BIOINFORMATICS

Vol. 19 Suppl. 2 2003, pages ii238–ii244
DOI: 10.1093/bioinformatics/btg1084

Extracting active pathways from gene expression data

Jean Philippe Vert [1,*] *and Minoru Kanehisa* [2]

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

Received on March 17, 2003; accepted on June 9, 2003

ABSTRACT

Motivation: A promising way to make sense out of gene expression profiles is to relate them to the activity of metabolic and signalling pathways. Each pathway usually involves many genes, such as enzymes, which can themselves participate in many pathways. The set of all known pathways can therefore be represented by a complex network of genes. Searching for regularities in the set of gene expression profiles with respect to the topology of this gene network is a way to automatically extract active pathways and their associated patterns of activity.

Method: We present a method to perform this task, which consists in encoding both the gene network and the set of profiles into two kernel functions, and performing a regularized form of canonical correlation analysis between the two kernels.

Results: When applied to publicly available expression data the method is able to extract biologically relevant expression patterns, as well as pathways with related activity.

Contact: Jean-Philippe.Vert@mines.org

INTRODUCTION

The microarray technology is a crucial tool to elucidate the genetic regulation mechanisms in a cell. By simultaneously monitoring the quantity of messenger RNA of virtually all genes of an organism submitted to various conditions, one should in principle be able to reconstruct some parts of the regulatory network at the transcription level, and much effort has been devoted to this task recently (de Jong, 2002). However the complexity of the mechanisms involved in genetic regulation, metabolic and signalling pathways, together with the still limited amount of expression data available, have limited the successes of genetic regulation reconstruction from microarray data to specific pathways or small numbers of genes.

On the other hand, many signalling and metabolic pathways have been experimentally characterized and recently integrated into databases such as KEGG (Kanehisa *et al.*, 2002). While still far from being complete, such curated databases contain the state-of-the-art of our knowledge about biochemical pathways. It is therefore natural to put in parallel these known pathways with the gene expression data obtained from microarray experiments, in order to validate the pathways, find new candidate pathways or check the quality of expression data. For example, it is possible to map gene clusters obtained from microarray data onto known metabolic networks to find pathways of interest (van Helden *et al.*, 2000); pathway scoring methods have been developed (Zien *et al.*, 2000) to assess the validity of candidate pathways in terms of expression data; more recently, methods were proposed to integrate expression data and pathway network at an early stage in order to extract clusters of genes which have similar expression profiles and participate in common pathways in the same time (Hanisch *et al.*, 2002).

In this report we study a different and possibly complementary approach which consists in looking for *correlations* between known genetic networks and microarray data. While expression data characterize each gene by a profile, i.e. a vector of real numbers, the pathway database provides a graph representation where genes are nodes and where two genes are linked when then catalyze successive reactions in some known pathway. The term 'correlation', usually used to assess the relationship between two random variables, must therefore be generalized to assess the relationship between two different random elements: a node of a graph on the one hand, a profile on the other hand.

In this paper we propose a method to give a sense to the term 'correlation' in this context, using the theory of kernel methods (Schölkopf and Smola, 2002) and spectral graph theory (Chung, 1997). Because of space limitations we present the technicalities of the approach in a companion paper (Vert and Kanehisa, 2002) and focus in this report on the possible applications of the methods to make sense out of expression data.

*To whom corespondence should be addressed.

METHOD

In this section we briefly describe an algorithm to extract correlations between nodes of a graph and real-valued vectors. The reader interested in the details and justifications of the method is referred to the companion paper (Vert and Kanehisa, 2002).

We model the set of genes by a discrete set $\mathcal{X}$ of cardinality $|\mathcal{X}| = n$. Each gene $x \in \mathcal{X}$ is supposed to be characterized by an expression profile $e(x) \in \mathbb{R}^p$, where p is the number of measurements available (supposed to be the same for all genes). By subtracting the mean profile from all genes, we suppose in the sequel that the set of profiles is centered, i.e. $\sum_{x\in\mathcal{X}} e(x) = 0$.

Independently of the gene expression profiles, we assume that a gene network has been extracted from a database of known biochemical pathways. More precisely, genes are nodes of this network, and two genes are linked whenever they encode proteins which have the possibility to catalyze two successive reactions in a pathway. This is mathematically represented by a simple graph $\Gamma = (\mathcal{X}, \mathcal{E})$, with the genes as vertices.

The main goal of our method is to automatically find profiles which exhibit some coherence with respect to the topology of the network. Formally speaking, a profile is a vector $v \in \mathbb{R}^p$. We don't require v to be any actual gene expression profile, but rather use it to represent some more abstract or hidden information, such as the quantity of some substance in the cell, or the activity of a pathway. Intuitively, if v represents the evolution of such a biological quantity, then expression profiles of genes participating in or affected by this event should exhibit some form of correlation with v.

For a zero-mean candidate profile $v \in \mathbb{R}^p$ (i.e. $\sum_{i=1}^p v_i = 0$), let us therefore call $f_v(x) \stackrel{\Delta}{=} v^T e(x)$ the correlation between v and $e(x)$. Typically, if v represents the activity level of a pathway where gene x plays a central regulatory role, then $f_e(x)$ is likely to be either strongly positive or strongly negative.

We now turn our attention to the problem of assessing how likely a candidate profile v is to be biologically relevant, and define two independent criteria which both reveal some form of relevance.

First of all, as microarray data are known to be noisy, and as the profiles v we are looking for are likely to be correlated to several genes, a statistical criterion such as the normalized variance of a feature $f_e(.)$, defined by:

$$V(f_e) = \frac{\sum_{x\in\mathcal{X}} f_e(x)^2}{||v||^2}, \tag{1}$$

quantifies how much a profile v explains the variations among expression profiles. This criterion only involves the set of expression profiles, and is maximized in principal component analysis (PCA). The larger $V(f_e)$, the more v explains the variations between profiles, and the more likely it is to correspond to some hidden biological event which influences many genes.

Second, and independently of the quantity (1), a criterion can be defined to assess how much the feature $f_e(.)$ correlates with the topology of the graph Γ. The criterion we choose to quantify is the smoothness of the feature with respect to the graph topology, i.e. how much $f_e(.)$ varies between adjacent nodes. Indeed, if a profile v corresponds to a biological event involving a pathway, then the linear feature $f_e(.)$ should exhibit some form of smoothness at least in the part of the gene network concerned by the event.

The smoothness of a feature can be quantified through its discrete Fourier transform (Chung, 1997). As an example, if $\hat{f} = \left(\hat{f}_1, \ldots, \hat{f}_n\right)$ is the discrete Fourier transform of a feature $f_e(.)$, then the smoothness of this feature can be measured by the following quantity which is large when $\hat{f}$ has a lot of energy at high frequency:

$$S(f_e) = \sum_{i=1}^{n} e^{\tau\lambda_i} \hat{f}_i^2, \tag{2}$$

where τ is a parameter and λ_i is the i-th eigenvalue of the graph Laplacian, which can be thought of as a discrete version of the frequency corresponding to the i-th element of the Fourier basis. The smoother f_e, the smaller $S(f_e)$.

Using (1) and (2), we can now reformulate the profile extraction problem as follows: find a profile v such that $V(f_e)$ be as large as possible, and $S(f_e)$ as small as possible. While this can be translated in many ways mathematically, we now present an approach derived from the theory of reproducible kernel Hilbert spaces (Schölkopf and Smola, 2002) which leads to a well-posed algorithm. The main trick to obtain such an algorithm is to express features, smoothness and variations in a dual form.

First of all, it can be shown that any profile of interest can be rewritten as a linear combination of expression profiles in the form:

$$v = \sum_{x\in\mathcal{X}} \alpha(x) e(x), \tag{3}$$

where $\alpha(.)$ is called the dual coordinate of v. If we call K the $n \times n$ matrix defined by $K_{x,y} = e(x)^T e(y)$, then a simple computation shows that the variation (1) captured by the feature f_e is given by:

$$V(f_e) = \frac{\alpha^T K^2 \alpha}{\alpha^T K \alpha}. \tag{4}$$

Second, let K' be the diffusion kernel Gram matrix of the graph Γ (Chung, 1997; Kondor and Lafferty, 2002),

i.e. the $n \times n$ matrix given by:

$$K' = \exp(-\tau L),$$

where L is the Laplacian matrix of the graph, $\tau > 0$ is a parameter, and 'exp' denotes the matrix exponential operation. K' being invertible any feature f can be uniquely written as $f = K'\beta$, or more explicitly:

$$f(.) = \sum_{x \in \mathcal{X}} \beta(x) K'(x, .). \quad (5)$$

With these notations, it can be shown that the smoothness functional $S(f)$ defined in (2) is equal to:

$$S(f) = \frac{\beta^T K' \beta}{\beta^T K'^2 \beta}. \quad (6)$$

Using the dual formulations (4) and (6), we can now formulate the correlation extraction problem as follows. Find dual coordinates α and β which maximize the functional:

$$\gamma(\alpha, \beta) \triangleq \frac{\alpha^T K K' \beta}{\left(\alpha^T \left(K^2 + \delta K\right) \alpha\right)^{\frac{1}{2}} \left(\beta^T \left(K'^2 + \delta K'\right) \beta\right)^{\frac{1}{2}}}, \quad (7)$$

where δ is a trade-off parameter. The maximization of (7) leads to a profile v given by (3), a corresponding feature f_e and a feature $f'(.)$ given by (5), such that:

- $V(f_e)$ be large,
- $S(f')$ be small,
- f_e and f' be as correlated as possible,

where δ controls the trade-off between these contradictory conditions. Indeed, when $\delta = 0$, the functional maximized in (7) is equal to the correlation coefficient between $f = K\alpha$ and $f' = K'\beta$. When δ increases, the correlation is penalized by $\alpha^T K \alpha$ and $\beta^T K' \beta$, which forces the solutions of (7) to have small $S(f')$ and large $V(f)$, by (4) and (6).

It turns out that (7) can be seen as a regularized form of canonical component analysis, equivalent to the following generalized eigenvalue problem:

$$\begin{pmatrix} 0 & KK' \\ K'K & 0 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \rho \begin{pmatrix} K^2 + \delta K & 0 \\ 0 & K'^2 + \delta K' \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \quad (8)$$

As pointed out in (Bach and Jordan, 2002; Vert and Kanehisa, 2002) this problem can be solved efficiently and results in a series of pairs of features:

$$\{(\alpha_i, \beta_i), i = 1, \ldots, n\}$$

with decreasing values of $\gamma(\alpha_i, \beta_i)$.

DATA

In order to test the method presented in the previous section we conducted experiments on publicly available expression data for the yeast *S.cerevisiae* using a curated database of metabolic pathways.

The gene graph was downloaded from the LIGAND database of chemical compounds of reactions in biological pathways (Goto *et al.*, 2002). This graph contains 774 nodes which represent 774 genes of the budding yeast *S.Cerevisiae*, linked through 16 650 edges, where two genes are linked when they code for proteins which have the possibility to catalyze two successive reactions in the LIGAND database (i.e. two reactions such that the main product of the first one is the main substrate of the second one).

We confronted this gene graph with two publicly available sets of expression data, downloaded from the Stanford Microarray Database (Sherlock *et al.*, 2001). The first data set is a collection of 18 measurements for 6198 yeast genes, collected every 7 minutes after cells were synchronized in G1 by addition of α factor (Spellman *et al.*, 1998). The analysis is restricted to the 756 genes of the LIGAND graph with an expression profile in this set. The second data set is a 7 time point series measured for 6199 yeast genes during the transition of an anaerobic growth to aerobic respiration, called diauxic shift (DeRisi *et al.*, 1997). Among these genes, 669 are present in the LIGAND graph. Following classical works (Eisen *et al.*, 1998) we work with the normalized logarithm of the ratio of expression levels of the genes between two experimental conditions. Missing values were estimated with the program KNNimput (Troyanskaya *et al.*, 2001). Each profile was then centered to zero mean and scaled to unit norm.

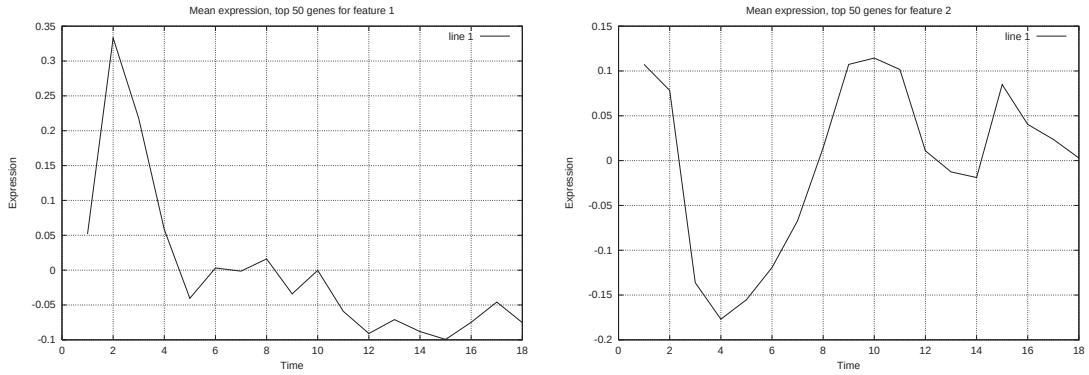
The generalized eigenvalue problem equivalent to (7) was solved with the free and publicly available program Octave[†]. Following experiments detailed in (Vert and Kanehisa, 2002) the regularization parameter δ of (7) was set to 0.01.

RESULTS

Alpha factor release dataset

This dataset was used in (DeRisi *et al.*, 1997) to detect genes whose expression exhibits periodicity related to the cell cycle. The profiles contain 18 points, hence 17 pairs or features with dual coordinates $(\alpha_i, \beta_i)_{i=1,\ldots,17}$ were extracted. The correlations between the corresponding pairs of features $(f_i, f'_i)_{i=1,\ldots,17}$ range from 0.62 to 0.36 (where $f_i = K\alpha_i$ and $f'_i = K'\beta_i$). This shows that the regularization parameter $\delta = 0.01$ is high enough to impose strong smoothness and relevance constraints on the fea-

[†] Available at http://www.octave.org



(a) First profile

(b) Second profile

Fig. 1. First 2 profiles extracted (α factor data set).

tures. Indeed, $\delta = 0$ (no smoothness or relevance constraint) would correspond to perfectly correlated features (all correlations being equal to 1), and the decrease from 1 to the actual correlations is the price to pay to ensure the smoothness of f_i' and the relevance of f_i in terms of variation captured.

Figure 1 shows the first two profiles extracted, and Table 1 contains a list representative of the genes with highest or lowest correlation with each profile, as well as the pathways they participate in in the KEGG database.

The first extracted profile is essentially a strong signal immediately following the begining of the experiment. Several pathways positively correlated with this pattern are involved in energy metabolism (oxidative phosphorylation, TCA cycle, glycerolipid metabolism), while pathways negatively correlated are mainly involved in protein synthesis (aminoacyl-tRNA biosynthesis, RNA polymerase, pyrimidine metabolism). Hence this profile clearly detects the sudden change of environment, and the priority to fuel the start of the cell cycle with fresh energetic molecules rather than to synthesize proteins.

The second extracted profile exhibits a strong sinusoidal shape corresponding to the progression in the cell cycle. Two cell cycles took place during the experiment, but the first one is more visible than the second one because the synchronization in the yeast colony decreased while the experiment progressed. Several genes directly involved in DNA synthesis (YNK1, RNR2, POL12) can be recognized in the list of genes anticorrelated with the second feature (corresponding to maximum expression in the S phase). Some pathways such as the starch metabolism have genes which exhibit either strong correlation or strong anticorrelation with the second profile, corresponding to the various regimes in the normal cell cycle (e.g. periods of energy storage alternate with periods of energy consumption).

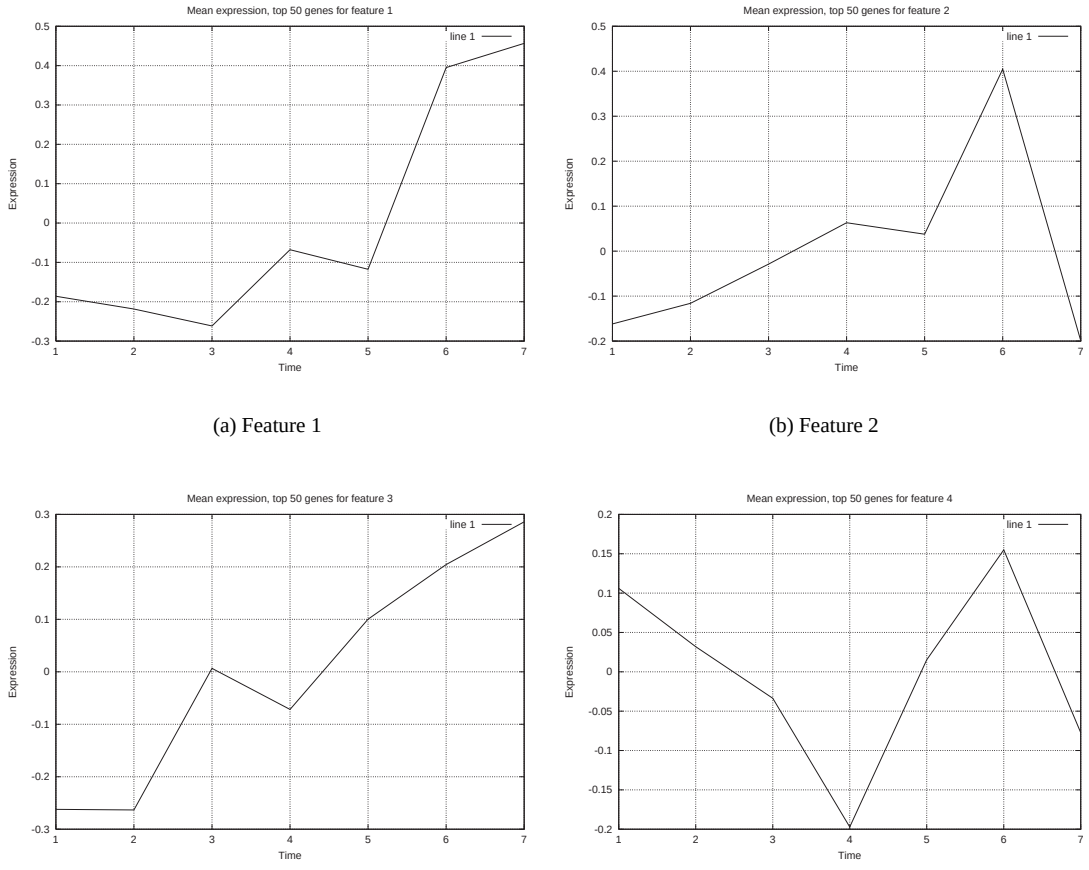
Diauxic shift data set

We performed a similar analysis of the diauxic shift data set (DeRisi *et al.*, 1997). Figure 2 shows the first four extracted profiles. We recover several typical curves already observed in (DeRisi *et al.*, 1997). The first feature f_1 is typical of an event which suddenly starts when all the glucose in consumed (between time points 5 and 6). f_2 corresponds to an event which progressively increases until time point 6, and suddenly decreases at the last time point (at the end of the diauxic shift), contrary to f_1. f_3 corresponds to a regular increase from the beginning until the last point, and can be thought of as an indicator of the diauxic shift progression. We also displayed feature 4, which is similar to feature 2 with the difference that the increase between points 1 and 6 is replaced by a two-stage process.

Table 2 shows the main pathways involved when one lists the set of genes with the highest and lowest 50 scores for each of the first 3 extracted features. As observed in (DeRisi *et al.*, 1997), we recover the activation of the TCA cycle (positive correlation with f_1 and f_3) used to oxidize pyruvate after all glucose is consumed, as well as the simultaneous activation of the urea cycle and the modification in the glycolysis and gluconeogenesis pathways studied in (DeRisi *et al.*, 1997). Observe that the correlation between f_2 and the TCA cycle is not detected here, which shows that the behavior at the last time point is an important difference between f_1 and f_2. Other observations include the fact that the porphyrin metabolism maps is also positively correlated with the third feature, at it is well known that

Table 1. Pathways and genes with highest and lowest scores on the first 2 features extracted

Feature	Correlation	Main pathways and genes
1	+	Glycolysis / Gluconeogenesis (PGK1, GPM2, ALD4,6), TCA cycle (CIT2, MDH1,2, SDH1, LSC1), Pentose phosphate pathway (RBK1, SOL4, ZWF1, YGR043C), Glycerolipid metabolism (GPD1,2,3, ALD4,6), Glyoxylate and dicarboxylate metabolism (MDH1,2, CIT2, ICL2), Sulfur metabolism (MET2,14,16,17).
1	-	Pyrimidine metabolism (RPA12,34,49,190, RPB2,5, RPC53, DUT1, TRR1, POL5, URK1, MIP1, PUS1), Purine metabolism (RPA12,34,49,190, RPB2,5, RPC53, CDC19, APT2, POL5, MIP1), Aminoacyl-tRNA biosynthesis (ILS1, FRS2, MES1, YHR020W, GLN4, ALA1, CDC60), Starch and sucrose metabolism (MPS1, HPR5, SWE1, HSL1, EXG1).
2	+	Pyrimidine metabolism (DEG1, PUS1,3,4, URA1,2, CPA1,2,FCY1), Folate biosynthesis (ENA1,5, BRR2, HPR5, FOL1), Starch and sucrose metabolism (ENA1,5, BRR2, HPR5, PGU1), Phenylalanine, tyrosine and tryptophan biosynthesis (TRP2,3,4, ARO2,7), Sterol biosynthesis (ERG7,12, HGM1,2).
2	-	Starch and sucrose metabolism (CDC7, ENA1, GIN4, HXK2, HPR5, SWE1, UGP1, HSL1, FKS1, MEK1), Purine and pyrimidine metabolism (POL12, ADK2, DUT1, RNR2, HYS2, YNK1, CDC21), Fructose and mannose metabolism (MNN1, PMI40, SEC53, HXK2), Cell cycle (CDC7, GIN4, SWE1, HSL1).



(a) Feature 1

(b) Feature 2

(c) Feature 3

(d) Feature 4

Fig. 2. First 4 features extracted (diauxic shift data set).

Table 2. Pathways and genes with highest and lowest scores on the first features extracted (diauxic shift data set)

Feature	Correlation	Main pathways and genes
1	+	TCA cycle (SDH1,2,3,4, LSC1,2, YJL045W, YLR164W, YMR118C), Oxidative phosphorylation (SDH1,2,3,4, ISP, Cytc1, COR1, QCR2,6,7,8,9), Urea cycle and metabolism of amino groups (ARG3,4,5,8, CAR1,2, SPE1)
1	-	Sterol biosynthesis (ERG1,7,8,12, HMG1), Purine metabolism (ADE1,5,6,7,8, AAH1, ADE2), N-Glycans biosynthesis (ROT2, ALG5,7, SEC59, DPM1), Aminoacyl-tRNA biosynthesis (ILS1, YDR341C, YER087W, MES1, FRS1, WRS1), Phenylalanine, tyrosine and tryptophan biosynthesis (TRP2,3,4, FRS1).
2	+	Glycolysis / Gluconeogenesis (GPM1,2,3, PGK1, TPI1), Glycerolipid metabolism (CRD1, TPI1, CHO1, DAK1,2, PCT1, GUT1), Pyrimidine metabolism (DUT1, TRR2, DCD1, CDD1, URA4, CDC21), Biotin metabolism (BPL1, YFR006W, BIO2, SPC3, RCE1)
2	-	Urea cycle and metabolism of amino groups (ARG3,5,8, SPE1), Glutamate metabolism (URA2, CPA1,2, GLN4), Phenylalanine, tyrosine and tryptophan biosynthesis (TRP2,3,4, ARO2,7), Starch and sucrose metabolism (ENA1,5, BRR2, HPR5, PGU1)
3	+	TCA cycle (SDH1,2,3,4, LSC1,2, YJL045W, YLR164W, YMR118C), Porphyrin metabolism (HEM2,4,14,15, CYC3).
3	-	Urea cycle and metabolism of amino groups (SPE1,3,4, PRO3, ARG8), Aminoacyl-tRNA biosynthesis (ILS1, YDR341C, YER087W, FRS1,2, VAS1, TYS1, YHR020W, THS1, MST1, ISM1, CDC60), Phenylalanine, tyrosine and tryptophan biosynthesis (TRP2,3,5, FRS1,2, TRP3,5, ARO2,7), Arginine and proline metabolism (YDR341C, PRO3, YER087W, YHR020W, SPE1,2,3,4), Pyrimidine metabolism (TRR1,2, DEG1, CDD1, PUS1,4), Valine, leucine and isoleucine biosynthesis (ILS1, VAS1, ILV1, ISM1, CDC60).

HEM2,4,14,15 and CYC3 participate in the porphyrin pathway which generates cytochrome c oxidase (which enhances the capability of yeast to produce ATP in the respiratory chain). In parallel, several pathways involved in biosynthesis of various molecules (sterol biosynthesis, purine and pyrimidine metabolism) exhibit a strongly negative correlation with profiles increasing along time, corresponding to the adaptation of the cell to a lower level of activity following the decrease of glucose in the environment.

DISCUSSION AND CONCLUSION

The method presented in this paper provides a way to compare a graph and a set of profiles. We focused on its ability to extract meaningful profiles of expression, as well as the corresponding metabolic pathways, in a fully automated way. It can therefore be used as a data mining tool, to automatically make sense out of expression data in terms of pathway activity.

A second possible application of this method is dimensionality reduction of microarray data. Indeed profiles which 'make sense' with respect to the graph topology are automatically detected: as a result, dimensionality reduction can be performed by simply keeping the first k extracted features, where k is smaller than the total number of points available, and representing a gene expression profile $e(x)$ by the smaller vector $(f_1(x), \ldots, f_k(x))$. This approach is explored in the companion paper (Vert and Kanehisa, 2002).

One particularity of our method is that it is able, up to some extent, to deal with noise and errors as well in the graph as in the expression data. Indeed, by looking for correlated features, the particularities of single genes disappear behind general trends of sets of genes. For example, if some edges are wrongly placed in the gene graph, the smoothness of features might still be detected as long as the topology of the graph in not too much modified. We plan to investigate in the future how the tools developed here can be used to automatically remove wrong edges or add new ones in the gene network, which is of particular importance as well for pathway analysis as for other networks such as protein interaction networks.

Finally, the algorithm presented in this paper should be considered as a first step toward integration of various kinds of data using kernel methods. Indeed, on an technical point of view, the only data required to perform our analysis are the diffusion kernel matrix K' computed from the graph on the one hand, and the Gram matrix of expression profiles inner products K. In other words, the graph topology as well as the set of expression profiles are encoded in a similar form (a Gram matrix of a kernel function), and a generalized form of CCA is performed between the two kernels. It turns out that many kernels for various types of gene representations

(different from expression profiles and nodes of a graph) have been developed in the last few years, including but not limited to kernels for aminoacid sequences (Jaakkola *et al.*, 2000), for phylogenetic profiles (Vert, 2002) or for promoter regions (Pavlidis *et al.*, 2001). Using such kernels in the place of K and K' gives a way to extract correlations not only between gene networks and expression data, but also with protein sequences, phylogenetic profiles or promoter regions. Moreover, as pointed out in (Bach and Jordan, 2002), the notion of correlation can be extended to more than two variables, so one can imagine looking for correlations between all these kinds of data simultaneously. A first attempt in this direction is presented in (Yamanishi *et al.*, 2003) for the purpose of operon detection in bacterial genomes, and we are currently investigating further extensions of this ideas.

ACKNOWLEDGMENTS

This work was supported by the Research for the Future Program of the Ministry of Education, Culture, Sport, Science and Technology of Japan.

REFERENCES

Bach,F.R. and Jordan,M.I. (2002) Kernel independent component analysis. *J. Machine Learning Res.*, **3**, 1–48.

Chung,F.R. (1997) Spectral graph theory. volume 92 of CBMS Regional Conference Series, American Mathematical Society, Providence.

de Jong,H. (2002) Modeling and simulation of genetic regulatory systems: a literature review. *J. Comput. Biol.*, **9**, 67–103.

DeRisi,J.L., Iyer,V.R. and Brown,P.O. (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, **278**, 680–686.

Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.

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

Hanisch,D., Zien,A., Zimmer,R. and Lengauer,T. (2002) Co-clustering of biological networks and gene expression data. *Bioinformatics*, **18**, 145S–154S.

Jaakkola,T., Diekhans,M. and Haussler,D. (2000) A discriminative framework for detecting remote protein homologies. *J. Comput. Biol.*, **7**, 95–114.

Kanehisa,M., Goto,S., Kawashima,S. and Nakaya,A. (2002) The KEGG databases at GenomeNet. *Nucleic Acid Res.*, **30**, 42–46.

Kondor,R.I. and Lafferty,J. (2002) Diffusion kernels on graphs and other discrete input. In *Proceedings of ICML 2002*. pp. 315–322.

Pavlidis,P., Furey,T.S., Liberto,M., Haussler,D. and Grundy,W.N. (2001) Promoter region-based classification of genes. In *Pacific Symposium on Biocomputing*. pp. 139–150.

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

Sherlock,G., Hernandez-Boussard,T., Kasarskis,A., Binkley,G., Matese,J., Dwight,S., Kaloper,M., Weng,S., Jin,H., Ball,C., Eisen,M. and Spellman,P. (2001) The Stanford Microarray Database. *Nucleic Acid Res.*, **29**, 152–155.

Spellman,P.T., Sherlock,G., Zhang,M.Q., Iyer,V.R., Anders,K., Eisen,M.B., Brown,P.O., Botstein,D. and Futcher,B. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell*, **9**, 3273–3297.

Troyanskaya,O., Cantor,M., Sherlock,G., Brown,P., Hastie,T., Tibshirani,R., Botstein,D. and Altman,R.B. (2001) Missing value estimation methods for NA microarrays. *Bioinformatics*, **17**, 520–525.

van Helden,J., Gilbert,D., Wemisch,L., Schroeder,K. and Wodak,S. (2000) Application of regulatory sequence analysis and metabolic network analysis to the interpretation of gene expression data. *Lecture Notes in Computer Science*, **2066**, 155–172.

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

Vert,J.-P. and Kanehisa,M. (2002) Graph-driven features extraction from microarray data. Preprint arXiv physics/0206055.

Yamanishi,Y., Vert,J.-P., Nakaya,A. and Kanehisa,M. (2003) Extraction of correlated gene clusters from multiple genomic data by generalized kernel canonical correlation analysis. *Bioinformatics*, To appear.

Zien,A., Küffner,R., Zimmer,R. and Lengauer,T. (2000) Analysis of gene expression data with pathway scores. In R.A. *et al.*, (ed.), *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology*. AAAI Press, pp. 114–120.