Mutual Information Estimation Reveals Global Associations between Stimuli and Biological Processes

Taiji Suzuki (s-taiji@stat.t.u-tokyo.ac.jp)
Department of Mathematical Informatics, The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan

Masashi Sugiyama (sugi@cs.titech.ac.jp)
Department of Computer Science, Tokyo Institute of Technology
2-12-1, O-okayama, Meguro-ku, Tokyo, 152-8552, Japan

Takafumi Kanamori (kanamori@is.nagoya-u.ac.jp)
Department of Computer Science and Mathematical Informatics,
Nagoya University
Furocho, Chikusaku, Nagoya 464-8603, Japan

Jun Sese (sesejun@is.ocha.ac.jp)
Department of Information Science, Ochanomizu University
2-1-1 Ohtsuka, Bunkyo-ku, Tokyo 112-8610, Japan

Abstract

Background: Although microarray gene expression analysis has become popular, it remains difficult to interpret the biological changes caused by stimuli or variation of conditions. Clustering of genes and associating each group with biological functions are often used methods. However, such methods only detect partial changes within cell processes. Herein, we propose a method for discovering global changes within a cell by associating observed conditions of gene expression with gene functions.

Results: To elucidate the association, we introduce a novel feature selection method called *Least-Squares Mutual Information (LSMI)*, which computes mutual information without density estimaion, and therefore LSMI can detect nonlinear associations within a cell. We demonstrate the effectiveness of LSMI through comparison with existing methods. The results of the application to yeast microarray datasets reveal that non-natural stimuli affect various biological processes, whereas others are no significant relation to specific cell processes. Furthermore, we discover that biological processes can be categorized into four types according to the responses of various stimuli: DNA/RNA metabolism, gene expression, protein metabolism, and protein localization.

Conclusions: We proposed a novel feature selection method called LSMI, and applied LSMI to mining the association between conditions of yeast and biological processes through microarray datasets. In fact, LSMI allows us to elucidate the global organization of cellular process control.

Background

Advances in microarray technologies enable us to explore the comprehensive dynamics of transcription within a cell. The current problem is to extract useful information from a massive dataset. The primarily used approach is clustering. Cluster analysis reveals variations of gene expression and reduces the complexity of large datasets. However, additional methods are necessary to associate genes in each cluster with genetic function using GO term finder [3], or to understand stimuli related to specific cellular status.

However, these clustering-association strategies cannot detect global cell status changes because of the division of clusters. Some stimuli activate a specific pathway, although others might change overall cellular processes. Understanding the effect of stimuli in cellular processes directly, in this paper, we introduce a novel feature selection method called Least-Squares Mutual Information (LSMI), which selects features using mutual information without density estimation. Mutual information has been utilized to measure distances between gene expressions [15]. To compute the mutual information in existing methods, density estimation or discritization is required. However, the estimation of gene expression is difficult because we have little knowledge about density function of gene expression profile. LSMI offers an analytic-form solution and avoid the estimation.

Feature selection techniques are often used in gene expression analysis [16]. Actually, LSMI has three advantages compared to existing methods: capability of avoiding density estimation which is known to be a hard problem [18], availability of model selection, and freedom from a strong model assumption. To evaluate the reliability of ranked features using LSMI, we compare receiver operating characteristic (ROC) curves [14] to those of existing methods: kernel density estimation (KDE) [20, 6], k-nearest neighbor (KNN) [13], Edgeworth expansion (EDGE) [11], and Pearson correlation coefficient (PCC). Thereby, we certify that our method has better performance than the existing methods in prediction of gene functions about biological processes. This fact implies that features selected using our method reflect biological processes.

Using the ranked features, we illustrate the associations between stimuli and biological processes according to gene expressions. Results show that stimuli damage essential processes within a cell, causing association with some cellular processes. From the response to stimuli, biological processes are divisible into four categories: DNA/RNA metabolic processes, gene expression, protein metabolic processes, and protein localization.

Results

Approach—Mutual Information Detection

In this study, we detect underlying dependencies between gene expressions obtained by groups of stimuli and gene functions. The dependencies are studied in various machine learning problems such as feature selection [8, 22] and independent component analysis [4]. Although classical correlation analysis would be useful for these problems, it cannot detect nonlinear dependencies with no correlation. On the other hand, *mutual information* (MI),

which plays an important role in information theory [5], enables us to detect general nonlinear dependencies. Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be a set of gene expressions and a set of known gene functions. A variant of MI based on the squared loss is defined by

$$I_s(X,Y) := \iint \left(\frac{p_{\mathrm{xy}}(\boldsymbol{x},\boldsymbol{y})}{p_{\mathrm{x}}(\boldsymbol{x})p_{\mathrm{y}}(\boldsymbol{y})} - 1 \right)^2 p_{\mathrm{x}}(\boldsymbol{x})p_{\mathrm{y}}(\boldsymbol{y}) d\boldsymbol{x} d\boldsymbol{y}. \quad (1)$$

Note that I_s vanishes if and only if $\boldsymbol{x}$ and $\boldsymbol{y}$ are independent. The use of MI allows us to detect no correlation stimulus with a specific gene function or process.

Estimating MI is known to be a difficult problem in practice [22, 13, 11]. Herein, we propose LSMI, which does not involve density estimation but directly models the *density ratio*:

$$w(\boldsymbol{x},\boldsymbol{y}) := \frac{p_{\mathrm{xy}}(\boldsymbol{x},\boldsymbol{y})}{p_{\mathrm{x}}(\boldsymbol{x})p_{\mathrm{y}}(\boldsymbol{y})}.$$

Given a density ratio estimator $\widehat{w}(\boldsymbol{x},\boldsymbol{y})$, squared loss MI can be simply estimated by

$$\widehat{I}_s(X,Y) = \frac{1}{n^2} \sum_{i,j=1}^{n} (\widehat{w}(\boldsymbol{x}_i,\boldsymbol{y}_j) - 1)^2.$$

Mathematical definitions related to LSMI are provided in the Methods section. LSMI offers an analytic-form solution, which allows us to estimate MI in a computationally very efficiently manner.

It is noteworthy that $\boldsymbol{x}$ includes a multi-dimensional vector. In fact, LSMI can handle a group of stimuli, although generic correlation indices such as Pearson correlation between parameters and target value are calculated independently. Therefore, we can elucidate which type of stimulus has no dependency to biological processes using LSMI.

Datasets and Feature Selection

In this section, we first prepare datasets to show the association between stimuli and biological process, and introduce feature selection using the datasets.

Biological Process

We compute mutual information between gene expression values grouped by stimuli and class of genes' biological processes. As the class, we use biological process terms in Gene Ontology (GO) categorization [1]. We select GO terms associated with more than 800 and less than 2,000 genes because terms having a small number of genes only describe a fraction of the cell status, whereas terms having a large number of genes indicate functions associated with almost all genes in yeast. Actually, GO has a directed acyclic graph (DAG) structure, and each term has child terms. The GO terms are classified into three categories; we use only biological process terms to identify the changes within a cell. Using this method, we select 12 GO terms.

Gene Expression Profiles

The gene expression profile is the best comprehensive dataset to associate stimuli and biological processes. We use two different microarray datasets. One is of 173 microarray data under stress conditions of various types [7]. We categorize the 173 stress conditions into 29 groups based on the type of condition such as heat shock, oxidizing condition, etc. The other is of 300 microarray data under gene-mutated conditions [10]. We categorize the genes into 146 groups based on associated GO terms. We use only the GO terms which are associated with 1,500 genes or fewer. We also use child terms on a GO layered structure if the term has more than 1,200 genes. When one gene belongs to multiple GO terms, we classify the gene into the the classification whose number of associated genes is smallest.

In both profiles, we remove genes whose expression values are obtained from fewer than 30% of all observed conditions. All missing values are filled out by the average of all the expression values.

Feature Selection using LSMI

We use a novel feature selection method called LSMI, which is based on MI, to associate stimuli with cellular processes. Here we consider the *forward* feature-group addition strategy, i.e., a feature-group score between each input feature-group and output cellular process is computed. The top m feature-groups are used for training a classifier. We predict 12 GO terms independently. We randomly choose 500 genes from among $6,116$ genes on the stress condition dataset for feature-group selection and for training a classifier; the rest are used for evaluating the generalization performance. For using the gene-mutated expression dataset, we select 500 genes from among $6,210$ genes. We repeat this trial 10 times. For classification, we use a Gaussian kernel support vector machine (GK-SVM) [18], where the kernel width is set at the median distance among all samples and the regularization parameter is fixed at $C = 10$. We explain the efficiency of feature selection of LSMI in the Discussion section.

Results

The association between stress conditions and biological processes in GO terms is shown in Fig.1. Each row and column respectively indicate a group of conditions and a GO term. Row and column dendrograms are clustering results by the Ward method according to cell values. Each cell contains an average ranking over 10 trials by LSMI. The red cell denotes that the parameter has a higher rank; that is, the parameter has association with the target GO term. A blue cell denotes that the parameter has a lower rank.

As shown in this figure, conditions are divided into two groups. Almost all conditions in the upper cluster have higher rank, whereas those in a lower cluster have higher rank only under specific conditions. The conditions in the upper cluster include strong heat shocks, dithiothreitol (DTT) exposure, nitrogen depletion, and diamide treatments, which

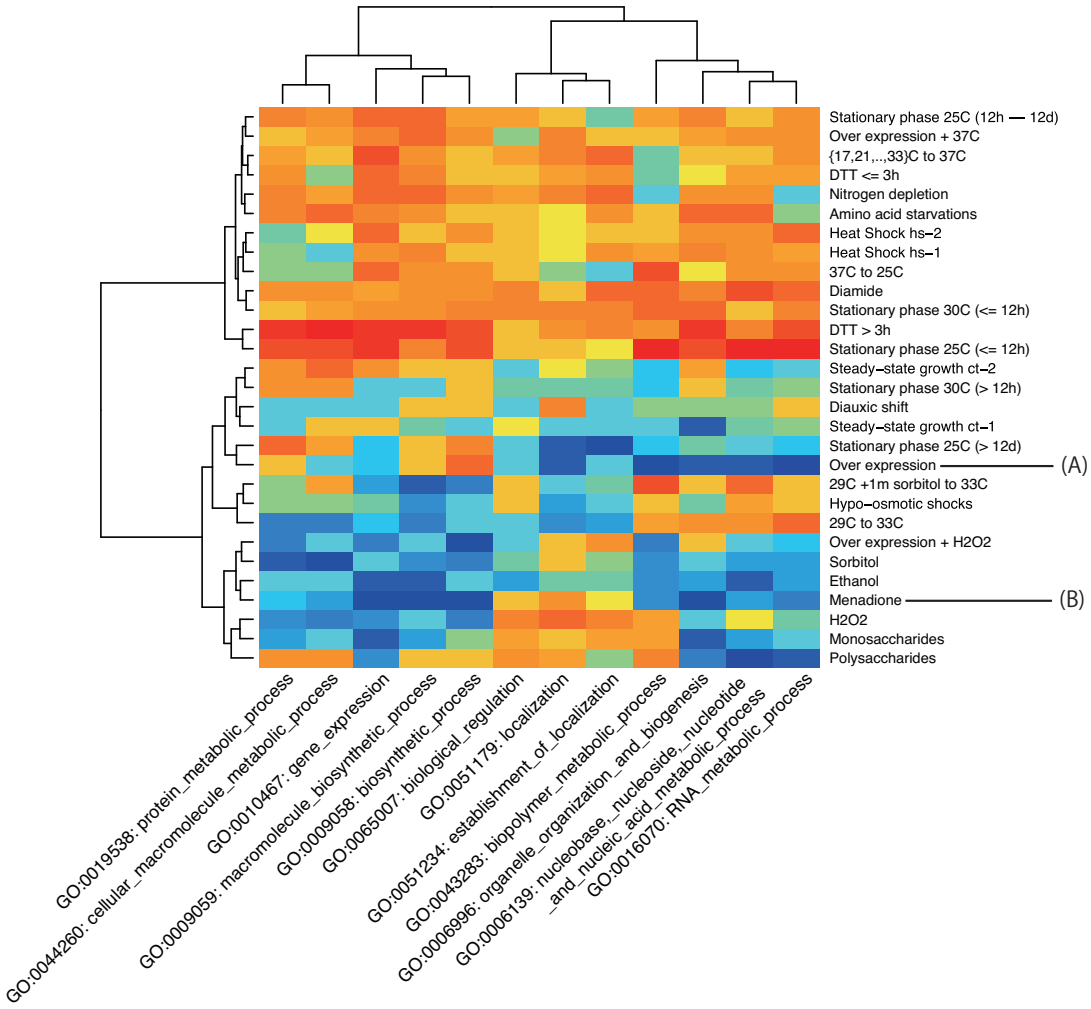


Figure 1: Matrix of stress conditions (rows) versus biological processes (columns). Red cells have higher correlation.

are non-natural conditions. The result reveals that non-natural conditions change overall cellular processes.

The GO term clusters are divided into three groups: DNA/RNA metabolism (right), localization of protein (middle), and others (left). The leftmost cluster contains bio synthesis, gene expression process, and protein metabolic process. From this figure, nucleic acid metabolism processes are inferred to be independent from amino acid metabolism processes. We will confirm the independence and consider the division of clusters by using other dataset later.

We herein investigate the details of difference among DNA metabolic process, protein metabolic process and localization of proteins. Under an overexpression condition indicated by sign (A) in Fig.1, DNA/RNA metabolisms show no correlation with expressions

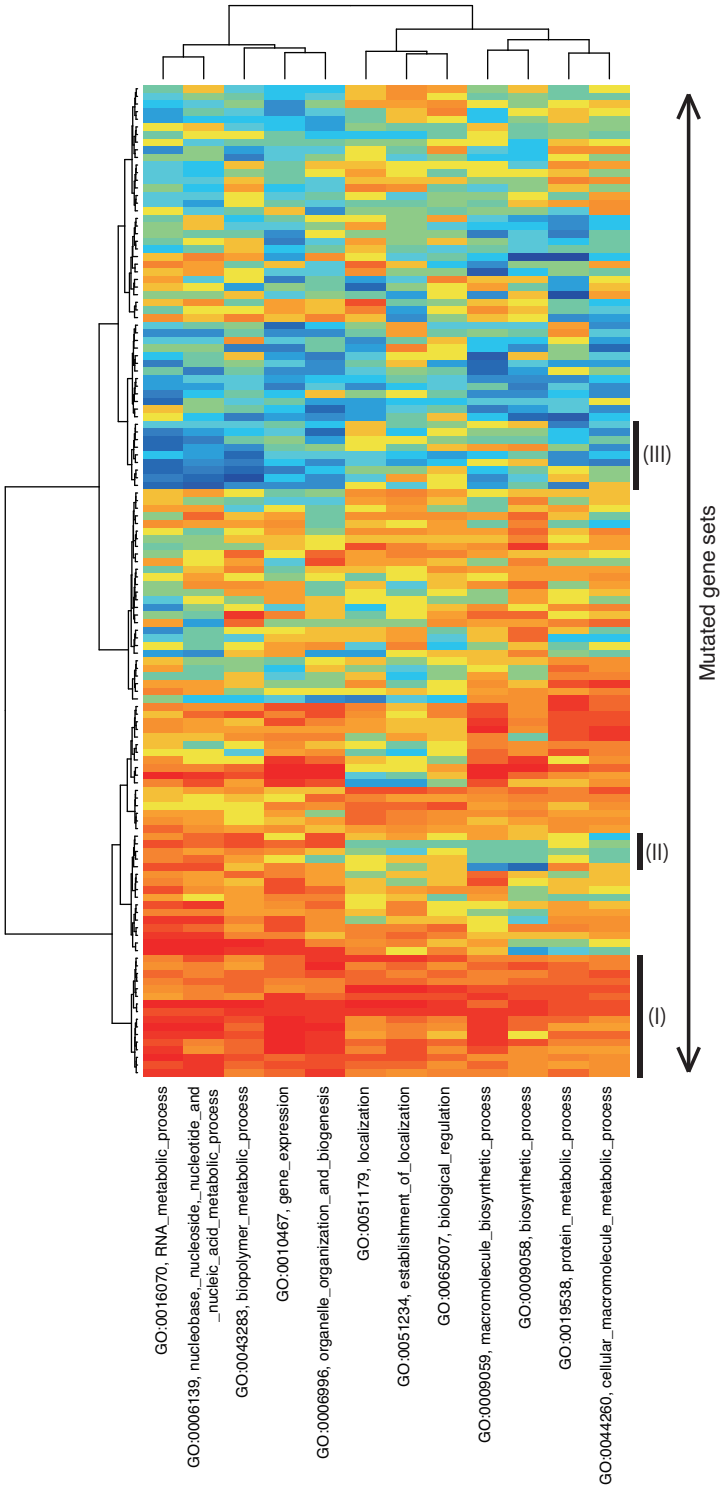


Figure 2: Overview: a matrix of mutated gene groups (rows) versus biological processes (columns)

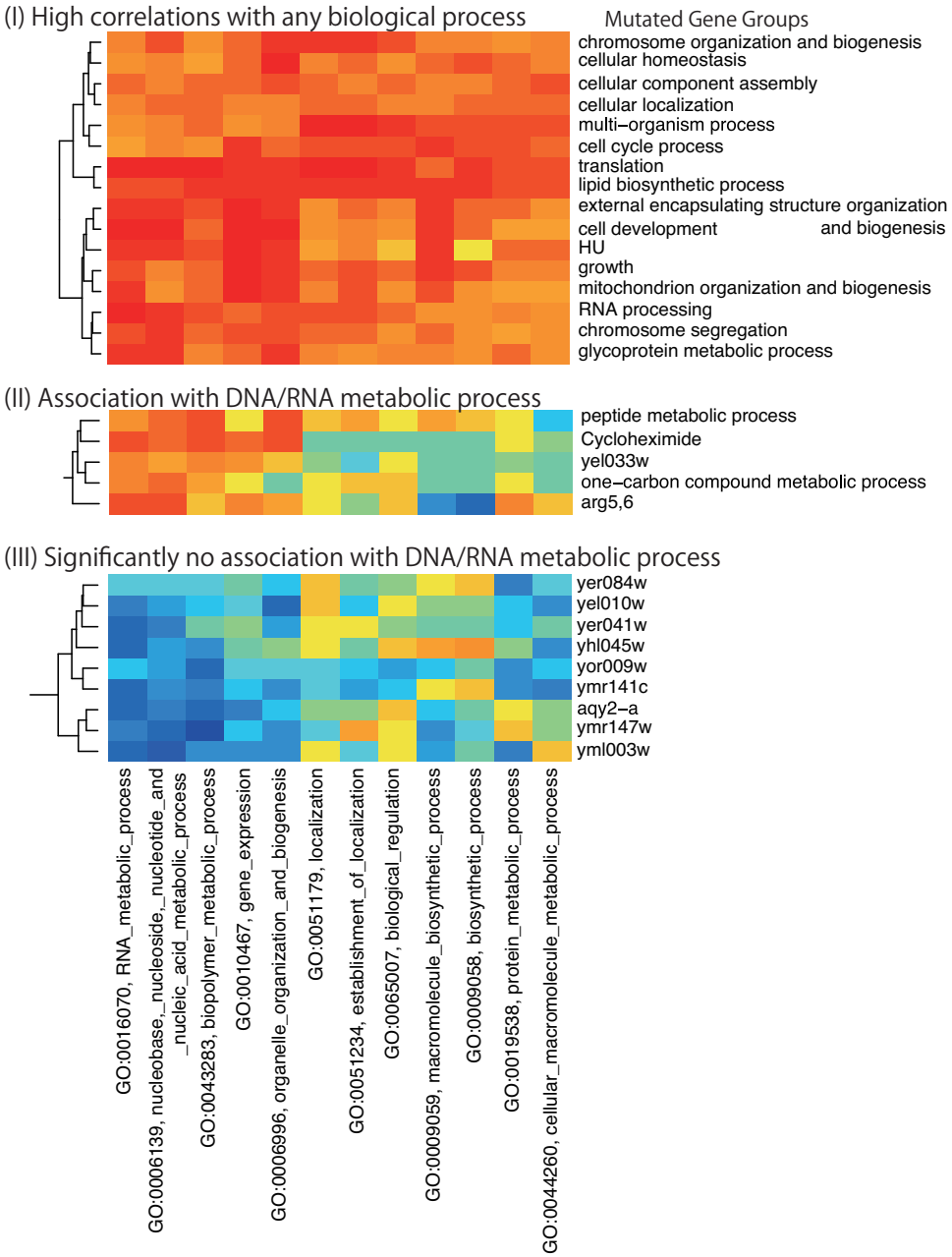


Figure 3: Sub-matrices of the full map

of genes belonging to over-expression genes. This finding of no correlation is one advantage of LSMI. The menadione (vitamin K) exposure condition indicated by (B) in Fig.1 is associated with localization of proteins. Menadione supplementation causes high toxicity; such toxicity might result from the violation of protein localizations.

Next, we compute the association using expressions of gene mutants. The results are shown in Fig.2. The stimulus can be categorized into two parts: high association under almost all processes and under particular conditions. The division is the same because

of stress condition associations. The GO terms also categorize three parts: DNA/RNA metabolic processes, protein metabolic processes, and localization. In this experiment, GO terms "gene expression" (GO:0010467) and "organelle organization and biogenesis" (GO:0006996) are in the DNA/RNA metabolic process cluster, although they are classified in protein metabolic processes cluster under stress conditions in Fig.1. Because the both divisions are close to ancestor division, we can conclude that the cluster about gene expression exists. From these results, GO terms are divisible into four categories: DNA/RNA metabolic process, protein metabolic process, localization, and gene expression.

In Fig.3, we present details of three clusters in Fig.2. In fact, Fig.3(I) presents a cluster whose members are correlated with any biological process. Furthermore, the functions of the mutated genes are essential processes for living cells, such as cellular localization, cell cycle, and growth. This result might indicate that the upper half stimulus in Fig.1 destroys the functions of these essential genes. Furthermore, Fig.3(II) includes the groups of genes associated with DNA/RNA metabolic processes. In this cluster, YEL033W/MTC1 is a gene with unknown function and is predicted to have a metabolic role using protein–protein interaction [17]. Our clustering result indicates that YEL033W would have some relation with metabolism, especially methylation (methylation is an important part of the one-carbon compound metabolic process). We show genes which have no significant association with DNA/RNA metabolic processes in Fig.3(III). In the cluster, all genes except AQY2 are of unknown function. No correlation clusters cannot be found by existing methods. Our result might provide clues to elucidate these genes' functions.

Discussion

A common analytical flow of the expression data is first clustering and then associating clusters with GO terms or pathways. Although clustering reduces the complexity of large datasets, the strategy might fail to detect changes of entire genes within a cell such as metabolic processes.

To interpret such gene expression changes, gene set enrichment analysis [21] has been proposed. This method treats microarrays independently. Therefore, housekeeping genes are often ranked highly. When gene expressions under various conditions are available, our method would show us the better changes of cellular processes because of the comparison between groups of conditions. The module map [19] gives a global association between a set of genes and a set of conditions. However, this method requires important changes of gene expressions because it uses hypergeometric distributions to compute correlations. Our correlation index is based on MI. Therefore, we can detect nonlinear dependencies with no correlation. An example is depicted in Fig.3(III).

The characteristics of LSMI and existing MI estimators are presented in Table 1. Detail comparisons are described in the Methods section. The *kernel density estimator* (KDE) [20, 6] is distribution-free. Model selection is possible by likelihood cross-validation (LCV). However, a hard task of density estimation is involved. Estimation of the *entropies*

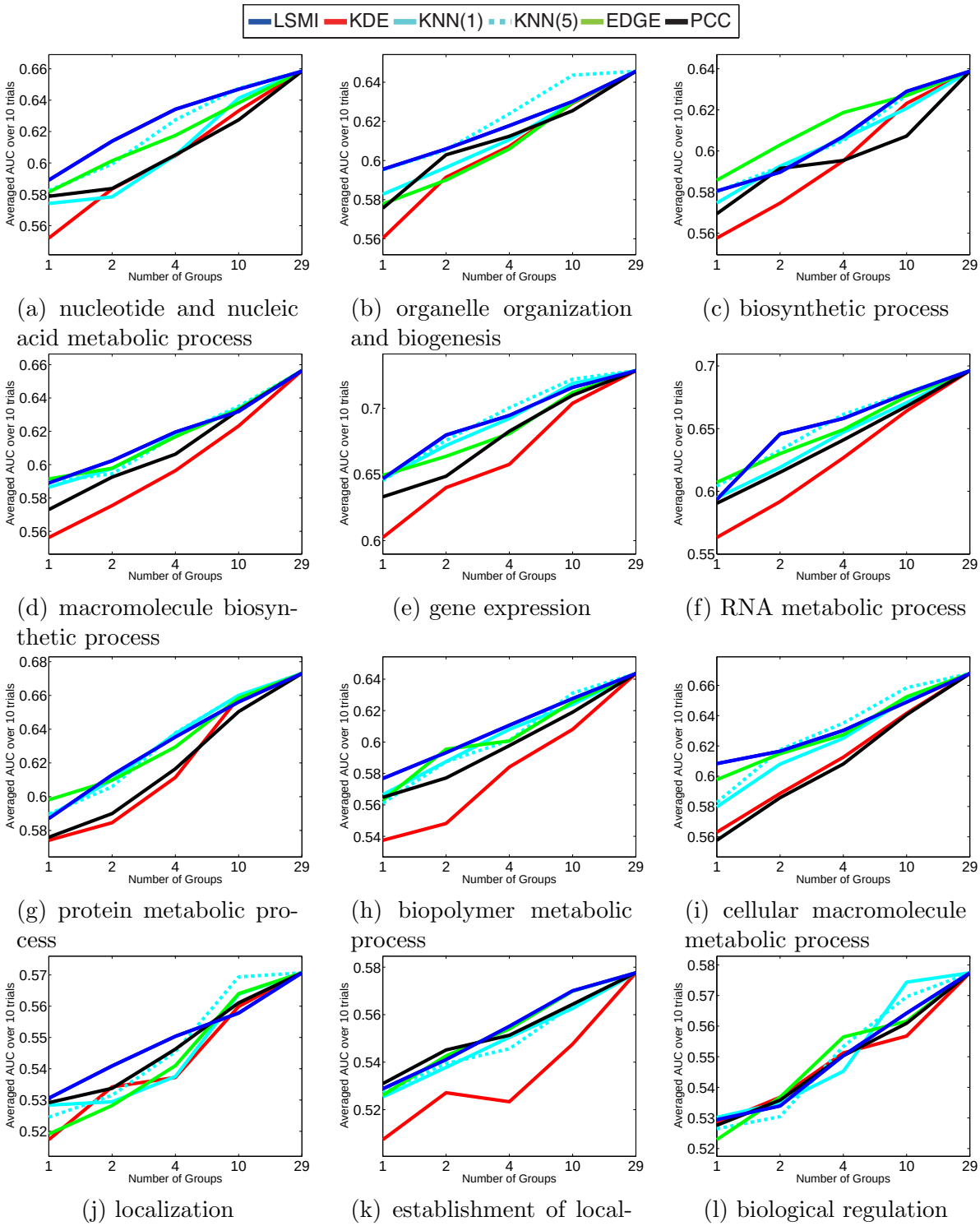
Table 1: Relation between existing and proposed MI estimators. If the order of the Edgeworth expansion is regarded as a tuning parameter, model selection of EDGE is expected to be 'Not available'.

	Density estimation	Model selection	Distribution
KDE	Involved	**Available**	**Free**
KNN	**Not involved**	Not available	**Free**
EDGE	**Not involved**	**Not necessary**	Nearly normal
LSMI	**Not involved**	**Available**	**Free**

using k-nearest neighbor (KNN) samples [13] is distribution-free and does not involve density estimation directly. However, no model selection method exists for determining the number of nearest neighbors. Edgeworth expansion (EDGE) [11] does not involve density estimation or any tuning parameters. However, it is based on the assumption that the target distribution is close to the normal distribution. On the other hand, LSMI is distribution-free; it involves no density estimation, and model selection is possible by cross-validation (CV). Therefore, LSMI overcomes limitations of the existing approaches. Within a cell, most processes have a nonlinear relation such as enzyme effects and feedback loops. The lack of one advantage might cause difficulty of application to biological datasets. By virtue of these advantages, LSMI can detect correlation or independence between features of complex cellular processes.

To investigate the efficiency of feature selection, we compare areas under the curve (AUCs) with LSMI (CV), KDE(LCV), KNN(k) for $k = 1, 5$, EDGE, and PCC. Details of these methods are described in the Methods section. Figure 4 depicts AUCs for 12 GO term classifications. The x-axis shows the number of stimulus groups used for the prediction. The y-axis means averaged AUC over 10 trials, where AUCs are calculated as the area under the receiver operating characteristic (ROC) curve, which is often used for diagnostic tests. Each figure shows AUC curves calculated using the six methods.

In the AUC figures, the higher curves represent better predictions. For example, Fig.4(a) shows that LSMI is the highest position, which means that LSMI achieves the best performance among the six methods. In Figs.4(b) and 4(d), KNN(1) and KNN(5), which are denoted by the light blue and dotted light blue lines, have the best performance. However, in Figs.4(i), 4(j) and 4(l), averaged AUCs of KNN using numerous groups are high, whereas the AUCs using small and few groups are low. No systematic model selection strategies exist for KNN and therefore KNN would be unreliable in practice. Figure 4(c) depicts that EDGE, which is indicated by the light green line, has the highest AUC. In fact, EDGE presumes the normal distribution. Consequently, it works well only on a few datasets. From these figures, LSMI indicated by the blue line appears to be the best feature selection method.



(a) nucleotide and nucleic acid metabolic process

(b) organelle organization and biogenesis

(c) biosynthetic process

(d) macromolecule biosynthetic process

(e) gene expression

(f) RNA metabolic process

(g) protein metabolic process

(h) biopolymer metabolic process

(i) cellular macromolecule metabolic process

(j) localization

(k) establishment of localization

(l) biological regulation

Figure 4: Classification error against the number of feature groups for the yeast cell datasets.

Conclusions

We provided a global view of the associations between stimuli and changes of biological processes based on gene expression profiles. The association is generally difficult to use for making models because of nonlinear correlation. To cope with this problem, we introduced a novel feature selection method called *LSMI*, which uses MI and can be computed efficiently. In comparison to other feature selection methods, LSMI showed better AUCs in prediction of biological process functions. Consequently, our feature selection results would be more reliable than those obtained using the other methods. We calculated the association between stimuli and GO biological process terms using gene expression profiles. The result revealed that the stimuli are categorized into four types: related to DNA/RNA metabolic process, gene expression, protein metabolic process, and protein localization. LSMI enabled us to reveal the global regulation of cellular processes from comprehensive transcription datasets.

Methods

Mutual Information Estimation

A naive approach to estimating MI is to use a KDE [20, 6], i.e., the densities $p_{\mathrm{xy}}(\boldsymbol{x}, \boldsymbol{y})$, $p_{\mathrm{x}}(\boldsymbol{x})$, and $p_{\mathrm{y}}(\boldsymbol{y})$ are separately estimated from samples and the estimated densities are used for computing MI. The band-width of the kernel functions could be optimized based on likelihood cross-validation (LCV) [9], so there remains no open tuning parameter in this approach. However, density estimation is known to be a hard problem [18] and therefore the KDE-based method may not be so effective in practice.

An alternative method involves estimation of entropies using KNN. The KNN-based approach was shown to perform better than KDE [12], given that the number k is chosen appropriately—a small (large) k yields an estimator with small (large) bias and large (small) variance. However, appropriately determining the value of k is not straightforward in the context of MI estimation.

Here, we propose a new MI estimator that can overcome the limitations of the existing approaches. Our method, which we call Least-Squares Mutual Information (LSMI), does not involve density estimation and directly models the *density ratio*:

$$w(\boldsymbol{x}, \boldsymbol{y}) := \frac{p_{\mathrm{xy}}(\boldsymbol{x}, \boldsymbol{y})}{p_{\mathrm{x}}(\boldsymbol{x}) p_{\mathrm{y}}(\boldsymbol{y})}. \tag{2}$$

The solution of LSMI can be computed by simply solving a system of linear equations. Therefore, LSMI is computationally very efficient. Furthermore, a variant of cross-validation (CV) is available for model selection, so the values of tuning parameters such as the regularization parameter and the kernel width can be adaptively determined in an objective manner.

A New MI Estimator

In this section, we formulate the MI inference problem as density ratio estimation and propose a new method of estimating the density ratio.

MI Inference via Density Ratio Estimation

Let $\mathcal{D}_{\mathrm{X}}$ ($\subset \mathbb{R}^{d_{\mathrm{x}}}$) and $\mathcal{D}_{\mathrm{Y}}$ ($\subset \mathbb{R}^{d_{\mathrm{y}}}$) be the data domains and suppose we are given n independent and identically distributed (i.i.d.) paired samples

$$\{(\boldsymbol{x}_i, \boldsymbol{y}_i) \mid \boldsymbol{x}_i \in \mathcal{D}_{\mathrm{X}},\ \boldsymbol{y}_i \in \mathcal{D}_{\mathrm{Y}}\}_{i=1}^{n}$$

drawn from a joint distribution with density $p_{\mathrm{xy}}(\boldsymbol{x}, \boldsymbol{y})$. Let us denote the marginal densities of $\boldsymbol{x}_i$ and $\boldsymbol{y}_i$ by $p_{\mathrm{x}}(\boldsymbol{x})$ and $p_{\mathrm{y}}(\boldsymbol{y})$, respectively. The goal is to estimate squared-loss MI defined by Eq.(1).

Our key constraint is that we want to avoid density estimation when estimating MI. To this end, we estimate the *density ratio* $w(\boldsymbol{x}, \boldsymbol{y})$ defined by Eq.(2). Given a density ratio estimator $\widehat{w}(\boldsymbol{x}, \boldsymbol{y})$, MI can be simply estimated by

$$\widehat{I}_s(X, Y) = \frac{1}{n^2} \sum_{i,j=1}^{n} (\widehat{w}(\boldsymbol{x}_i, \boldsymbol{y}_j) - 1)^2.$$

We model the density ratio function $w(\boldsymbol{x}, \boldsymbol{y})$ by the following linear model:

$$\widehat{w}_{\boldsymbol{\alpha}}(\boldsymbol{x}, \boldsymbol{y}) := \boldsymbol{\alpha}^{\top} \boldsymbol{\varphi}(\boldsymbol{x}, \boldsymbol{y}),$$

where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_b)^{\top}$ are parameters to be learned from samples, $^{\top}$ denotes the transpose of a matrix or a vector, and

$$\boldsymbol{\varphi}(\boldsymbol{x}, \boldsymbol{y}) = (\varphi_1(\boldsymbol{x}, \boldsymbol{y}), \varphi_2(\boldsymbol{x}, \boldsymbol{y}), \ldots, \varphi_b(\boldsymbol{x}, \boldsymbol{y}))^{\top}$$

are basis functions such that

$$\boldsymbol{\varphi}(\boldsymbol{x}, \boldsymbol{y}) \geq \mathbf{0}_b \quad \text{for all} \quad (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{D}_{\mathrm{X}} \times \mathcal{D}_{\mathrm{Y}}.$$

$\mathbf{0}_b$ denotes the b-dimensional vector with all zeros. Note that $\boldsymbol{\varphi}(\boldsymbol{x}, \boldsymbol{y})$ could be dependent on the samples $\{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^{n}$, i.e., *kernel* models are also allowed. We explain how the basis functions $\boldsymbol{\varphi}(\boldsymbol{x}, \boldsymbol{y})$ are chosen in the later section.

A Least-squares Approach to Direct Density Ratio Estimation

We determine the parameter $\boldsymbol{\alpha}$ in the model $\widehat{w}_{\boldsymbol{\alpha}}(\boldsymbol{x}, \boldsymbol{y})$ so that the following squared error J_0 is minimized:

$$\begin{aligned} J_0(\boldsymbol{\alpha}) :=& \frac{1}{2} \iint (\widehat{w}_{\boldsymbol{\alpha}}(\boldsymbol{x}, \boldsymbol{y}) - w(\boldsymbol{x}, \boldsymbol{y}))^2 p_{\mathrm{x}}(\boldsymbol{x}) p_{\mathrm{y}}(\boldsymbol{y}) d\boldsymbol{x} d\boldsymbol{y} \\ =& \frac{1}{2} \iint \widehat{w}_{\boldsymbol{\alpha}}(\boldsymbol{x}, \boldsymbol{y})^2 p_{\mathrm{x}}(\boldsymbol{x}) p_{\mathrm{y}}(\boldsymbol{y}) d\boldsymbol{x} d\boldsymbol{y} - \iint \widehat{w}_{\boldsymbol{\alpha}}(\boldsymbol{x}, \boldsymbol{y}) p_{\mathrm{xy}}(\boldsymbol{x}, \boldsymbol{y}) d\boldsymbol{x} d\boldsymbol{y} + C, \end{aligned}$$

where $C = \frac{1}{2}\iint w(\boldsymbol{x},\boldsymbol{y})p_{\mathrm{xy}}(\boldsymbol{x},\boldsymbol{y})d\boldsymbol{x}d\boldsymbol{y}$ is a constant and therefore can be safely ignored. Let us denote the first two terms by J:

$$J(\boldsymbol{\alpha}) := J_0(\boldsymbol{\alpha}) - C = \frac{1}{2}\boldsymbol{\alpha}^\top \boldsymbol{H}\boldsymbol{\alpha} - \boldsymbol{h}^\top \boldsymbol{\alpha},$$

where

$$\boldsymbol{H} := \iint \boldsymbol{\varphi}(\boldsymbol{x},\boldsymbol{y})\boldsymbol{\varphi}(\boldsymbol{x},\boldsymbol{y})^\top p_{\mathrm{x}}(\boldsymbol{x})p_{\mathrm{y}}(\boldsymbol{y})d\boldsymbol{x}d\boldsymbol{y},$$
$$\boldsymbol{h} := \iint \boldsymbol{\varphi}(\boldsymbol{x},\boldsymbol{y})p_{\mathrm{xy}}(\boldsymbol{x},\boldsymbol{y})d\boldsymbol{x}d\boldsymbol{y}.$$

Approximating the expectations in $\boldsymbol{H}$ and $\boldsymbol{h}$ by empirical averages, we obtain the following optimization problem:

$$\widetilde{\boldsymbol{\alpha}} := \underset{\boldsymbol{\alpha}\in\mathbb{R}^b}{\operatorname{argmin}} \left[\frac{1}{2}\boldsymbol{\alpha}^\top \widehat{\boldsymbol{H}}\boldsymbol{\alpha} - \widehat{\boldsymbol{h}}^\top \boldsymbol{\alpha} + \lambda\boldsymbol{\alpha}^\top\boldsymbol{\alpha}\right], \tag{3}$$

where we included a regularization term $\lambda\boldsymbol{\alpha}^\top\boldsymbol{\alpha}$ and

$$\widehat{\boldsymbol{H}} := \frac{1}{n^2}\sum_{i,j=1}^{n} \boldsymbol{\varphi}(\boldsymbol{x}_i,\boldsymbol{y}_j)\boldsymbol{\varphi}(\boldsymbol{x}_i,\boldsymbol{y}_j)^\top,$$
$$\widehat{\boldsymbol{h}} := \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\varphi}(\boldsymbol{x}_i,\boldsymbol{y}_i).$$

Differentiating the objective function (3) with respect to $\boldsymbol{\alpha}$ and equating it to zero, we can obtain an analytic-form solution:

$$\widetilde{\boldsymbol{\alpha}} = (\widehat{\boldsymbol{H}} + \lambda\boldsymbol{I}_b)^{-1}\widehat{\boldsymbol{h}},$$

where $\boldsymbol{I}_b$ is the b-dimensional identity matrix.

We call the above method *Least-Squares Mutual Information (LSMI)* . Thanks to the analytic-form solution, the LSMI solution can be computed very efficiently.

Convergence Bound

Here, we show a non-parametric convergence rate of the solution of the optimization problem (3).

Let $\mathcal{G}$ be a general set of functions on $\mathcal{D}_{\mathrm{X}} \times \mathcal{D}_{\mathrm{Y}}$. For a function g ($\in \mathcal{G}$), let us consider a non-negative function $R(g)$ such that

$$\sup_{\boldsymbol{x},\boldsymbol{y}}[g(\boldsymbol{x},\boldsymbol{y})] \leq R(g).$$

Then the problem (3) can be generalized as

$$\widehat{w} := \underset{g\in\mathcal{G}}{\operatorname{argmin}} \left[\frac{1}{2n^2}\sum_{i,j=1}^{n} g_{i,j}^2 - \frac{1}{n}\sum_{i=1}^{n} g_{i,i} + \lambda_n R(g)^2\right],$$

where $g_{i,j} := g(\boldsymbol{x}_i, \boldsymbol{y}_j)$. We assume that the true density ratio function $w(\boldsymbol{x}, \boldsymbol{y})$ is contained in model $\mathcal{G}$ and satisfies

$$w(\boldsymbol{x}, \boldsymbol{y}) < M_0 \quad \text{for all} \quad (\boldsymbol{x}, \boldsymbol{y}) \in D_X \times D_Y.$$

We also assume that there exists γ $(0 < \gamma < 2)$ such that

$$\mathcal{H}_{[]}(\mathcal{G}_M, \epsilon, L_2(p_X p_Y)) = O((M/\epsilon)^\gamma),$$

where

$$\mathcal{G}_M := \{g \in \mathcal{G} \mid R(g) \leq M\}$$

and $\mathcal{H}_{[]}$ is the *bracketing entropy* of $\mathcal{G}_M$ with respect to the $L_2(p_{\mathrm{x}}p_{\mathrm{y}})$-norm [25, 24]. This means the function class $\mathcal{G}$ is not too much complex.

Then we have the following theorem. Its proof is omitted due to lack of space.

Theorem 1 *Under the above setting, if* $\lambda_n \to 0$ *and* $\lambda_n^{-1} = o(n^{2/(2+\gamma)})$ *then*

$$\|\widehat{w} - w\|_2 = \mathcal{O}_p(\lambda_n^{1/2}),$$

where $\|\cdot\|_2$ *means the* $L_2(p_{\mathrm{x}}p_{\mathrm{y}})$*-norm and* $\mathcal{O}_p$ *denotes the asymptotic order in probability.*

This theorem is closely related to [23, 2]. The paper [23] considers least squares estimators for nonparametric regression, and related topics can be found in Section 10 of [24].

CV for Model Selection and Basis Function Design

The performance of LSMI depends on the choice of the model, i.e., the basis functions $\boldsymbol{\varphi}(\boldsymbol{x}, \boldsymbol{y})$ and the regularization parameter λ. Here we show that model selection can be carried out based on a variant of CV.

First, the samples $\{\boldsymbol{z}_i \mid \boldsymbol{z}_i = (\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^n$ are divided into K disjoint subsets $\{\mathcal{Z}_k\}_{k=1}^K$. Then a density ratio estimator $\widehat{w}_k(\boldsymbol{x}, \boldsymbol{y})$ is obtained using $\{\mathcal{Z}_j\}_{j \neq k}$ and the cost J is approximated using the held-out samples $\mathcal{Z}_k$ as

$$\widehat{J}_k^{(K\text{-CV})} = \sum_{\boldsymbol{x}', \boldsymbol{y}' \in \mathcal{Z}_k} \frac{\widehat{w}_k(\boldsymbol{x}', \boldsymbol{y}')^2}{2n_k^2} - \sum_{(\boldsymbol{x}', \boldsymbol{y}') \in \mathcal{Z}_k} \frac{\widehat{w}_k(\boldsymbol{x}', \boldsymbol{y}')}{n_k},$$

where n_k is the number of pairs in the set $\mathcal{Z}_k$. $\sum_{\boldsymbol{x}', \boldsymbol{y}' \in \mathcal{Z}_k}$ is the summation over all combinations of $\boldsymbol{x}'$ and $\boldsymbol{y}'$ (i.e., n_k^2 terms), while $\sum_{(\boldsymbol{x}', \boldsymbol{y}') \in \mathcal{Z}_k}$ is the summation over all pairs $(\boldsymbol{x}', \boldsymbol{y}')$ (i.e., n_k terms). This procedure is repeated for $k = 1, 2, \ldots, K$ and its average $\widehat{J}^{(K\text{-CV})}$ is used as an estimate of J:

$$\widehat{J}^{(K\text{-CV})} = \frac{1}{K} \sum_{k=1}^K \widehat{J}_r^{(K\text{-CV})}.$$

We can show that $\widehat{J}^{(K-\mathrm{CV})}$ is an almost unbiased estimate of the true cost J, where the 'almost'-ness comes from the fact that the number of samples is reduced in the CV procedure due to data splitting [18].

A good model may be chosen by CV, given that a family of promising model candidates is prepared. As model candidates, we propose using a Gaussian kernel model:

$$\varphi_\ell(\boldsymbol{x}, \boldsymbol{y}) = \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{u}_\ell\|^2}{2\sigma^2}\right)\delta(\boldsymbol{y} = \boldsymbol{v}_\ell),$$

where

$$\{(\boldsymbol{u}_\ell, \boldsymbol{v}_\ell)\}_{\ell=1}^b$$

are 'center' points randomly chosen from

$$\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^n.$$

$\delta(\boldsymbol{y} = \boldsymbol{v}_\ell)$ is a indicator function, which is 1 if $\boldsymbol{y} = \boldsymbol{v}_\ell$ and 0 otherwise.

In the experiments, we fix the number of basis functions at

$$b = \min(100, n),$$

and choose the Gaussian width σ and the regularization parameter λ by CV with grid search.

Relation to Existing Methods

In this section, we discuss the characteristics of existing and proposed approaches.

Kernel Density Estimator (KDE)

KDE [20, 6] is a non-parametric technique to estimate a probability density function $p(\boldsymbol{x})$ from its i.i.d. samples $\{\boldsymbol{x}_i\}_{i=1}^n$. For the Gaussian kernel, KDE is expressed as

$$\widehat{p}(\boldsymbol{x}) = \frac{1}{n(2\pi\sigma^2)^{d/2}}\sum_{i=1}^n \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{x}_i\|^2}{2\sigma^2}\right).$$

The performance of KDE depends on the choice of the kernel width σ and it can be optimized by *likelihood CV* as follows [9]: First, divide the samples $\{\boldsymbol{x}_i\}_{i=1}^n$ into K disjoint subsets $\{\mathcal{X}_k\}_{k=1}^K$. Then obtain a density estimate $\widehat{p}_{\mathcal{X}_k}(\boldsymbol{x})$ from $\{\mathcal{X}_j\}_{j\neq k}$ and compute its hold-out log-likelihood for $\mathcal{X}_k$:

$$\frac{1}{|\mathcal{X}_k|}\sum_{\boldsymbol{x}\in\mathcal{X}_k}\log\widehat{p}_{\mathcal{X}_k}(\boldsymbol{x}).$$

This procedure is repeated for $k = 1, 2, \ldots, K$ and choose the value of σ such that the average of the hold-out log-likelihood over all k is maximized. Note that the average

hold-out log-likelihood is an almost unbiased estimate of the Kullback-Leibler divergence from $p(\boldsymbol{x})$ to $\widehat{p}(\boldsymbol{x})$, up to an irrelevant constant.

Based on KDE, MI can be approximated by separately estimating the densities $p_{\mathrm{xy}}(\boldsymbol{x}, \boldsymbol{y})$, $p_{\mathrm{x}}(\boldsymbol{x})$ and $p_{\mathrm{y}}(\boldsymbol{y})$ using $\{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^n$. However, density estimation is known to be a hard problem and therefore the KDE-based approach may not be so effective in practice.

k-nearest Neighbor Method (KNN)

Let $\mathcal{N}_k(i)$ be the set of k-nearest neighbor samples of $(\boldsymbol{x}_i, \boldsymbol{y}_i)$, and let

$$\begin{aligned}
\epsilon_{\mathrm{x}}(i) &:= \max\{\|\boldsymbol{x}_i - \boldsymbol{x}_{i'}\| \mid (\boldsymbol{x}_{i'}, \boldsymbol{y}_{i'}) \in \mathcal{N}_k(i)\}, \\
\epsilon_{\mathrm{y}}(i) &:= \max\{\|\boldsymbol{y}_i - \boldsymbol{y}_{i'}\| \mid (\boldsymbol{x}_{i'}, \boldsymbol{y}_{i'}) \in \mathcal{N}_k(i)\}, \\
n_{\mathrm{x}}(i) &:= \#\{\boldsymbol{z}_{i'} \mid \|\boldsymbol{x}_i - \boldsymbol{x}_{i'}\| \leq \epsilon_{\mathrm{x}}(i)\}, \\
n_{\mathrm{y}}(i) &:= \#\{\boldsymbol{z}_{i'} \mid \|\boldsymbol{y}_i - \boldsymbol{y}_{i'}\| \leq \epsilon_{\mathrm{y}}(i)\}.
\end{aligned}$$

Then the KNN-based MI estimator is given as follows [13]:

$$\widehat{I}(X, Y) = \psi(k) + \psi(n) - \frac{1}{k} - \frac{1}{n} \sum_{i=1}^{n} \left[\psi(n_x(i)) + \psi(n_y(i))\right],$$

where ψ is the *digamma* function.

A practical drawback of the KNN-based approach is that the estimation accuracy depends on the value of k and there seems no systematic strategy to choose the value of k appropriately.

Edgeworth Expansion (EDGE)

MI can be expressed in terms of the entropies as

$$I(X, Y) = H(X) + H(Y) - H(X, Y),$$

where $H(X)$ denotes the entropy of X:

$$H(X) := -\int p_{\mathrm{x}}(\boldsymbol{x}) \log p_{\mathrm{x}}(\boldsymbol{x}) d\boldsymbol{x}.$$

Thus MI can be approximated if the entropies above are estimated.

In the paper [11], an entropy approximation method based on the *Edgeworth expansion* is proposed, where the entropy of a distribution is approximated by that of the normal distribution and some additional higher-order correction terms. More specifically, for a d-dimensional distribution, the entropy is approximated by

$$H \approx H_{\mathrm{normal}} - \frac{1}{12} \sum_{i=1}^{d} \kappa_{i,i,i}^2 - \frac{1}{4} \sum_{i,j=1, i \neq j}^{d} \kappa_{i,i,j}^2 - \frac{1}{72} \sum_{i,j,k=1, i<j<k}^{d} \kappa_{i,j,k}^2,$$

where H_{normal} is the entropy of the normal distribution with covariance matrix equal to the target distribution and $\kappa_{i,j,k}$ $(1 \leq i, j, k \leq d)$ is the standardized third cumulant of the target distribution. In practice, all the cumulants are estimated from samples.

If the underlying distribution is close to the normal distribution, the above approximation is quite accurate and the EDGE method works very well. However, if the distribution is far from the normal distribution, the approximation error gets large and therefore the EDGE method may be unreliable.

In principle, it is possible to include the fourth and even higher cumulants for further reducing the estimation bias. However, this in turn increases the estimation variance; the expansion up to the third cumulants would be reasonable.

Competing interests

The authors declare that they have no competing interests.

Authors contributions

TS developed the method, implemented the algorithm and wrote the manuscript. MS and TK discussed the method and revised the manuscript. JS discussed the method, interpreted the results and wrote the manuscript.

Acknowledgements

This work was partially supported by KAKENHI (Grant-in-Aid for Scientific Research) on Priority Areas "Systems Genomics" from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

T.S. was supported by the JSPS Research Fellowships for Young Scientists.

References

[1] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, et al. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nature Genetics*, 25(1):25–29, May 2000.

[2] L. Birgé and P. Massart. Rates of convergence for minimum contrast estimators. *Probability Theory and Related Fields*, 97:113–150, 1993.

[3] E. I. Boyle, S. Weng, J. Gollub, H. Jin, D. Botstein, J. M. Cherry, and G. Sherlock. Go::termfinder—open source software for accessing gene ontology information and finding significantly enriched gene ontology terms associated with a list of genes. *Bioinformatics*, 20:3710–3715, 2004.

[4] P. Comon. Independent component analysis, a new concept? *Signal Processing*, 36(3):287–314, 1994.

[5] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, Inc., N. Y., 1991.

[6] A. M. Fraser and H. L. Swinney. Independent coordinates for strange attractors from mutual information. *Physical Review A*, 33(2):1134–1140, 1986.

[7] A. P. Gasch, P. T. Spellman, C. M. Kao, O. Carmel-Harel, M. B. Eisen, G. Storz, D. Botstein, and P. O. Brown. Genomic expression programs in the response of yeast cells to environmental changes. *Molecular Biology of the Cell*, 11(12):4241–4257, 2000.

[8] I. Guyon and A. Elisseeff. An introduction to variable feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

[9] W. Härdle, M. Müller, S. Sperlich, and A. Werwatz. *Nonparametric and Semiparametric Models*. Springer Series in Statistics. Springer, Berlin, 2004.

[10] T. R. Hughes, M. J. Marton, A. R. Jones, et al. Functional discovery via a compendium of expression proiles. *Cell*, 102:109–126, 2000.

[11] M. M. Van Hulle. Edgeworth approximation of multivariate differential entropy. *Neural Computation*, 17(9):1903–1910, 2005.

[12] S. Khan, S. Bandyopadhyay, A. Ganguly, and S. Saigal. Relative performance of mutual information estimation methods for quantifying the dependence among short and noisy data. *Physical Review E*, 76:026209, 2007.

[13] A. Kraskov, H. Stögbauer, and P. Grassberger. Estimating mutual information. *Physical Review E*, 69:066138, 2004.

[14] M. S. Pepe. *Evaluation of Medical Tests for Classification and Prediction*. Oxford Press, 2003.

[15] I. Priness, O. Maimon, and I. Ben-Gal. Evaluation of gene-expression clustering via mutual information distance measure. *BMC Bioinformatics*, 8:111, 2007.

[16] Y. Saeys, I. Inza, and P. Larranaga. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19):2507–2517, 2007.

[17] B. T. Schlitt, K. Palin, J. Rung, S. Dietmann, M. Lappe, E. Ukkonen, and A. Brazma. From gene networks to gene function. *Genome Research*, 13:2568–2576, 2003.

[18] B. Schölkopf and A. J. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.

[19] E. Segal, N. Friedman, D. Koller, and A. Regev. A module map showing conditional activity of expression modules in cancer. *Nature Genetics*, 36(10):1090–1098, 2004.

[20] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman & Hall/CRC, April 1986.

[21] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the USA*, 102:15545–15550, 2005.

[22] K. Torkkola. Feature extraction by non-parametric mutual information maximization. *Journal of Machine Learning Research*, 3:1415–1438, 2003.

[23] S. van de Geer. Estimating a regression function. *The Annals of Statistics*, 18(2):907–924, 1990.

[24] S. van de Geer. *Empirical Processes in M-Estimation*. Cambridge University Press, 2000.

[25] A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes. With Applications to Statistics*. Springer, New York, 1996.