[x_i\right], \;\; \eta_{ij} = E\left[x_i x_j\right], \;\; (i<j), ..., \\ \eta_{12 \cdots n} &= E\left[x_1 \cdots x_n\right] \end{aligned}$$

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 log-expanded by

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

where the indices of θ_{ijk}, etc. satisfy $i < j < k$, etc and ψ is the normalization term, corresponding to $-\log p(x_1 = x_2 =, ... = x_n = 0)$. The terms θ_{ijk} together have $2^n - 1$ components, forming another coordinate system, called θ-coordinates (Amari and Nagaoka, 2000). Each term can be easily computed (see Sec 2.3).

Given gene expression data, both of the above coordinates can be easily estimated in principle. Information geometry assures us that the η-coordinates and θ-coordinates are dually orthogonal coordinates. This property remarkably simplifies an investigation on dependency of random variables, as shown below.

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. For example, 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

We discuss here the relation of conditional independence (under the weak definition) with the information geometric measure, i.e., θ-coordinates in Eq 1. One of the key advantages to use the information geometric measure is that it allows a succinct expression of the weak conditional independence, as shown in Theorem 1 below.

Let us divide the indices, $i = 1, ..., n$, into the three mutually exclusive, non-empty subsets, denoted by (A, B, C). Below, we indicate any elements of each of A, B etc by small letters (e.g., $a \in A$). Given $X = (X_1, \cdots, X_n)$, we use the notation X_A, which refers to the set of X_i, whose indices belong to A. Then, we have $X = X_A \cup X_B \cup X_C$.

[2] In our analysis, we do not actually take gene expression to be binary. Section 2.6 describes how we convert real-valued data to the coordinates of the log linear model.

Consider the following equation,

$$P(X_A, X_B|\boldsymbol{x}_C) = P(X_A|\boldsymbol{x}_C)P(X_B|\boldsymbol{x}_C), \tag{2}$$

where we used $\boldsymbol{x}_C$, instead of X_C, to indicate that we consider the above equation with respect to a specific value of X_C, i.e., $\boldsymbol{x}_C$. Given $\boldsymbol{x}_C$, we divide the indices of C into two terms,

$$C_0 = \{i; x_i = 0, i \in C\}, \;\; C_1 = C - C_0. \tag{3}$$

Each component of the θ-coordinates (Eq 1) corresponds to the set of indices (e.g., θ_{12457} corresponds to the set of indices $\{1, 2, 4, 5, 7\}$). Using this correspondence, We now define the subset of θ-coordinates as follows,

$$\begin{aligned}\Theta(A, B; C_0) = \{&\text{the components whose} \\ &\text{indices include } a, b \text{ but not } \; c_0\}.\end{aligned} \tag{4}$$

Then, we have

Theorem 1

Eq 2 $\Longleftrightarrow \sum_{\theta \in \Theta(A,B;C_0)} \theta = 0$.

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, Eq 2 provided $X = X_A \cup X_B \cup X_C$, for the presentation simplicity. It suffices for the present paper. The generalization is possible. For example, when there are more than two conditioned variables, e.g., $P(X_A, X_B, X_C|\boldsymbol{x}_D) = P(X_A|\boldsymbol{x}_D)P(X_B|\boldsymbol{x}_D) P(X_C|\boldsymbol{x}_D)$, 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.

2.3.1 Three variable case

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

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

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., $\theta_1 = \log \frac{p_{100}}{p_{000}}$, $\theta_{123} = \log \frac{p_{111}p_{100}p_{010}p_{001}}{p_{110}p_{101}p_{011}p_{000}}$. As an example, let us seek the condition of $P(X_1, X_2|x_3) = P(X_1|x_3)P(X_2|x_3)$ with respect to θ-coordinates. By Theorem 1, we find $\theta_{12} = 0$ and $\theta_{12} + \theta_{123} = 0$ in the cases of $x_3 = 0$ and $x_3 = 1$, respectively.

Obviously, we then have $P(X_1, X_2|X_3) = P(X_1|X_3)P(X_1|X_3) \iff \theta_{12} = \theta_{123} = 0$, that is, the relation under the strong definition of the conditional independence. GMs exploit this relationship.

In contrast, using θ-coordinates with the weak definition, we can dig into a finer structure of gene interaction. θ_{12} and θ_{123} have distinct meanings: θ_{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, θ_{123} indicates the difference of the interaction between the cases of $X_3 = 0$ and $X_3 = 1$. Hence, 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 ($X_3 = 0$), 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. For example, the sign of θ_{12} indicates the positive or negative interaction respectively when $X_3 = 0$.

2.3.2 Four variable case

The log linear model now becomes

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

Let us consider one example,

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

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, we note that $\theta_{124} + \theta_{1234}$ indicates the difference of the interaction of X_1 and X_2 between two cases of $X_4 = 0$ and $= 1$, conditional to $X_3 = 1$. We will use these features in an iterative procedure below.

Let us remark on the relation between three and four variable cases, using example of $\theta^3_{123} = \log \frac{p_{111}p_{100}p_{010}p_{001}}{p_{110}p_{101}p_{011}p_{000}}$ for the three variable model (and θ^4_{123} for the four variable model). First, we must note $\theta^3_{123} \neq \theta^4_{123}$, since $\theta^4_{123} = \log \frac{p_{1110}p_{1000}p_{0100}p_{0010}}{p_{1100}p_{1010}p_{0110}p_{0000}}$. This is because the three variable model ignores the fourth variable influence on the three variables. In general, a smaller model may miss a finer interaction which can be represented in a larger model. By inspection, we see that θ^4_{123} indicates the third-order interaction among (X_1, X_2, X_3) conditional to $X_4 = 0$, while θ^3_{123} indicates the third-order interaction regardless of the value of X_4. We also note that a simple calculation yields $\theta^4_{123} + \theta^4_{1234} = \log \frac{p_{1111}p_{1001}p_{0101}p_{0011}}{p_{1101}p_{1011}p_{0111}p_{0001}}$ so that it indicates the third-order interaction conditional to $X_4 = 1$. Thus, there is a hierarchical nature in the θ-coordinates with respect to the number of variables.

2.4 Quantifying estimates

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, this is why GMs need to '*propagate beliefs*' and/or '*estimate graph structure*' (Lauritzen, 1996; Pearl, 1988). Taking into account such dependencies, we need to quantify significance of its estimated values. Here we show how we can do so for the information geometric measure. Due to the limited space, here, we only discuss a simple case in the three variable model. See (Nakahara and Amari, 2002b) for more details and also (Bishop et al., 1975; Whittaker, 1990).

Let us first write θ- and η-coordinates;

$$\begin{aligned} \boldsymbol{\theta} &= (\theta_1, .., \theta_7) = (\theta_1, \theta_2, \theta_3; \theta_{12}, \theta_{13}, \theta_{23}; \theta_{123}), \\ \boldsymbol{\eta} &= (\eta_1, .., \eta_7) = (\eta_1, \eta_2, \eta_3; \eta_{12}, \eta_{13}, \eta_{23}; \eta_{123}). \end{aligned}$$

Estimation of these coefficients from data 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 independently of the complementary subset of η-coordinates. Quantification of the estimated values can be done via the Fisher information matrix of the mixed coordinates. This Fisher information matrix has a simple structure again due to the dual orthogonality, resulting in an easy procedure of quantification (Nakahara and Amari, 2002b).

To use the above properties of the dually orthogonal 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 in singling out different dependency. Let us now focus on the case of $\boldsymbol{\zeta}_7$; $\boldsymbol{\zeta}_7$ 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 $(\eta_1, ...\eta_6)$.

Let our estimate of expression data $\hat{\boldsymbol{\zeta}}_7 = (\hat{\eta}_1 ., , .\hat{\eta}_6, \hat{\theta}_7)$ and suppose our null hypothesis, denoted by $\boldsymbol{\zeta}^0_7$, is given by $\boldsymbol{\zeta}^0_7 = (\hat{\eta}_1 ., , .\hat{\eta}_6, \theta^0_7)$, where $\theta^0_7 = 0$. The difference in ζ_7, i.e., $\Delta\zeta_7 = \hat{\theta}_{123} - \theta^0_{123} = \hat{\theta}_7$, is not quantified yet. This is because the metric varies at each probability distribution $p(\boldsymbol{x}; \boldsymbol{\zeta}_7)$. We can, however, quantify the discrepancy between $p(\boldsymbol{x}; \boldsymbol{\zeta}^0_7)$ (null hypothesis) and $p(\boldsymbol{x}; \hat{\boldsymbol{\zeta}}_7)$ (our estimate) by the KL divergence,

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

The KL divergence works as a quasi-distance measure. We have, asymptotically,

$$\begin{aligned} 2ND(\boldsymbol{\zeta}_7^0;\hat{\boldsymbol{\zeta}}_7) &\approx I_{77}(\boldsymbol{\zeta}_7^0)(\Delta\zeta_7)^2 \\ &= I_{77}(\boldsymbol{\zeta}_7^0)(\hat{\theta}_7)^2 \equiv \lambda_7 \end{aligned}$$

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 $\sum_{i,j=1}^{7} I_{ij}(\boldsymbol{\zeta})\zeta_i\zeta_j$. 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, 2002b). Hence, we can easily quantify the estimation of θ_{123} and then, following the formulation of the likelihood ratio test, can even get the p-value by $\lambda_7 \sim \chi^2(1)$, where $\chi^2(k)$ denotes χ^2 probability distribution with k degree of freedom.

In a similar manner, we can inspect any of θ_k $(k = 1, ..., 7)$ by using $\boldsymbol{\zeta}_k$. Furthermore, the similar procedure is available to single out any subset of $\boldsymbol{\theta}$ i.e. more than two θ 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., 2002). 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. A simple way to do so is to use a full log linear model having all genes of interest and inspect the estimated values of the θ-coordinates. However, it is likely that the number of samples is limited (Sec 2.2) and also that candidate genes of interest may not be known (that is, all genes in the data, $\sim \mathcal{O}(10^4)$, may be potentially interesting). Therefore, it is rather impractical to start with the full model.

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 itself 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_{1*}, X_{2*}). The number of the remaining genes is $n-2$ so that they are renumbered as $X_3, ..., X_n$ by omitting X_{1*} and X_{2*} 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_{3*}, which gives the maximum value of θ_{12k}^3 given $X = (X_{1*}, X_{2*}, X_k)$, that is[3],

$$X_{3*} = \arg\max_{X_k} \theta_{12k}^3 \ \ (k = 3, 4, .., n).$$

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

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

$$X_{4*} = \arg\max_{X_k} \theta_{12k}^4 + \theta_{123k}^4 \ \ (k = 4, .., n).$$

[3] More precisely, we should search for $\arg\max_{X_k} \lambda_7$ (see Sec 2.4 for this notation). Here and below, however, for notational simplicity, we write the equations in terms of the 'raw' values (e.g. θ_{123}), but they should be interpreted as seeking the variable maximizing the quantified values (e.g. λ_7). In the Results section, we maximized the quantified values.

This X_{4*} gives the maximal change in the interaction of (X_{1*}, X_{2*}) conditional to $X_{3*} = 1$ (see Sec 2.3.2). Similarly, given the five variable models, when we find

$$X_{5*} = \arg\max_{X_k} \theta^5_{12k} + \theta^5_{123k} + \theta^5_{124k} + \theta^5_{1234k},$$

where $k = 5, .., n$, this X_{5*} gives the maximum change in the interaction conditional to $X_{3*} = X_{4*} = 1$.

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

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

$$X_{m+1*} = \arg\max_{X_k} \sum_{\theta \in \mathit{\Delta}_m} \theta, \quad (k = m+1, .., n) \tag{5}$$

where we define $\mathit{\Delta}_m$ by

$$\begin{aligned} \mathit{\Delta}_m &= \mathit{\Theta}(\{1*\}, \{2*\}; C_0^{m+}) \\ &\quad - \mathit{\Theta}(\{1*\}, \{2*\}; C_0^{m-}). \end{aligned} \tag{6}$$

Here, C_0^{m+} and C_0^{m-} denote the set of indices each of which gives $x_i = 0$ among the indices in the specific values of $\boldsymbol{x}^{m+} \equiv (\boldsymbol{x}^m, x_{m+1} = 1)$ and $\boldsymbol{x}^{m-} \equiv (\boldsymbol{x}^m, x_{m+1} = 0)$, respectively (modified from the notation in Eq 3). As defined in Eq 4, $\mathit{\Theta}(\{1*\}, \{2*\}; C_0^{m+})$ denotes the subset of components in the θ-coordinates under the log linear model of $m+1$ variables, that is, the set of components whose indices include $1*$ and $2*$ but do not include any indices belonging to C_0^{m+}. $\mathit{\Theta}(\{1*\}, \{2*\}; C_0^{m-})$ is defined similarly and $\mathit{\Delta}_m$ is the set of components that is included in $\mathit{\Theta}(\{1*\}, \{2*\}, C_0^{m+})$ but not in $\mathit{\Theta}(\{1*\}, \{2*\}; C_0^{m-})$.

Notably, the evaluation of $\sum_{\theta \in \mathit{\Delta}_m} \theta$ amounts to evaluating the conditional probability of the three variables (i.e., X_{1*}, X_{2*}, X_k), that is, $P(X_{1*}, X_{2*}, X_k | \boldsymbol{x}^m)$. In other words, once we re-parameterize this conditional probability by $\boldsymbol{\theta}' = (\theta'_1, .., \theta'_7) = (\theta'_1, .., \theta'_{123})$, then, we have $\theta'_7 = \sum_{\theta \in \mathit{\Delta}_m} \theta$. Therefore, the quantitive evaluation derived in Sec 2.4 can be easily done in each iteration. Eq 5 can be also regarded as maximizing the difference between $P(X_{1*}, X_{2*} | \boldsymbol{x}^{m+})$ and $P(X_{1*}, X_{2*} | \boldsymbol{x}^{m-})$. This property can be used by other exploratory procedures, one of which is given by (Nakahara et al., 2002). Finally, although we presented IPIG as a 'strictly' iterative procedure (i.e. building $\boldsymbol{x}^m$ iteratively), Eq 5 can be performed with any pre-chosen $\boldsymbol{x}^m$. In other words, if there is any prior knowledge(e.g. known regulatory interaction), we can take it into account in choosing $\boldsymbol{x}^m$, which may be called a modified IPIG.

2.6 Preprocessing

There are two steps in our preprocessing procedure. The first step is to discard genes with dubious expression and then normalize the data. The data treated in the present study is obtained by Affymetrix microarray. First, genes with low variance (below 10000) in the Average Difference (AvgDiff) value were omitted. Then, the AvgDiff values were given by 'Log-intensity = log10(max(10, AvgDiff))'. Between-array normalization was performed so that the mean expression intensities of the genes for each of the arrays become equal.

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 (e.g. normal cell/cancer cell). We wish to re-represent this relative degree as bounded in $[0, 1]$, where 0 and 1 indicate zero and full expression, respectively. Then we 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 as shown below.

How then can we convert a real value of each gene into a value bounded in $[0, 1]$ There are several possibilities, and we provide one approach below. This approach is applicable to families of linear rank statistics in general (Nakahara et al., 2002) and in the present study, we adopted a simple one among them, namely rank order statistics.

We denote the samples for the variable X_i by $X_i^N = (x_{i1}, x_{i2}, ., x_{ij}, ..x_{iN})$, where each x_{ij} has a real value. We denote the rank order of x_{ij} by $x_{i(j)}$ and then construct the corresponding $Z_i^N = (z_{i1}, z_{i2}, ., z_{ij}, ..z_{iN})$, where $z_{ij} = x_{i(j)}/N$. Now, each Z_i^N expresses the relative degree of expression, bounded by $[0, 1]$. Using these Z_i^Ns, 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. Thus, using rank order statistics, we converted real-values of gene expression to numbers bounded in $[0, 1]$ and further mapped the bounded numbers to the θ-coordinates (and other coordinates).

3 RESULTS

To demonstrate and validate our proposed method (particularly IPIG), we investigated an Affymetrix microarray dataset of human acute lymphoblastic leukemia (ALL) cells (Yeoh et al., 2002)[4], which contains 327 microarray experiments of 12,558 genes[5]. ALL is a malignant disease of the bone marrow in which lymphoid precursor cells proliferate and replace the normal hematopoietic cells of the marrow. The malignant lymphoid precursor cells have malfunctions in the differentiation process.

In the present study, we chose to investigate the interaction between the genes XBP-1 and IGHM. XBP-1 (X-box-binding protein-1) is a gene for CREB-like transcription factor and is required for plasma cell differentiation (Reimold et al., 2001), while IGHM (constant region of heavy chain of IgM) 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, Reimold et al (2001) reported 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.

(Figure 1 is around here)

In fact, there was no evident correlation between XBP-1 and IGHM, as shown in Fig 1 A[6]. IPIG was then employed to discover such other genes. In the first iteration of IPIG, we found that the ADPRT gene gives the maximal change[7] in the interaction between XBP-1 and IGHM. Thus, we set

$$(X_{1*}, X_{2*}; X_{3*}) = (\text{XBP-1}, \text{IGHM}; \text{ADPRT}).$$

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

IPIG next searched for the fourth gene in two different conditions, namely $X_{3*} = 0$

[4] from http://www.stjuderesearch.org/data/ALL1

[5] After the first step of our preprocessing, the number of genes became 9,887.

[6] In this and following figures, the scale of each gene is bounded by $[0, 1]$ due to our preprocessing.

[7] More precisely, ADPRT gene gives the largest λ_7, 4.299, (which yields to $p = 0.0382$).

and $X_{3*} = 1$[8]. The fourth genes were found to be TM4SF2 and ZFP36L1, conditional to $X_{3*} = 0$ and $X_{3*} = 1$, respectively.

We visualized each gene's modulation as follows. The modulation by TM4SF2 was shown for the data-points with low values of ADPRT (roughly corresponding to $X_{3*} = 0$)[9]. In other words, the data-points in Fig 1 B (a) were divided into groups with low and high values of TM4SF2, and re-plotted in Fig 1 C (a) and (b), respectively. We observe that the correlation in Fig 1 B (a) is modulated by the expression of TM4SF2 into Fig 1 C (a) and (b). TM4SF2 expression tends to induce the positive correlation (conditional to low ADPRT expression).

Similarly, The modulation by ZFP36L1 is shown in Fig 1 D (a, b) so that data-points in Fig 1 B (b) were divided into groups with low and high values of ZFP36L1. ZFP36L1 expression is facilitatory to ADPRT and acts to strengthen the negative correlation between XBP-1 and IGHM (conditional to high ADPRT expression).

The fifth gene was then searched by IPIG in two conditions, namely $(X_{3*} = 0, X_{4*} = 1)$, where X_{4*} =TM4SF2, and $(X_{3*} = 1, X_{4*} = 1)$, where X_{4*} =ZFP36L1. We then obtained AF1Q and DFKZp586C1019, respectively; the corresponding figures are Fig 1 E and F. We can make observations on the modulation of each gene, similarly to the above. AF1Q expression tends to cause a stronger positive correlation, and DFKZp586C1019 expression a stronger negative correlation, between XBP-1 and IGHM (also see below).

(Figure 2 is around here)

Figure 2 summarizes the relation of genes found by IPIG with respect to their modulation on the interaction between XBP-1 and IGHM. This diagram should not be taken rigidly but should be considered as a guide for further biological examination.

We don't have a complete biological explanation for the relationship of genes found by IPIG and the two genes but here describe several fragmented but relevant pieces of biological information, referring to some observations on Fig 1.

First note in Fig 1 B (b) that a cloud of points in the upper-left corner (i.e. with low XBP-1 and high IGHM) seems to induce the negative correlation, conditional to high ADPRT expression. This may suggest that other pathways may lead to high IGHM expression independent of XBP-1. Interestingly, these points in the upper-left corner are mostly preserved when ZFP36L1, which is identified in the next iteration of IPIG, is highly expressed (Fig 1 D b). Then, ZFP36L1 may be a candidate gene that regulates IGHM expression independent of XBP-1. ZFP36L1 is a gene that has zinc finger domain, and its putative role is a transcription factor regulating the response to growth factors and cytokines. Interactions between ZFP36L1 and XBP-1 / IGHM / ADPRT is unknown, but it is reported that this gene plays a role in B-cell apoptosis and proliferation. Furthermore, the points in the upper-left corner are again preserved, when the DFKZp586C1019, identified in the next iteration of IPIG, is highly expressed (Fig 1 F b). The function of the DFKZp586C1019 is unknown and the nucleotide-nucleotide BLAST search gave no homologue. Yet, DFKZp586C1019 may work as a positive regulatory factor to the ZFP36L1.

Second, going back to Fig 1 B, we observe that low ADPRT expression leads to a positive correlation between XBP-1 and IGHM. There is no report of ADPRT regulation on XBP-1 to the authors' knowledge, but ADPRT-dependent silencing of transcription factors including CREB has been reported (Oei et al., 1998). Since XBP-1 is a CREB-like transcription factor, ADPRT-dependent silenc-

[8] We emphasize that the actual calculation to get the θ-coordinates is done by using the $[0, 1]$-bounded values of X_{3*}, i.e. ADPRT. See Sec 2.6.

[9] We emphasize that all data-points in Fig 1 A are used in each iteration of IPIG. The visualization in Fig 1, showing only subsets of data-points with the corresponding gene values in each figure, is only for presentation simplicity. In each figure, the correlation is computed only with the points in the figure.

ing of XBP-1 may occur, i.e. ADPRT expression may suppress functions of XBP-1. Then the positive correlation appeared under low ADPRT expressions (Fig 1 B a), whereas this suppression is involved in the negative correlation under high ADPRT (Fig 1 B b) given other pathways discussed above.

TM4SF2 is found to modulate the correlation between XBP-1 and IGHM under low ADPRT (Fig 1 C). AF1Q is then found as modulatory, conditional to low ADPRT and high TM4SF2 (Fig 1 E). Notably, the two correlations shown in Fig 1 E (a,b) are smaller than that in Fig 1 C (b). Furthermore the scattered points in Fig 1 C (b) are somewhat concentrated in the lower-left corner or in the upper-right corner. These suggest that the correlation in Fig 1 C (b) is possibly a false one, that is, it emerged simply due to combining the two set of points in the two corners. We note that IPIG may successfully discover such a false correlation, by inspecting higher-order interactions. This kind of observation cannot be made if we inspect only the pair-wise correlation, as hierarchical clustering does. TM4SF2 is a gene for transmembrane protein, belonging to the Transmembrane 4 Superfamily, and associates with various surface molecules to build a network of molecular interactions (Hemler, 2001). AF1Q is also a transmembrane protein. Thus, TM4SF2 and AF1Q may work together but their exact co-function is unknown.

(Figure 3 is around here)

Remark 1. Gene data sorted by hierarchical clustering is shown in Fig 3. 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 (see Discussion).

Remark 2. One concern in using IPIG is the stability, that is, a question of how IPIG behaves if the data is somewhat perturbed. We show a preliminary result on this issue. We have examined how IPIG would choose the third gene, when a number of data points is removed from the original 327 samples, up to 5% removal (i.e, the removal of 16 data points). The result is shown in Fig 4 (see the legend for the details of this examination). Only two genes were chosen as the third gene in all conditions. ADPRT, which is the gene chosen by IPIG with the full sample, was chosen in almost all conditions. TM4SF2 started to occupy 3 % around 5 points removal and became 32.7 % with 16 points removal. No other genes were chosen in any conditions. This result indicates that IPIG is to an extent 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 and in that only two genes, including the ADPRT, were selected. Further discussion is given in the next section.

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 gene interaction, where gene expressions are given by real values, to the information geometric measure of a binary random vector, or the θ-coordinates (and the η-coordinates). This re-representation is very simple, one of the strengths of our method. Using the properties of this dual orthogonal coordinates under the weak definition of conditional independence, we can investigate the fine structure of gene interaction. In particular, we proposed an iterative procedure, called IPIG, to investigate hidden causes (i.e. other genes) for the interaction of two genes of interest. IPIG is useful in discovering a gene interaction hidden in data and selecting candidate genes for further biological investigation. Using dataset of ALL, we demonstrated the validity of IPIG.

Let us discuss the relation between our approach and two other related methods, namely hierarchical clustering and graphical models (GMs)[10]. In general, an approach with the stronger assumption can treat a larger number of variables, however, only with the more lmited probabilistic structure of the variables. Hierarchical clustering investigates only the pair-wise interaction, which basically corresponds to assuming the θ-coordinates of third-order and higher-order interaction as zero in our approach. It is attractive, with its low computational load, when a primary interest is to look over the interaction of as many genes as possible. However, it cannot discover the fine gene interactions treated by our approach. Both GMs and our approach address the high-order interaction with a more computational load. GMs, however, use a stronger assumption (i.e. the strong definition of the conditional independence) than our approach (the weak definition). The GM framework is for a more general class of probability distributions. Our approach focuses on the probability of a binary random vector (but see below) and thereby fully utilizes its properties (e.g. θ-coordinates), so that it is easy to implement. Also, IPIG systematically provides a set of candidate genes important for the fine interaction. Taking these together, we consider that our approach is particularly suitable for finding and exploring the fine interaction of a relatively small set of genes (e.g. pre-screening of a biological test). Conversely, GMs are particularly suitable for making a prediction based on its estimate of a relatively large gene network (e.g. disease diagnosis). Finally, we just mention that the class of the model trivially becomes equivalent between GMs and our approach, if we add to GMs a sufficient number of latent variables, provided GMs treats a binary random vector.

Let us discuss the limitation of IPIG and its future extension. First, IPIG investigates the interaction and rigorously speaking, cannot distinguish cause and effect. For example, when we say that one gene is 'modulatory' on other two genes (e.g. Results section), it should be understood as indicating not that the gene is causative on the interaction of the other two but that the interaction is correlated with the degree of expression of the gene. To address a question of cause and effect, IPIG should be combined with some other approaches. Second, we represented gene interaction through the coordinates of binary random vector but in principle, this mapping cannot fully capture the 'true' interaction in original data. This is simply because the probability distribution of real values may have more dimensions (infinite in principle) than that of binary random vectors. It is possible to extend our method from the binary to k-discrete random variable case. The current binary model, however, may be sufficient for selecting genes of interest as pre-screening of biological test. Examination on this issue is needed. Third, IPIG is found to be to an extent susceptible to statistical fluctuation of data (see Remark 2 in Results). From a theoretical viewpoint, this fact motivates an exploration of robust statistical techniques in IPIG. From a practical viewpoint, the result in Fig 4 suggests that a modified IPIG (see the end of Sec 2.5) may be useful in practice, combined with a perturabation test. In Fig 4, we found that the TM4SF2 is another potential candidate for the third gene in addition to the ADPRT. On the other hand, with the full sample, after choosing ADPRT in the first iteration, the IPIG chose the TM4SF2 in the next iteration, as the forth gene *conditional to low ADPRT*. These two facts together suggest that it may be worth to also examine the TM4SF2 in the case of high ADPRT. In other words, this suggest using a modified IPIG, which means considering both two genes modulatory before proceeding to the next iteration and which departs from the original IPIG that is strictly iterative. Finally, IPIG is only one instantiation of using the information geometric measure for DNA

[10] Here, we mostly discuss on GMs with undirected graphs.

microarray data analysis. We must further explore its possibilities.

Acknowledgments

The authors are grateful for the comments of reviewers. HN thanks J. Read for her comments on the draft.

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.

N. R. Draper and H. Smith. 1998. *Applied Regression Analysis.* John-Wiley & Sons, New York.

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.

Martin E. Hemler. 2001. Specific tetraspanin functions. *J. Cell Biol.*, 155(7):1103–1108.

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

H. Nakahara and S. Amari. 2002a. 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. 2002b. Information geometric measure for neural spikes. *Neural Computation*, 14(10):2269–2316.

H. Nakahara, S. Nishimura, M. Inoue, G. Hori, and S. Amari. 2002. Log linear model and gene interaction in dna microarray data. Technical report, as RIKEN BSI BSIS Tech Report No02-2 avaiable at www.mns.brain.riken.go.jp/ nakahara/papers/TR_IGDNA.ps, TR_IGDNA.pdf.

S. L. Oei, J. Griesenbeck, M. Schweiger, and M. Ziegler. 1998. Regulation of rna polymerase ii-dependent transcription by poly(adp-ribosyl)ation of transcription factors. *Journal of Biological Chemistry*, 273(48):31644–31647.

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.

A. M. Reimold, N. N. Iwakoshi, J. Manis, P. Vallabhajosyula, E. Szomolanyi-Tsuda, E. M. Gravallese, D. Friend, M. J. Grusby, F. Alt, and L. H. Glimcher. 2001. Plasma cell differentiation requires the transcription factor xbp-1. *Nature*, 412(6844):300–307.

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

E.-J. Yeoh, M. E. Ross, S. A. Shurtleff, W. K. Williams, D. Patel, R. Mahfouz, F. G. Behm, S. C. Raimondi, M. V. Relling, A. Patel, C. Cheng, D. Campana, D. Wilkins, X. Zhou, J. Li, H. Liu, C.-H. Pui, W. E. Evans, C. Naeve, L. Wong, and J. R. Downing. 2002. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell*, 1(2):133–143.

Figure legends

Figure 1: IPIG gives decomposition of the interaction between two genes, XBP-1 and IGHM.

Figure 2: Scheme drawn from the results in Fig 1. The arrow and flat heads indicate the positive and negative influence onto the gene (and/or the gene interaction), respectively.

Figure 3: 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).

Figure 4: 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.

Figure 1

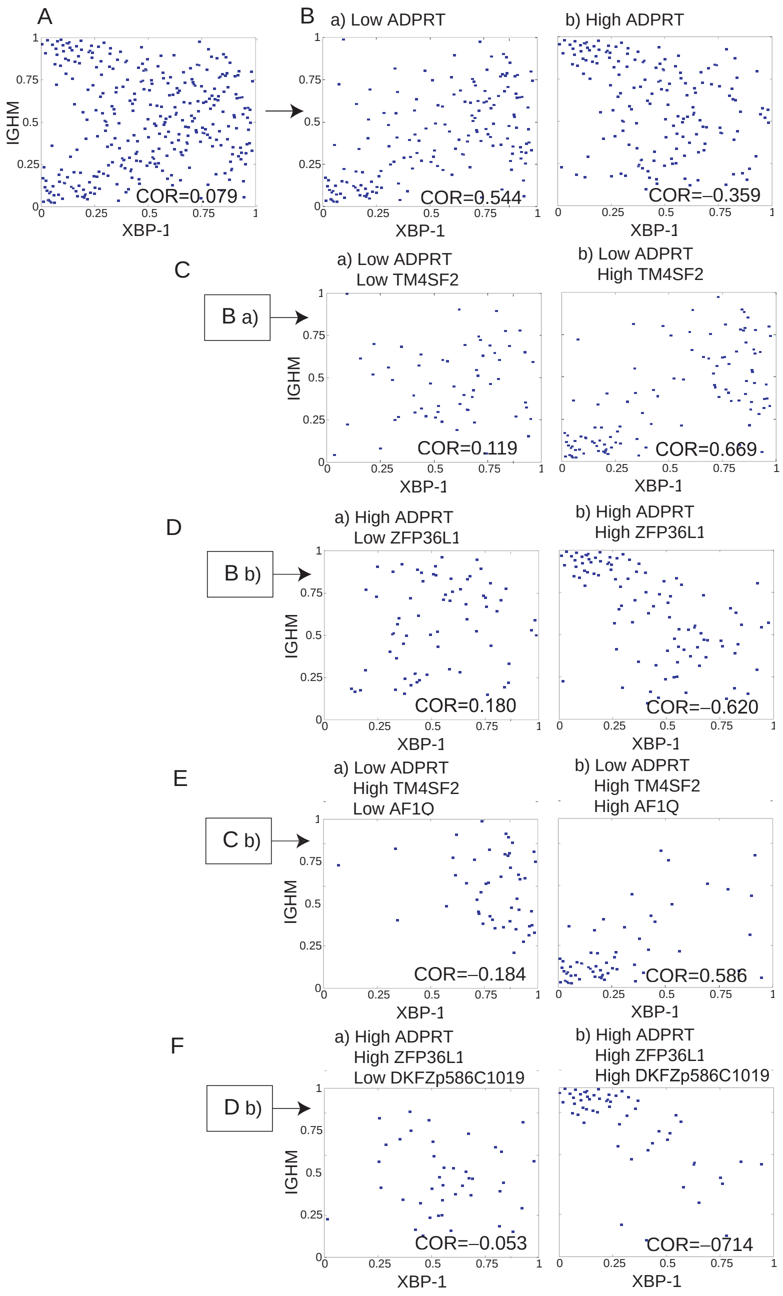

Figure 1

Figure 2

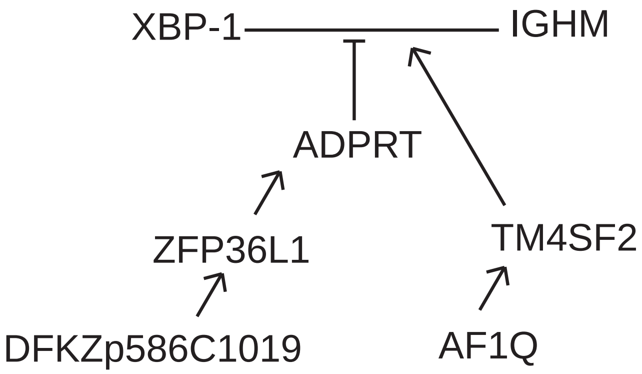

Figure 2

Figure 3

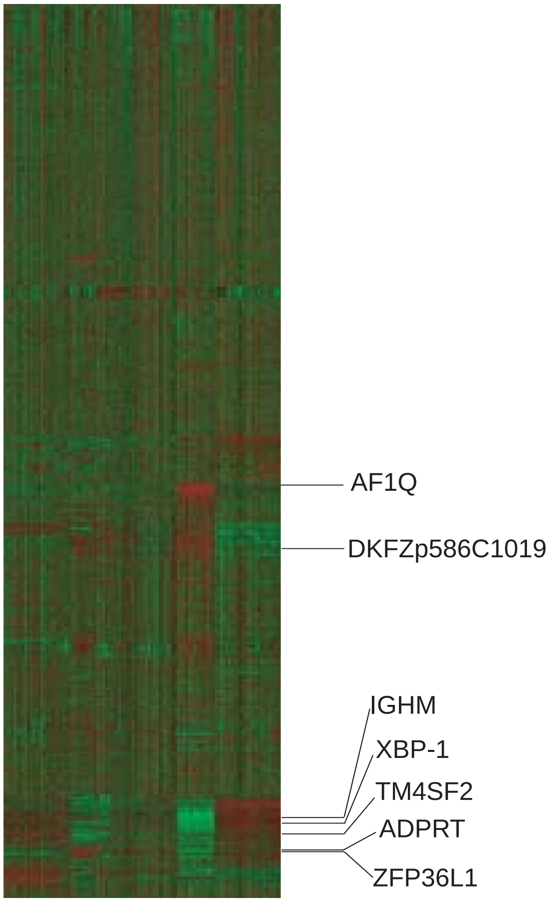

Figure 3

Figure 4

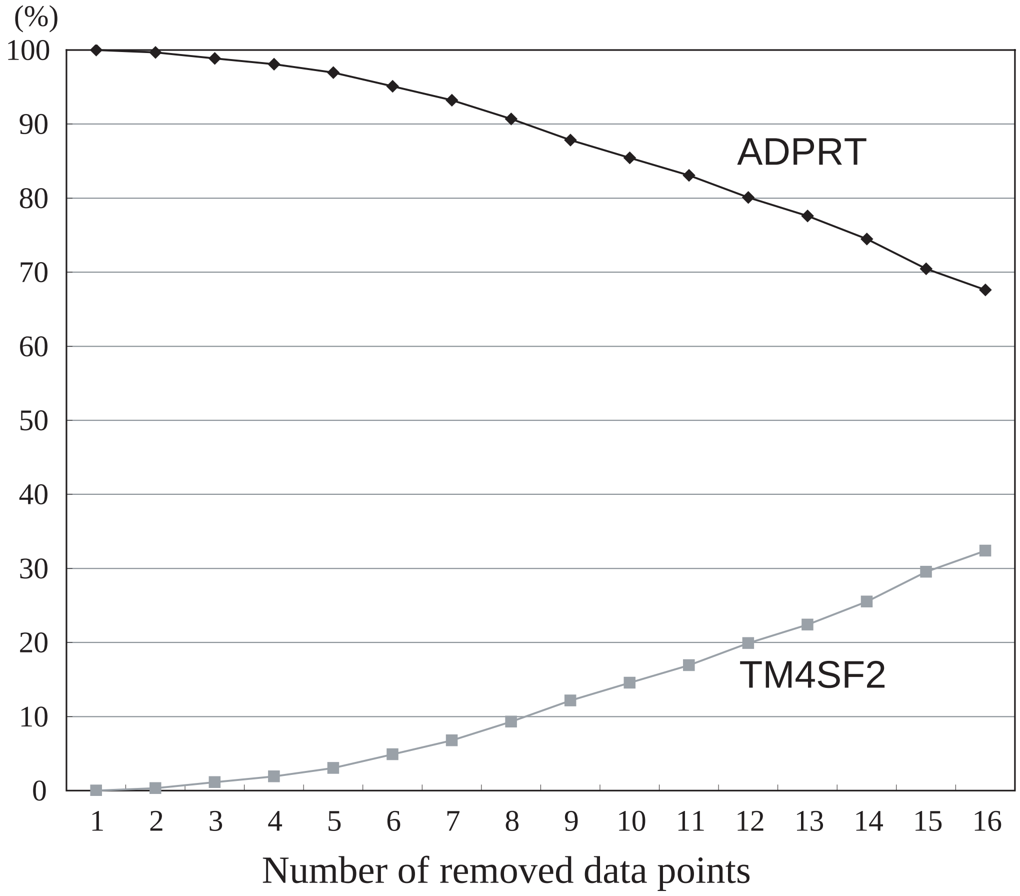

Figure 4