BIOINFORMATICS
Vol. 1 no. 1 2003
Pages 1–8

Extraction of Correlated Gene Clusters from Multiple Genomic Data by Generalized Kernel Canonical Correlation Analysis

Y. Yamanishi *[1], J.-P. Vert[2], A. Nakaya[1] and M.Kanehisa[1]

[1]Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto 611-0011, Japan and [2]Centre de Géostatistique, Ecole des Mines de Paris 35 rue Saint-Honoré, 77305 Fontainebleau cedex, France

ABSTRACT

Motivation: A major issue in computational biology is the reconstruction of pathways from several genomic datasets, such as expression data, protein interaction data and phylogenetic profiles. As a first step toward this goal, it is important to investigate the amount of correlation which exists between these data.

Method: We present new methods to measure the correlation between several heterogeneous datasets, and to extract sets of genes which share similarities with respect to multiple biological attributes. The originality of our approach is the extension of the concept of correlation for non-vectorial data, which is made possible by the use of generalized kernel canonical correlation analysis (KCCA), and the method we propose to extract groups of genes responsible for the detected correlations. Moreover, two variants of KCCA are proposed when more than two datasets are available.

Result: These methods are successfully tested on their ability to recognize operons in the *E. coli* genome, from the comparison of three datasets corresponding to *functional* relationships between genes in metabolic pathways, *geometrical* relationships along the chromosome, and *co-expression* relationships as observed by gene expression data.

Contact: yoshi@kuicr.kyoto-u.ac.jp

INTRODUCTION

Recent developments in high-throughput technologies have filled biological databases with many kinds of genomic data, such as pathway knowledge (Kanehisa *et al.*, 2002), microarray gene expression data (Eisen *et al.*, 1998), protein-protein interaction data (Ito *et al.*, 2001), phylogenetic profiles (Pellegrini *et al.*, 1999), and several more. The problem of reconstructing pathways from such genomic datasets is a major issue in computational biology, because pathways represent a higher level of biological functions than single genes. As a first step toward this goal, it is crucial to investigate the correlation which exists between multiple biological attributes, and eventually to use this correlation in order to extract biologically meaningful features from heterogeneous genomic data. Indeed, a correlation detected between multiple datasets is likely to be due to some hidden biological phenomenon. Moreover, by selecting the genes responsible for the correlation, one can expect to select groups of genes which play a special role in or are affected by the underlying biological phenomenon. As an example, the existence of operons in prokaryotes is responsible for a form of correlation between several datasets, because genes which form operons are close to each other along chromosomes, have similar expression profiles, and can catalyze successive reactions in a pathway. Conversely, one can start from three datasets containing the localization of the genes on the genome, their expression profiles, and the chemical reactions they catalyze in known pathways, and look for correlations between these datasets, in order to finally recover groups of genes, which may form operons.

The integration of different kinds of data has been investigated with a variety of approaches so far. Using graph-theoretical arguments, clusters of genes have been extracted from several biological networks using multiple graph comparison by Ogata *et al.* (2000) and Nakaya *et al.* (2001). Several approaches using kernel methods have also been proposed, such as the combination of kernel matrices of expression data and phylogenetic profiles (Pavlidis *et al.*, 2001) or the extraction of features from microarray data using a gene network as side information (Vert *et al.*, 2003). In both cases, the goal was to improve the performance of gene function prediction algorithms.

A well-known statistical method to investigate the correlation between different real-valued attributes is canonical correlation analysis (CCA) (Hotelling , 1936). However, ordinary CCA cannot be applied to non-vectorial genomic data, such as pathways, protein-protein interactions or gene positions in a chromosome. In this paper we over-

*To whom correspondence should be addressed.

come this issue by using a generalization of CCA, known as kernel CCA (KCCA) (Akaho, 2001; Bach *et al.*, 2001), which provides a way to perform a generalized form of CCA on any data type as soon as a kernel function can be defined. KCCA finds directions simultaneously in the two feature spaces defined by the kernel functions with maximum correlation. If some biological phenomenon is responsible for this correlation, then significantly high/low-scoring genes in each direction can be considered functionally related in some biological meaning.

In this paper we derive two variants of KCCA in order to perform CCA on more than two datasets. The first one is a multiple KCCA, which is a natural generalization of KCCA to more than two kernel functions, already suggested by Bach *et al.* (2001). The second one is an integrated KCCA, which is a normal KCCA carried out with two kernels which are themselves sums of primary kernels. Integrated KCCA can therefore be useful to extract correlations between two sets of data types, represented by two sets of kernel functions. These methods are tested on their ability to extract operons from the *E.coli* genome, by detecting correlations between the KEGG/pathways dataset, the positions of the genes on the genome, and microarray expression data.

MATERIAL AND METHODS

Data

The dataset of pathways is constructed from the KEGG/LIGAND database (Goto *et al.*, 2000). The LIGAND database of chemical compounds and reactions in biological pathways is part of KEGG database (Kanehisa *et al.*, 2002). It contains thousands of metabolic reactions known to take place in various organisms, together with the substrates involved and the classification of the catalyzing enzyme as an EC number. From this database we created an undirected graph with genes of an organism as vertices, and where two vertices are linked when the genes catalyze two successive reactions in a pathway.

The dataset containing the position of the genes on a sequenced genome is available from KEGG/BRITE database (Goto *et al.*, 1996). BRITE is a database of binary relations for computation and comparison of graphs involving genes and proteins. From the gene position information we created a graph whose nodes correspond to genes and whose edges encode the neighboring association between two genes on a chromosome.

The data of microarray expression are downloaded from ExpressDB (Aach *et al.*, 2000). ExpressDB is a relational database containing yeast and *E.coli* RNA expression data and information loaded from numerous expression studies. We created a multivariate dataset whose individuals correspond to genes and whose variables correspond to four experimental conditions.

Ordinary kernel canonical correlation analysis

Kernel CCA (Akaho, 2001; Bach *et al.*, 2001) is a method which generalizes classical CCA and which we now recall. Its goal is to detect correlations between two datasets $\{\mathbf{x}^i\}_{i=1}^N$ and $\{\mathbf{y}^i\}_{i=1}^N$, where N is the number of objects and each object $\mathbf{x}^i$ (resp. $\mathbf{y}^i$) belongs to some set $\mathcal{X}$ (resp. $\mathcal{Y}$). To this end, the objects $\mathbf{x}^i$ (resp. $\mathbf{y}^i$) are mapped to some Hilbert space H_x (resp. H_y) by a mapping $\phi_x(.)$ (resp. $\phi_y(.)$). These objects correspond to genes or proteins in this study, and each data corresponds to one representation of the genes. Classical CCA can then be performed between the images $\{\phi_x(\mathbf{x}^i)\}_{i=1}^N$ and $\{\phi_y(\mathbf{y}^i)\}_{i=1}^N$ as follows. The goal is to find two directions $f_x \in H_x$ and $f_y \in H_y$ such that the features

$$\begin{aligned} u_x &= \langle f_x, \phi_x(\mathbf{x}) \rangle, \\ u_y &= \langle f_y, \phi_y(\mathbf{y}) \rangle, \end{aligned} \tag{1}$$

be maximally correlated. As directions orthogonal to the linear spans of the points do not contribute to any correlation, f_x and f_y can be restricted to belong to these linear spans. They can therefore be expressed as:

$$\begin{aligned} f_x &= \textstyle\sum_i \alpha_x^i \phi_x(\mathbf{x}^i), \\ f_y &= \textstyle\sum_i \alpha_y^i \phi_y(\mathbf{y}^i), \end{aligned} \tag{2}$$

in which case the corresponding u_x and u_y can be rewritten as

$$\begin{aligned} u_x &= \textstyle\sum_i \alpha_x^i \langle \phi_x(\mathbf{x}^i), \phi_x(\mathbf{x}) \rangle, \\ u_y &= \textstyle\sum_i \alpha_y^i \langle \phi_y(\mathbf{y}^i), \phi_y(\mathbf{y}) \rangle. \end{aligned} \tag{3}$$

The f_x and f_y can now be found by solving the Lagrangean

$$\begin{aligned} L_0 &= E[(u_x - E[u_x])(u_y - E[u_y])] \\ &- \tfrac{\rho_x}{2} E[(u_x - E[u_x])^2] - \tfrac{\rho_y}{2} E[(u_y - E[u_y])^2], \end{aligned} \tag{4}$$

where ρ_x and ρ_y are Lagrange multipliers. However, the Lagrangean is ill-posed when the dimensionalities of the Hilbert spaces are too large. To overcome the difficulty, penalty terms $PEN(f_x)$ and $PEN(f_y)$ are introduced to form a new Lagrangean :

$$L = L_0 + \frac{\lambda_x}{2} PEN(f_x) + \frac{\lambda_y}{2} PEN(f_y), \tag{5}$$

where λ_x and λ_y are regularization parameters. The importance of regularization in CCA in high dimension has been discussed in detail (see Hastie *et al.*, 1995; Leurgans*et al.*, 1993).

Any kernel function $k_{\mathcal{X}}(\cdot,\cdot)$ on $\mathcal{X}^2$ defines a Hilbert space and a mapping $\phi_x(.)$ (Schölkopf *et al.*, 2002) such that $\forall(\mathbf{x}^1,\mathbf{x}^2) \in \mathcal{X}^2, k_x(\mathbf{x}^1,\mathbf{x}^2) =< \phi(\mathbf{x}^1), \phi(\mathbf{x}^2) >$. Examples of the kernel functions are the following:

$$k_x(\mathbf{x}^1,\mathbf{x}^2) = (\mathbf{x}^1 \cdot \mathbf{x}^2 + 1)^d, \tag{6}$$

$$k_x(\mathbf{x}^1,\mathbf{x}^2) = \exp\{- \parallel \mathbf{x}^1 - \mathbf{x}^2 \parallel^2 /2\sigma^2\}. \tag{7}$$

Eq.(6) is a polynomial kernel with degree d, and eq.(7) is a Gaussian radial basis function (RBF) kernel with width σ. Let now $(\mathbf{K}_x)_{ij} := k_x(\mathbf{x}^i,\mathbf{x}^j)$ and $(\mathbf{K}_y)_{ij} := k_y(\mathbf{y}^i,\mathbf{y}^j)$ be two kernel matrices, assumed to be centered (Bach *et al.*, 2001). Then L can be rewritten as

$$\begin{aligned} L &= \alpha_x^T \mathbf{K}_x \mathbf{K}_y \alpha_y \\ &- \tfrac{\rho_x}{2}\alpha_x^T(\mathbf{K}_x + \lambda_x \mathbf{I})^2\alpha_x - \tfrac{\rho_y}{2}\alpha_y^T(\mathbf{K}_y + \lambda_y \mathbf{I})^2\alpha_y, \end{aligned} \tag{8}$$

where $\mathbf{I}$ is an identity matrix and $\alpha_x = (\alpha_x^1, \cdots, \alpha_x^N)^T$ and $\alpha_y = (\alpha_y^1, \cdots, \alpha_y^N)^T$. This shows that regularization parameters λ_x and λ_y control the trade-off between maximizing the correlation and penalizing the complexity of f_x and f_y. Maximizing this Lagrangean can be done by solving the following generalized eigenvalue problem:

$$\begin{aligned} &\begin{pmatrix} \mathbf{0} & \mathbf{K}_x\mathbf{K}_y \\ \mathbf{K}_y\mathbf{K}_x & \mathbf{0} \end{pmatrix} \begin{pmatrix} \alpha_x \\ \alpha_y \end{pmatrix} \\ &= \rho \begin{pmatrix} (\mathbf{K}_x + \lambda_x \mathbf{I}_x)^2 & \mathbf{0} \\ \mathbf{0} & (\mathbf{K}_y + \lambda_y \mathbf{I}_y)^2 \end{pmatrix} \begin{pmatrix} \alpha_x \\ \alpha_y \end{pmatrix}. \end{aligned} \tag{9}$$

After finding α_x and α_y, canonical correlation scores (CC scores) can be recovered by $u_x = \mathbf{K}_x\alpha_x$ and $u_y = \mathbf{K}_y\alpha_y$.

Multiple kernel canonical correlation analysis

Here we propose an extension of the model when more than two attributes are available. This method was suggested by Bach *et al.* (2001) for the purpose of independent component analysis. We refer to it as multiple kernel canonical correlation analysis (MKCCA). It is a straightforward extension of the ordinary KCCA model described in the previous section.

Suppose that we have P datasets $\{\mathbf{x}_p^i\}_{i=1}^N\,(p = 1, 2, \cdots, P)$, and mappings $\phi_{x_p}(\mathbf{x}_p)$ to some Hilbert spaces H_{x_p}. Then we can look for directions $f_p \in H_p\,(p = 1, 2, \cdots, P)$ such that the sum of all pairwise correlations between features

$$u_p = \langle f_p, \phi_x(\mathbf{x}_p)\rangle, \quad p = 1, \cdots, P, \tag{10}$$

be the largest possible. Since f_p is expressed as

$$f_p = \sum_i \alpha_p^i \phi_{x_p}(\mathbf{x}_p^i), \tag{11}$$

the corresponding u_p can be rewritten as

$$u_p = \sum_i \alpha_p^i \langle \phi_{x_p}(\mathbf{x}_p^i), \phi_{x_p}(\mathbf{x}_p)\rangle, \tag{12}$$

where $\alpha_p = (\alpha_p^1, \cdots, \alpha_p^N)^T$. The f_p can be found by solving the Lagrangean

$$\begin{aligned} L_0 &= \textstyle\sum_{p,q} E[(u_p - E[u_p])(u_q - E[u_q])] \\ &- \textstyle\sum_p \frac{\rho_p}{2} E[(u_p - E[u_p])^2]. \end{aligned} \tag{13}$$

Like in ordinary KCCA, we introduce a penalty term $PEN(f_{x_p})$ and we get

$$L = L_0 + \frac{\lambda_p}{2}\sum_p PEN(f_{x_p}), \tag{14}$$

where λ_p is a regularization parameter. Using the kernel trick, we can work with P kernel matrices as $(\mathbf{K}_{x_p})_{ij} := k_{x_p}(\mathbf{x}_p^i, \mathbf{x}_p^j)\,(p = 1, \cdots, P)$, assumed to be centered. Then L can be rewritten in the kernel form as

$$\begin{aligned} L &= \textstyle\sum_{p,q} \alpha_p^T \mathbf{K}_{x_p}\mathbf{K}_{x_q}\alpha_q \\ &- \textstyle\sum_p \frac{\rho_p}{2}\alpha_p^T(\mathbf{K}_{x_p} + \lambda_{x_p}\mathbf{I})^2\alpha_p. \end{aligned} \tag{15}$$

Finally, the estimation of canonical correlation scores (CC scores) is reduced to the following generalized eigenvalue problem:

$$\begin{aligned} &\begin{pmatrix} \mathbf{0} & \dots & \mathbf{K}_1\mathbf{K}_P \\ \vdots & \ddots & \vdots \\ \mathbf{K}_P\mathbf{K}_1 & \dots & \mathbf{0} \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_P \end{pmatrix} \\ &= \rho \begin{pmatrix} (\mathbf{K}_1 + \lambda_1\mathbf{I}_1)^2 & \dots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \dots & (\mathbf{K}_P + \lambda_P\mathbf{I}_P)^2 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_P \end{pmatrix}. \end{aligned} \tag{16}$$

After finding α_p, CC scores can be obtained by $u_p = \mathbf{K}_p\alpha_p\ (p = 1, 2, \cdots, P)$.

Integrated kernel canonical correlation analysis

Instead of maximizing the sum of all pairwise correlations as in multiple KCCA, one can prefer to maximize the correlation between one type of attribute and a combination of other types, or even between two combinations of attributes. A simple combination of attributes via kernel is obtained by summing two (or more) kernels, which is equivalent to take the direct sum of the feature spaces associated with each kernel as a new feature space. Hence it is a way to combine implicitly two attributes into a single one. This possibility was used for instance by Pavlidis

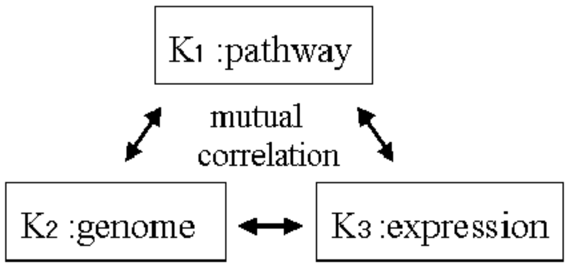

Fig. 1. Mutual correlation model in multiple KCCA (MKCCA).

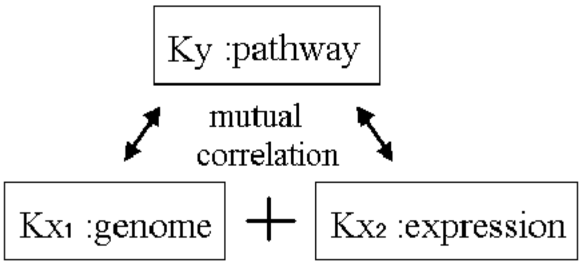

Fig. 2. Mutual correlation model in integrated KCCA (IKCCA).

et al. (2001), where kernel matrices of expression data and phylogenetic profiles are combined into one kernel matrix in support vector machine (SVM) for functional classification of genes.

More formally, suppose we have two major attributes x and y, where x has several sub-attributes $x_p \, (p = 1, 2, \cdots, P)$ and y has several sub-attributes $y_q \, (q = 1, 2, \cdots, Q)$. If a kernel is defined on each set of sub-attributes, then heterogeneous kernel matrices for x and y are computed as

$$\mathbf{K}_x = \sum_{p=1}^{P} \mathbf{K}_{x_p}, \quad \mathbf{K}_y = \sum_{q=1}^{Q} \mathbf{K}_{y_q}. \tag{17}$$

Classical KCCA is then performed on the heterogeneous kernel matrices, by solving the following generalized eigenvalue problem:

$$\begin{pmatrix} \mathbf{0} & \sum_p \mathbf{K}_{x_p} \sum_q \mathbf{K}_{y_q} \\ \sum_q \mathbf{K}_{y_q} \sum_p \mathbf{K}_{x_p} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \alpha_x \\ \alpha_y \end{pmatrix}$$
$$= \rho \begin{pmatrix} (\sum_p \mathbf{K}_{x_p} + \lambda_x \mathbf{I}_x)^2 & \mathbf{0} \\ \mathbf{0} & (\sum_q \mathbf{K}_{y_q} + \lambda_y \mathbf{I}_y)^2 \end{pmatrix} \begin{pmatrix} \alpha_x \\ \alpha_y \end{pmatrix} \tag{18}$$

Then, CC scores can be obtained as $u_x = \sum_p \mathbf{K}_{x_p} \alpha_x$ and $u_y = \sum_q \mathbf{K}_{y_q} \alpha_y$.

Diffusion kernel

When the dataset is a network of genes, it must be transformed into a kernel function to be analyzed by our methods. This operation can be performed with the diffusion kernel, proposed in Kondor *et al.* (2002), which we now recall.

Suppose that we have an undirected, unweighted graph $\Gamma = (V, E)$. The (opposite) Laplacian of this graph is the matrix

$$\mathbf{H}_{ij} = \begin{cases} 1 & \text{for } i \sim j, \\ -d_i & \text{for } i = j, \\ 0 & \text{otherwise}, \end{cases} \tag{19}$$

where $i \sim j$ means that the i-th and j-th genes are joined by an edge on the graph, and d_i is the number of genes emanating from the i-th gene. The exponential of the matrix $\mathbf{H}$ is defined as

$$\exp(\beta \mathbf{H}) = \lim_{m \to \infty} \left(1 + \frac{\beta \mathbf{H}}{m} \right)^m, \tag{20}$$

where β is a positive constant. This is equivalent to the following expansion:

$$\exp(\beta \mathbf{H}) = \mathbf{I} + \beta \mathbf{H} + \frac{\beta^2}{2} \mathbf{H}^2 + \frac{\beta^3}{3!} \mathbf{H}^3 + \cdots. \tag{21}$$

The resulting matrix is symmetric and positive definite. It is therefore a valid kernel called the diffusion kernel (Kondor *et al.*, 2002), which can be thought of as a generalization of the Gaussian RBF kernel to a discrete setting.

RESULTS

An operon is a characteristic structure of prokaryotic genomes. Genes belonging to the same operon are coregulated, often play successive roles in pathways, and are closely located on genomes. We therefore use pathways, genome positions, and microarray expression data for *E.coli* as original datasets, because those attributes are expected to exhibit some form of correlation between each other due to the presence of operons. Multiple and integrated KCCAs (MKCCA and IKCCA) are applied to extract operon structures of *E.coli*. Figure 1 and Figure 2 show the illustration of mutual correlation models in MKCCA and IKCCA, respectively.

In the application of MKCCA, the kernel matrices $\mathbf{K}_1$, $\mathbf{K}_2$ and $\mathbf{K}_3$ correspond to gene-gene similarities in

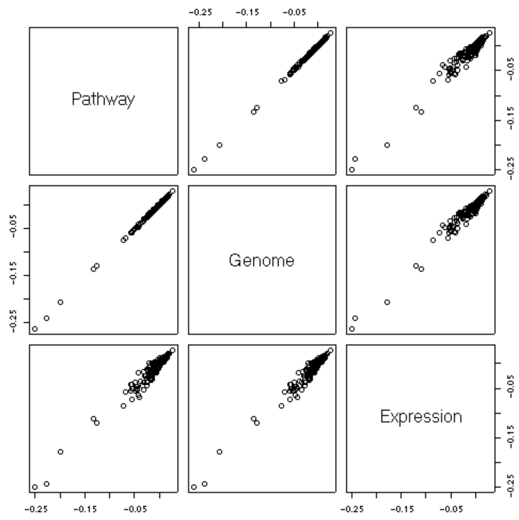

Fig. 3. CC1 scores in MKCCA: pathway vs. genome vs. expression

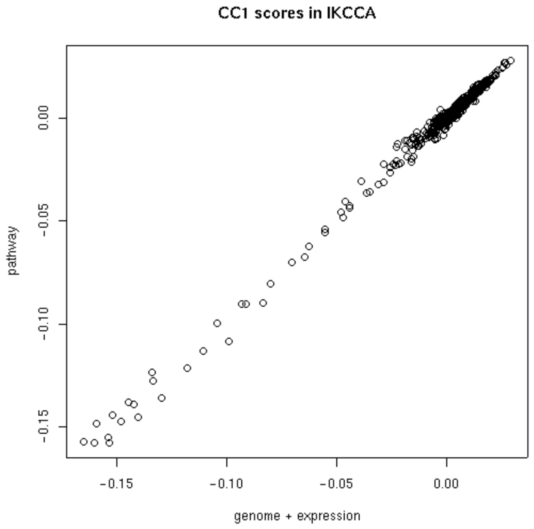

Fig. 4. CC1 scores in IKCCA: pathway vs. genome+expression

pathways, genome, and expression. The kernel matrices $\mathbf{K}_1$ and $\mathbf{K}_2$ are computed using diffusion kernel, where the parameter β is set to 1. The kernel matrix $\mathbf{K}_3$ is computed from a Gaussian RBF kernel with unit width. The regularization parameters λ in the algorithm of KCCAs are set to 0.1. In the application of IKCCA, we set $\mathbf{K}_y = \mathbf{K}_1$ (pathway) and $\mathbf{K}_{x_1} = \mathbf{K}_2$ (genome) and $\mathbf{K}_{x_2} = \mathbf{K}_3$ (expression), hence $Q = 1$ and $P = 2$. This is motivated by the hypothesis that gene-gene similarity in pathways can be predicted by a combination of gene-gene similarities in genome and expression, because successive reactions on pathways are often implied by neighboring relationships on chromosome or coexpression relationship in microarray experiment. For comparison, ordinary KCCA is applied twice. First, ordinary KCCA is applied to pathways and genome positions, which we refer to as OKCCA(a). Second, ordinary KCCA is applied to pathways and expressions, which we refer to as OKCCA(b). In each experiment, the maximum value in each kernel matrix is set to 1 by scaling of the matrices.

Figure 3 shows multiple cross-scatter plots of the first canonical correlation scores (CC1 scores) for genes in MKCCA between pathway, genome, and expression. Figure 4 shows a scatter plot of CC1 scores for genes in IKCCA between pathway and genome/expression. In each case a correlation has clearly been detected between different attributes. The correlations detected are mostly due to the genes with high or low score, which can be suspected of forming clusters simultaneously in several feature spaces. Such clusters are operon candidates, in the sense that they would correspond to genes close to each other with respect to their positions in the pathways, in the genome, and to their expression profiles. To validate this hypothesis we selected the upper and lower 5 $\%$ genes on the CC1 computed by each method, mapped and visualized them on the KEGG/pathway database, and compared them with known operons obtained from the Operon Data Library (http://cib.nig.ac.jp/dda/taitoh/operondata.html).

In this study we focus on operons in the metabolic pathways available from the KEGG database. Each gene is represented by the EC number of its product enzyme in this database, so we compare the EC numbers of selected genes with those of genes in known operons. Table 1 shows the number of genes selected by each method (OKCCA(a), OKCCA(b), MKCCA, and IKCCA) which belong to 9 major known operons. For those operons, IKCCA provides the best overall rate of gene detection, followed by MKCCA. Figure 5 shows an example of a known operon involved in the biotin metabolism, which contains 3 genes marked with bold lines. Figure 6 shows the genes selected by the IKCCA method which belong to the biotin metabolism, colored in gray. We observe that in this case, the gene selected correspond almost perfectly to the operon, except for the bioA gene (EC:2.6.1.62) which is selected but absent from the Operon Data Library. One can observe that the four genes selected form a more appropriate operon candidate than the three genes listed in the Operon database, because they catalyze four

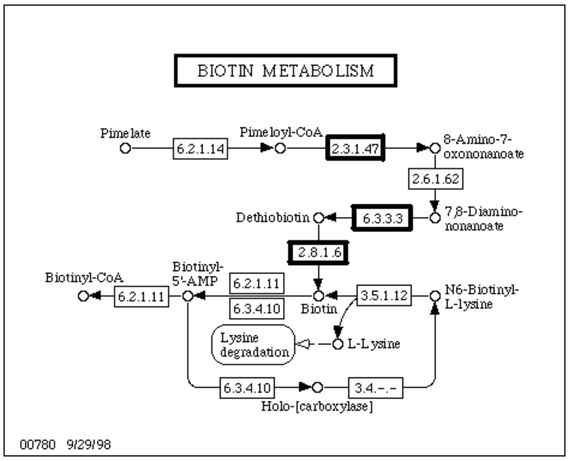

Fig. 5. An example of known operons in Operon Data Library

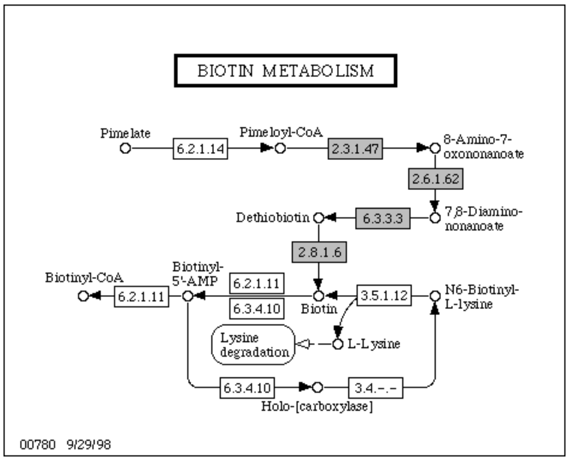

Fig. 6. An example of operons predicted by IKCCA

successive reaction in the biotin pathway.

DISCUSSION AND CONCLUTION

In this paper we presented new approaches to investigate the correlation between heterogeneous genomic data. We proposed several generalized formulations of ordinary canonical correlation analysis and derived the algorithm for computing CC scores. The integration of different types of genomic data (e.g., biochemical pathways, genomes, and expression data) is a key problem in computational biology nowadays. When data types are different (e.g., graphs, strings, and vectors), integration strategies often rely on various heuristic approaches, which depend on the types of the data. The originality of our approach is the extension of the concept of correlation for non-vectorial data and integration of genomic data in a rigorous mathematical framework common to all types.

The proposed methods (MKCCA and IKCCA) enable us to automatically find correlated directions, along which high/low scoring genes share similarities with respect to multiple biological attributes. These methods worked well to recognize the genes that belong to operons in the *E. coli* genome, by comparing three datasets corresponding to *functional* relationships between genes in metabolic pathways, *geometrical* relationships along the chromosome, and *co-expression* relationships as observed by gene expression data. We observed that generalized KCCAs (MKCCA and IKCCA) behaved better than ordinary KCCAs in terms of the numbers of correctly extracted operon candidates. In this work we used the 0.05 percentile as a threshold and confirmed that extracted genes correspond well to known operons of the *E.coli* genome. It would be necessary to determine an appropriate threshold for a more specific purpose. In our preliminary results it seems that the number of correct operon candidates selected by MKCCA tends to be smaller than that selected by IKCCA. One explanation for this difference in performance might be the fact that MKCCA looks for correlations simultaneously among all pairs of datasets. It would work well if the genes in an operon were systematically similar to each other with respect to all three sources of information we used. To the contrary, in our IKCCA setting, we relax the constraint of having a correlation between gene positions along the genome and gene expression, and rather focus on detection of correlations between positions on the pathways on the one hand, positions on the genome OR expression profile on the other. Due to noise and errors in the data, this less constrained problem might detect biological phenomena (operons in our case) more easily than the MKCCA approach. We conjecture that as the number of datasets increases, the performance of MKCCA might decrease because it becomes too difficult to impose correlation constraints between any two datasets. In that case it might be more efficient to try to detect correlations between a smaller number of datasets, obtained themselves by combining the initial datasets available, as we did in IKCCA.

From the viewpoint of algorithms, our method first starts by transforming each dataset into a kernel matrix whose elements represent similarities between genes. This process enables us to deal with different data types elegantly and in a unified framework. We showed that CC scores can be easily computed by solving the generalized eigenvalue problem. However, the performance of kernel methods often depends on the definition of the kernel

Table 1. Acurracy for operon prediction based on the first canonical correlation scores (CC1 scores) in each KCCA.

No. Operon (#gene products)	OKCCA(a)	OKCCA(b)	MKCCA	IKCCA
1 Fructose uptake (3)	2/3	2/3	2/3	2/3
2 Galactose metabolism (3)	0/3	2/3	3/3	3/3
3 Ubiquinone biosynthesis (2)	1/2	2/2	2/2	0/2
4 Fatty acid biosynthesis (2)	0/2	2/2	2/2	2/2
5 Purine nucleotide biosynthesis (2)	2/2	1/2	2/2	2/2
6 Pyruvate dehydrogenase (3)	3/3	3/3	3/3	3/3
7 Biotin metabolism (3)	0/3	1/3	1/3	3/3
8 Valine biosynthesis (5)	3/5	2/5	2/5	3/5
9 Peptidoglycan biosynthesis (3)	3/3	1/3	1/3	1/3
Total prediction rate	14/26	16/26	18/26	19/26

function and its parameters. In kernel methods such as support vector machines (SVM), a kernel function between two objects should be determined a priori (Müller *et al.*, 2001). For that reason, kernel engineering for various genomic data has been investigated by several methods recently in bioinformatics (e.g., Tsuda *et al.*, 2002; Vert, 2002). The performance of our method could be improved by using a more specific kernel function for each data type in actual application. In addition, it is necessary to develop appropriate normalization methods across different kernel matrices in KCCA algorithm, because scales are different across data types.

The biological motivation of our approach is similar to the work on multiple graph comparison (Ogata *et al.*, 2000; Nakaya *et al.*, 2001), where gene-gene relationships on all the attributes are regarded as graphs whose nodes correspond to genes and whose edges encode the presence of the association between two genes. In graph comparison methods, finding correlated gene clusters can be formalized as a subgraph isomorphism problem, but the method is based on the assumption that the relationship between genes are linear across different biological attributes. That is why the addition of too many graphs is sometimes too restrictive to uncover biologically meaningful findings. The other approach related to our study is the graph-driven features extraction method studied by Vert *et al.* (2003), where ordinary KCCA is applied to microarray expression data and biochemical pathways of yeast. They propose to use the resulting CC scores as feature vectors for functional classification by SVM. Their method improved the performance in functional classification. This suggests that the integration of different kinds of data is beneficial to improve the accuracy and reliability of the prediction of gene functions. It is also expected that such a good behavior can be obtained in functional prediction by using CC scores extracted by generalized KCCAs.

The proposed methods enable us to gain some understandings of how pathways are correlated with several genomic datasets. However, it is just a first step toward our final goal, of pathway prediction. The quantity of pathway data is by far fewer than that of other genomic datasets (e.g., gene expression, protein interaction, genome position, and phylogenetic profiles). That is why the computational reconstruction of unkown pathways from other genomic data is an important issue in bioinformatics. We are currently working on developing methods for pathway prediction by expanding the framework described in this paper.

ACKNOWLEDGEMENTS

This work was supported by grants from the Ministry of Education, Culture, Sports, Science and Technology of Japan, the Japan Society for the Promotion of Science, and the Japan Science and Technology Corporation. The computational resource was provided by the Bioinformatics Center, Institute for Chemical Research, Kyoto University.

REFERENCES

Aach,J., Rindone,W. and Church,G.M. (2000) Systematic management and analysis of yeast gene expression data. *Genome Res.*, **10(4)**, 431–445.

Akaho,S. (2001) A kernel method for canonical correlation analysis. *International Meeting of Psychometric Society (IMPS2001).*

Bach,F.R. and Jordan,M.I. (2001) Kernel independent component analysis. *Technical Report UCB//CSD-01-1166,UC Berkeley*.

Ito,T., Chiba,T., Ozawa,R., Yoshida,M., Hattori,M. and Sakaki,Y. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interacome. *Proc. Natl. Acad. Sci. USA*, **98(8)**, 4569–4574.

Eisen,M.B., Spellman,P.T., Patrick,O.B. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns.

Proc. Natl. Acad. Sci. USA, **95**, 14863–14868.

Goto,S., Bono,H., Ogata,H., Fujibuchi,W., Nishioka,T., Sato,K. and Kanehisa,M. (1996) Organizing and computing metabolic pathway data in terms of binary relations. *Pacific. Symp. Biocomputing*, **3**, 175–186.

Goto,S., Okuno,Y., Hattori,M. and Kanehisa,M. (2002) LIGAND: database of chemical compounds and reactions in biological pathways. *Nucleic Acids Res.*, **30**, 402–404.

Hotelling,H. (1936) Relation between two sets of variates. *Biometrika*, **28**, 322–277.

Hastie,T., Buja,A. and Tibshirani,R. (1995) Penalized discriminant analysis. *Annals of Statistics*, **23**, 73–102.

Kanehisa,M., Goto,S., Kawashima,S. and Nakaya,A. (2002) The KEGG databases at GenomeNet. *Nucleic Acids Res.*, **30**, 42–45.

Kondor,R.I. and Lafferty,J. (2002) Diffusion kernels on graphs and other discrete input. *International Conference on Machine Learning (ICML2002)*.

Leurgans,S., Moyeed,R. and Silverman,B. (1993) Canonical correlation analysis when the data are curves. *J. Royal Statist. Soc.*, **B55**, 725–740.

Müller,K.-R., Mika,S., Rätsch,G., Tsuda,K. and Schölkopf,B. (2001) An introduction to kernel-based learning algorithms. *IEEE Trans. Neural Networks*, **12**, 181–201.

Nakaya,A., Goto,S. and Kanehisa,M. (2001) Extraction of correlated gene clusters by multiple graph comparison. *Genome Informatics*, **12**, 44–53.

Ogata,H., Fujibuchi,W., Goto,S. and Kanehisa,M. (2000) A heuristic graph comparison algorithm and its application to detect functionally related enzyme clusters. *Nucleic Acids Res.*, **28**, 4021–4028.

Pellegrini,M., Marcotte,E.M., Thompson,M.J., Eisenberg,D. and Yeates,T.O. (1999) Assigning protein functions by comparative genome analysis:Protein phylogenetic profiles. *Proc. Natl. Acad. Sci. USA*, **96**, 4285–4288.

Pavlidis,P., Weton,J., Cai,J. and Grundy,W.N. (2001) Gene functional classification from heterogeneous data. *International Conference on Research in Computational Molecular Biology (RECOMB2001)*, 249–255.

Schölkopf,B. and Smola,A.J. (2002) *Learning with Kernels: Support Vector Machines, Regularization,Optimization, and Beyond*. MIT Press, Cambridge, MA.

Tsuda,K., Kin,T. and Asai,K. (2002) Marginalized kernels for biological sequences. *Bioinformatics*, **18**, S268–S275.

Vert,J.-P. (2002) A tree kernel to analyze phylogenetic profiles. *Bioinformatics*, **18**, S276–S284.

Vert,J.-P. and Kanehisa,M. (2003) Graph-driven features extraction from microarray data using diffusion kernels and kernel CCA. *To appear in Advances in Neural Information Processing Systems 15, MIT Press, Cambridge, MA*.