BIOINFORMATICS

Vol. 1 no. 1 2004
Pages 1–8

Protein Network Inference from Multiple Genomic Data: A Supervised Approach

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

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

ABSTRACT

Motivation: An increasing number of observations support the hypothesis that most biological functions involve the interactions between many proteins, and that the complexity of living systems arises as a result of such interactions. In this context the problem of inferring a global protein network for a given organism, using all available genomic data about the organism, is quickly becoming one of the main challenges in current computational biology.

Results: This paper presents a new method to infer protein networks from multiple types of genomic data. Based on a variant of kernel canonical correlation analysis, its originality is in the formalization of the protein network inference problem as a supervised learning problem, and in the integration of heterogeneous genomic data within this framework. We present promising results on the prediction of the protein network for the yeast *Saccharomyces cerevisiae* from four types of widely available data: gene expressions, protein interactions measured by yeast two-hybrid systems, protein localizations in the cell, and protein phylogenetic profiles. The method is shown to outperform other unsupervised protein network inference methods. We finally conduct a comprehensive prediction of the protein network for all proteins of the yeast, which enables us to propose protein candidates for missing enzymes in a biosynthesis pathway.

Availability: Softwares are available upon request.

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

INTRODUCTION

An increasing number of observations support the hypothesis that most biological functions involve the interactions between many proteins, and that the complexity of living systems arises as a result of such interactions. In this context the problem of inferring a global protein network for a given organism, using all available genomic data about the organism, is quickly becoming one of the main challenges addressed in current computational biology. By protein network, we mean in this paper a graph with proteins as vertices and with edges that correspond to various binary relationships between proteins. More precisely we consider below the protein network with edges between two proteins if (i) the proteins interact physically, or (ii) the proteins are enzymes that catalyze two successive chemical reactions in a pathway, or (iii) one of the proteins regulates the expression of the other. This definition of protein network involves various forms of interactions between proteins which should be taken into account for the study of the behavior of biological systems.

Unfortunately, the experimental determination of this protein network remains very challenging nowadays, even for the most basic organisms. The lack of reliable information contrasts with the wealth of genomic data generated by high-throughput technologies such as gene expression data (Eisen *et al.*, 1998), physical protein interactions (Ito *et al.*, 2001), protein localization (Huh *et al.*, 2003), phylogenetic profiles (Pellegrini *et al.*, 1999), or pathway knowledge (Kanehisa *et al.*, 2004). There is therefore an incentive to develop methods to predict the protein network from such data.

A variety of computational methods for this problem have been investigated so far. Some methods perform the protein network inference from a single type of genomic data, such as Bayesian networks (Friedman *et al.*, 2000) and Boolean networks (Akutsu *et al.*, 2000), which aim at inferring gene regulation networks from gene expression data, or the mirror tree method (Pazos *et al.*, 2001), which predicts protein interactions from evolutionary similarities. Other methods combine different sources of data to infer the network: this is for example the case in the joint graph method (Marcotte *et al.*, 1999), where graphs representing similarities with respect to various types of genomic information are overlapped in order to detect strong associations between proteins.

These methods share the particularity of being *unsupervised*, in the sense that the whole protein network is inferred from the data. Inference typically relies on the assumption that proteins sharing similarity according to a dataset (e.g., co-expression or co-evolution), are more likely to be linked than others. The reliable a priori knowl-

*To whom correspondence should be addressed.

edge about the protein network, such as experimentally determined protein interactions, is usually not used in the inference process itself, but rather as a way to assess the accuracy of the inference engine.

In this paper, we propose a method to infer protein networks from multiple heterogeneous genomic datasets in a *supervised* context. By supervised we mean that the reliable a priori knowledge about parts of the true protein network is used in the inference process itself. The supervised approach is a two-step process. First, a model is learned to explain the "gold standard" from available datasets. Second, this model is applied to proteins absent from the "gold standard", in order to infer their interactions. While supervised classification is a classical paradigm in machine learning and statistics, most methods can not be adapted directly to the network inference problem, because the goal is to predict properties *between* proteins, not about individual proteins. In order to develop an algorithm adapted to this context, we propose a method borrowing ideas from spectral clustering (Ng *et al.*, 2002) and finally equivalent to kernel canonical correlation analysis (CCA) (Akaho , 2001) in order to detect correlations between heterogeneous datasets, in particular between a protein network and other genomic attributes. Kernel CCA has received a lot of attention in computational biology recently, appearing as a useful approach to predict gene functions (Vert *et al.*, 2003a), extract active metabolic pathway from gene expression (Vert *et al.*, 2003b), or detect operon structures from pathways, genomes and expression data (Yamanishi *et al.*, 2003).

The method is tested on its ability to predict the protein network of *Saccharomyces cerevisiae* from four datasets for proteins: gene expression data obtained from DNA microarrays, noisy protein interaction data obtained by yeast two-hybrid systems, localization data, and sequence data encoded into phylogenetic profiles. It compares favorably to two other unsupervised method we propose, one based on the assumption that similar proteins (in the sense of the available datasets) should interact, the other based on a spectral clustering approach. The systematic experiments we conduct highlight the accuracy improvement resulting from the integration of heterogeneous data, and from the supervised learning approach. Finally we perform a comprehensive prediction of the protein network for all proteins of the yeast, which enables us to propose protein candidates for missing enzymes in biosynthesis pathways.

MATERIALS

Protein network data As a gold standard for part of the protein network of *Saccharomyces cerevisiae*, we take the KEGG/PATHWAY database (Kanehisa *et al.*, 2004) which is a graph with proteins as vertices, and with three types of edges: enzyme-enzyme relations when two proteins are enzymes that catalyze successive reactions in a known pathway, direct physical protein-protein interactions, and gene expression regulation between a transcription factor and its target gene products. The resulting protein network, that contains 769 nodes and 7404 edges, is regarded as a reliable part of the global protein network to be inferred below.

Expression data Expression data corresponding to 157 experiments (77 experiments in Spellman *et al.*, 1998 and 80 experiments in Eisen *et al.*, 1998) were used. To each protein is therefore associated a vector of dimension 157.

Protein interaction data We used 5470 interacting protein pairs, detected from several yeast two hybrid (Y2H) experiments (Ito *et al.*, 2001; Uetz *et al.*, 2000). Because the Y2H method is known to introduce many false positives, this dataset should be considered as a very noisy version of the physical interaction part of the true protein network.

Localization data The localization data were obtained from the budding yeast localization (Huh *et al.*, 2003). This dataset describes localization information of proteins in 23 intracellular locations such as mitochondrion, Golgi, and nucleus. To each protein is therefore attached a string of 23 bits, in which the presence and absence of the protein in a certain intracellular location is coded as 1 and 0, respectively, across the 23 intracellular locations.

Phylogenetic profile Phylogenetic profiles were constructed from the ortholog clusters in the KEGG database, which describes the sets of orthologuous proteins in 145 organisms. In this study, we focus on the organisms with fully sequenced genomes, including 11 eukaryotes, 16 archaea, and 118 bacteria. Each phylogenetic profile consists of a string of bits, in which the presence and absence of an orthologous protein is coded as 1 and 0, respectively, across the above 145 organisms.

METHODS

Data representation and integration by kernels

Kernel representation In order to represent each type of genomic information described in the previous section into a coherent and useful mathematical framework, we first transform each dataset into a symmetric positive definite kernel function (simply called *kernel* below), that is, a real-valued function $K(\mathbf{x}, \mathbf{y})$ satisfying $K(\mathbf{x}, \mathbf{y}) = K(\mathbf{y}, \mathbf{x})$ for any two proteins $\mathbf{x}$ and $\mathbf{y}$, and $\sum_{i=1}^{n} a_i a_j K(\mathbf{x}_i, \mathbf{x}_j) \geq 0$ for any integer n, set of proteins $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ and set of real numbers $(a_1, \ldots, a_n)$ (Schölkopf *et al.*, 2002). Intuitively, the kernel corresponding to a given dataset can be thought of as a measure of similarity between proteins with respect to the dataset. For example, when a dataset assigns a vector to each protein (such as expression, localization data or phylogenetic profiles), the Gaussian RBF kernel $k(\mathbf{x},\mathbf{y}) = \exp(- \parallel \mathbf{x}-\mathbf{y} \parallel^2 / 2\sigma^2)$ or the linear kernel $k(\mathbf{x},\mathbf{y}) = \mathbf{x}\cdot\mathbf{y}$ are natural candidates. When a dataset consists in a graph of proteins (such as the gold standard protein network or the noisy protein interactions), then a natural choice is the diffusion kernel (Kondor *et al.*, 2002) defined as the matrix $K = \exp(\beta H)$, where $\beta > 0$ is a parameter and H is the opposite Laplacian matrix of the graph ($H = A - D$ where A is the adjacency matrix and D is the diagonal matrix of node connectivity). The two main motivations behind representing all datasets by kernel functions are first that all types of data are encoded in the same mathematical framework even though they might be different by nature (e.g., vectors, strings, graphs), and second that this choice paves the way to the use of kernel methods (Schölkopf *et al.*, 2002).

Data integration In this study we use $P \geq 1$ sorts of heterogeneous genomic data in order to predict the protein network, which are represented by P kernels $K_1, \ldots, K_P$. The function K_p measures the similarity of proteins with respect to the p-th dataset. A simple data integration can be performed by creating a new kernel as the sum the kernels corresponding to different genomic data, namely $K = \sum_{p=1}^{P} K_p$. While more complex approaches can be imagined to combine heterogeneous data through kernel operation, this simple operation has proved to be useful in (Pavlidis *et al.*, 2001; Yamanishi *et al.*, 2003) and is used below.

Direct approach for protein network prediction

We consider the problem of predicting the protein network of *S. cerevisiae* from several genomic datasets. As a first direct inference method, under the assumption that connected proteins are likely to share similarities in the datasets, we propose to predict an edge between two proteins $\mathbf{x}$ and $\mathbf{y}$ when the value $K(\mathbf{x},\mathbf{y})$ is large enough. Depending on the choice of K, this covers the situations of selecting proteins with correlated expression, similar profiles, similar localization, or all of these simultaneously. For a fixed choice of K, a predicted network can be progressively built by starting from isolated nodes and adding edges between pairs of proteins with decreasing kernel values. The discrete version of this approach, which we call the direct approach below, is related to the joint graph method (Marcotte *et al.*, 1999).

Unsupervised spectral approach for protein network prediction

Spectral clustering (Ng *et al.*, 2002) has attracted a lot of attention recently and led to impressive results in complex clustering tasks. Given a set of points (e.g., proteins) to cluster, the idea of spectral clustering is to map them onto a feature space where clusters are easier to detect, before applying a classical clustering algorithm. The feature space is defined as the linear span of the first eigenvectors of a similarity matrix between the points. In case one has a kernel to define the similarity between points, then kernel principal component analysis (PCA) (Schölkopf *et al.*, 1998) is known to be related to spectral clustering: the feature space spanned by the first few principal components (PCs) is also a space where clusters can be easier to detect (Bengio *et al.*, 2003). Kernel PCA can be shortly summarized as follows. Given a set of N proteins $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ and a kernel function $K : \mathcal{X}^2 \to \mathbf{R}$, one considers the set $\mathcal{H}$ of real-valued functions $\left\{ f(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i K(\mathbf{x}_i, \mathbf{x}), (\alpha_1, \ldots, \alpha_N) \in \mathbf{R}^N \right\}$ endowed with the norm $||f||_{\mathcal{H}} = \sum_{i,j} \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j)$. The projection onto the first principal direction is defined up to a scaling factor as the function $f^{(1)} \in \mathcal{H}$ that minimizes $||f^{(1)}||_{\mathcal{H}}$ under the constraint $\sum_{i=1}^{N} f^{(1)}(\mathbf{x}_i)^2 = 1$. The projections onto the following principal directions are defined recursively in the same way with the additional orthogonality constraint $\sum_{i=1}^{N} f^{(l)}(\mathbf{x}_i) f^{(m)}(\mathbf{x}_i) = 0$ if $l < m$. If $K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i \cdot \mathbf{x}_j$ for vectors, then one recovers the classical PCA method. As a result, spectral clustering suggests to represent the point $\mathbf{x}_i$ by the vector $(f^{(1)}(\mathbf{x}_i), \ldots, f^{(L)}(\mathbf{x}_i))^\top$ with $L < N$, before performing classical clustering on these representations.

Even though our concern is not directly on gene clustering, the problem of network reconstruction bears similarity with clustering. It can be thought of as an extreme clustering problem, where one looks for clusters of two points (that correspond to connected protein pairs in the network). Given a kernel K between proteins, this suggests an alternative to the direct approach: first project all proteins onto the subspace defined by the first few PCs obtained by kernel PCA, and then select pairs of similar points in this feature space.

Supervised approach for protein network prediction

The actual problem we are confronted with is illustrated in Figures 1 and 2: we would like to infer a protein network from a lot of noisy data about the proteins in Fig. 2, and we already know with some confidence part of the network to be inferred. This prior knowledge is depicted in Fig. 1, where we assume that the protein network restricted to $n < N$ proteins is known, N being the total number of proteins. Both the direct approach and the spectral approach are unsupervised, in the sense that they don't use the prior information illustrated in Fig. 1 but rather directly infer a network from the data illustrated in Fig. 2.

In contrast, we propose in this section a *super-*

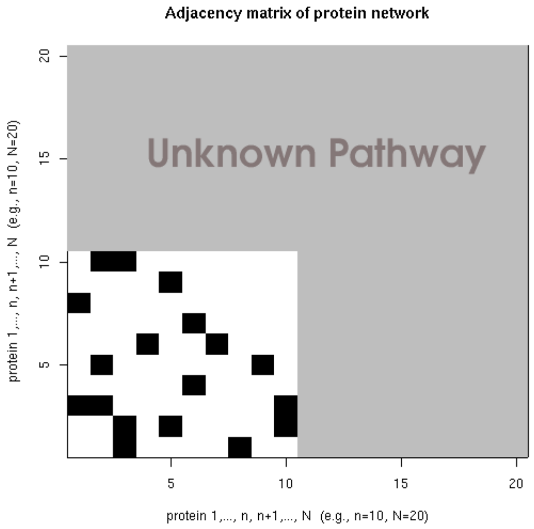

Fig. 1. An example of adjacency matrix of proteins in protein network

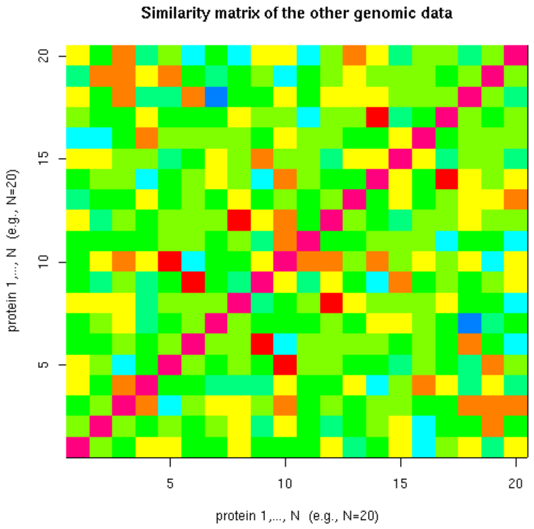

Fig. 2. An example of similarity matrix of proteins in the other genomic data

vised method to infer the network from both the data and the prior knowledge. The method is a slight modification of the unsupervised spectral approach described in the previous section. In the spectral approach, each protein $\mathbf{x}$ is first represented by a vector $f(\mathbf{x}) = \left(f^{(1)}(\mathbf{x}), \ldots, f^{(L)}(\mathbf{x})\right)^{\top}$, where $L < N$ and $f^{(l)}(\mathbf{x})$ is the projection of $\mathbf{x}$ onto the l-th principal component. The goal of this projection is to define a feature space where pairs of interacting proteins have similar projection, so that it becomes possible to infer interaction from similarity in the feature space. Hence, whenever $\mathbf{x}_i$ interacts with $\mathbf{x}_j$, we would like $f(\mathbf{x}_i)$ to be similar to $f(\mathbf{x}_j)$, which ideally would be fulfilled if $f^{(l)}(\mathbf{x}_i)$ was close to $f^{(l)}(\mathbf{x}_j)$ for each $l = 1, \ldots, L$. Consequently an "ideal" feature space, if the protein network was known beforehand, would be a subspace defined by functions $f^{(l)}$ $(l = 1, \ldots, L)$ that vary slowly between adjacent nodes of the protein network. Such functions are usually called *smooth*, and it is known (Vert *et al.*, 2003a) that the norm $||f||_{\mathcal{H}}$ associated with a diffusion kernel on a graph exactly quantifies this smoothness: the smoother f, the smaller $||f||_{\mathcal{H}}$. As a result, if the protein network was known, an ideal feature space would be defined by the projection onto the first principal directions defined by kernel PCA with a diffusion kernel on the graph.

As the total protein network is not known beforehand, the projections onto this ideal feature space can not be computed, as opposed to the projections in the unsupervised spectral approach. In order to improve the representation provided by the spectral approach, we propose to constrain it to somehow fit the ideal feature space, at least on the part of the network known beforehand. This can be done as follows. Let $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ be the n proteins in the "gold standard", and $\{\mathbf{x}_{n+1}, \ldots, \mathbf{x}_N\}$ be the remaining proteins whose participation in the protein network must be inferred (Fig. 1). Let K_1 be the kernel representing the genomic information restricted to the n first proteins, and K_2 be the diffusion kernel derived from the known protein network. Both K_1 and K_2 are then $n \times n$ matrices. For any function f defined on $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, let $||f||_1$ and $||f||_2$ be the corresponding norms. In order to define a feature f such that $||f||_1$ be small, as in the spectral approach, and $||f||_2$ be small simultaneously, as in the ideal representation, we propose to use the following trick: find two functions f_1 and f_2 such that $\sum_{i=1}^{n} f_k(\mathbf{x}_i)^2 = 1$ for $k = 1, 2$, and that maximize the functional

$$corr\,(f_1, f_2) \times \frac{1}{\sqrt{1 + \lambda_1 ||f_1||_1^2}} \times \frac{1}{\sqrt{1 + \lambda_2 ||f_2||_2^2}}, \tag{1}$$

where λ_1 and λ_2 are positive regularization parameters, and $corr\,(f_1, f_2)$ is the correlation coefficient between f_1 and f_2. The first term of this product ensures that f_1 "fits" f_2 on the a priori known part of the network, while the second and last terms ensure that $||f_1||_1$ and $||f_2||_2$ are small simultaneously. Subsequent features can

Table 1. List of experiments of direct approach, spectral approach based on kernel PCA, and supervised approach based on kernel CCA

Approach	Kernel (Predictor)
Direct	K_{exp} (Expression)
Direct	K_{ppi} (Protein interaction)
Direct	K_{loc} (Localization)
Direct	K_{phy} (Phylogenetic profile)
Direct	K_{exp}+K_{ppi}+K_{loc}+K_{phy} (Integration)

Approach	Kernel (Predictor)
Spectral	K_{exp} (Expression)
Spectral	K_{ppi} (Protein interaction)
Spectral	K_{loc} (Localization)
Spectral	K_{phy} (Phylogenetic profile)
Spectral	K_{exp}+K_{ppi}+K_{loc}+K_{phy} (Integration)

Approach	Kernel (Predictor)	Kernel (Target)
Supervised	K_{exp} (Expression)	K_{gold} (Protein network)
Supervised	K_{ppi} (Protein interaction)	K_{gold} (Protein network)
Supervised	K_{loc} (Localization)	K_{gold} (Protein network)
Supervised	K_{phy} (Phylogenetic profile)	K_{gold} (Protein network)
Supervised	K_{exp}+K_{ppi}+K_{loc}+K_{phy} (Integration)	K_{gold} (Protein network)

be defined recursively by minimizing the same functional with additional orthogonality conditions. The main reason for using the functional (1) is that it can be shown (Akaho, 2001; Bach and Jordan, 2002) to be equivalent to the following generalized eigenvalue problem:

$$\begin{pmatrix} \mathbf{0} & K_1K_2 \\ K_2K_1 & \mathbf{0} \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} = \rho \begin{pmatrix} (K_1+\lambda_1 I)^2 & \mathbf{0} \\ \mathbf{0} & (K_2+\lambda_2 I)^2 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix}, \quad (2)$$

where I is the identity matrix. Indeed, the successive solutions to eq.(1) can be written as $f_1 = K_1\alpha_1$ and $f_2 = K_2\alpha_2$, where α_1 and α_2 are the eigenvectors of eq.(2) with decreasing eigenvalue ρ. This problem is usually called kernel canonical correlation analysis (CCA) (Akaho, 2001). If one now focuses on the first L solutions $\alpha_1^{(1)}, \ldots, \alpha_1^{(L)}$ of eq.(2) (sorted by decreasing value of ρ), then they define L features of interest by $f^{(l)} = K_1\alpha_1^{(l)}$, for $l = 1, \ldots, L$. These features are built from the genomic dataset kernel K_1 only, and are expected to fit the ideal features on the gold standard set of proteins. These features can now be generalized to any protein $\mathbf{x}$ by the following equation:

$$f^{(l)}(\mathbf{x}) = \sum_{k=1}^{n} \alpha_1^{(l)}(\mathbf{x}_k) K(\mathbf{x}_k, \mathbf{x}). \quad (3)$$

This is the set of features we propose to map the proteins to before inferring protein interactions.

In both the spectral method and this supervised method, each protein $\mathbf{x}$ is mapped to a feature space as a L-dimensional vector $\mathbf{u} = (u_1, \ldots, u_L)^\top = (f^{(1)}(\mathbf{x}), \ldots, f^{(L)}(\mathbf{x}))^\top$. To assess the similarity of protein $\mathbf{x}$ and protein $\mathbf{y}$ in this feature space, we simply follow the spirit of the direct approach and quantify the similarity between points $\mathbf{u} = (u_1, \ldots, u_L)^\top$ and $\mathbf{v} = (v_1, \ldots, v_L)^\top$ by their correlation:

$$\widehat{corr}(\mathbf{u}, \mathbf{v}) = \frac{\widehat{cov}(\mathbf{u}, \mathbf{v})}{\sqrt{\widehat{var}(\mathbf{u})}\sqrt{\widehat{var}(\mathbf{v})}} = \frac{\frac{1}{L}\sum_{l=1}^{L}(u_l - \bar{\mathbf{u}})(v_l - \bar{\mathbf{v}})}{\sqrt{\frac{1}{L}\sum_{l=1}^{L}(u_l - \bar{\mathbf{u}})^2}\sqrt{\frac{1}{L}\sum_{l=1}^{L}(v_l - \bar{\mathbf{v}})^2}}, \quad (4)$$

where $\bar{\mathbf{u}}$ and $\bar{\mathbf{v}}$ are the averages of $\mathbf{u}$ and $\mathbf{v}$.

RESULTS

All genomic datasets are transformed into kernels as follows. The gold standard protein network and the noisy protein interaction datasets are represented by a diffusion kernel with parameter $\beta = 1$, and respectively denoted K_{gold} and K_{ppi}. For the gene expression data, we used the Gaussian RBF kernel with $\sigma = 5$, and denote the resulting kernel K_{exp}. For both localization data and the phylogenetic profiles, a simple linear kernel, denoted respectively K_{loc} and K_{phy}. All kernels are then normalized to 1 on the diagonal and centered in the feature space (Schölkopf *et al.*, 2002).

We tested the direct and spectral approaches either on single types of genomic datasets, or on the integrated kernel representing all datasets. For the spectral approach, we arbitrarily kept the first $L = 50$ principal components

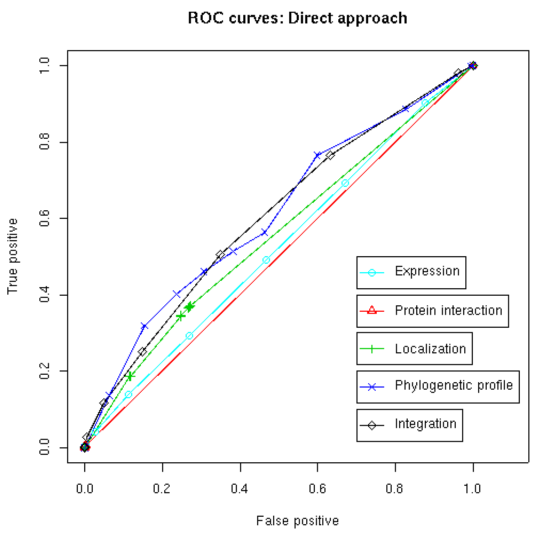

Fig. 3. ROC curves: Direct approach

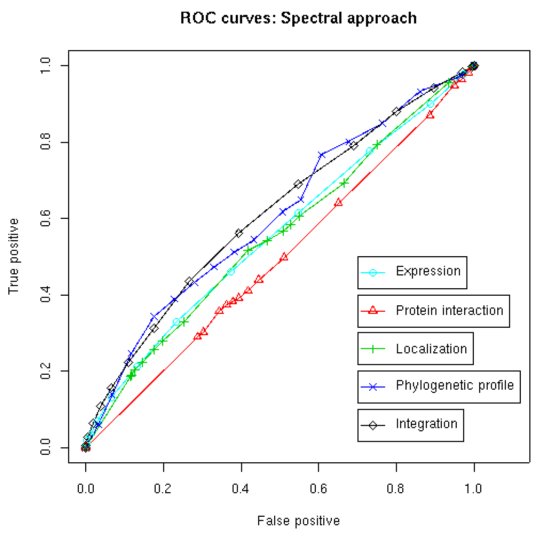

Fig. 4. ROC curves: Spectral approach

to define the feature space. The accuracy of both methods is assessed on the gold standard dataset, by their capacity to recover the protein network. Starting from isolated nodes, each method can be used to progressively build a network by adding edges between pairs of proteins sorted by decreasing similarity. At each addition, we recorded the number of true positives (predicted edges that indeed are present in the gold standard) and false positives (predicted edges that are absent from the gold standard). Figures 3 and 4 show the ROC curves representing the numbers of true positives as a function of the number of false positives for the two methods. In both cases, the overall accuracy of the inference method is very limited. Little information seems to be caught by the direct approach, while the spectral approach gives slightly better results, in particular when used in combination with the kernel that integrates all genomic datasets, but remains useless in practice due to the large rate of false positives at any rate of true positives. These negative results, in particular for the direct approach, confirm that the problem of protein network reconstruction is far from trivial.

We then tested the supervised approach. The parameters λ_1 and λ_2 were set to 0.1, and again we kept $L = 50$ features to define the feature space. We tested various combinations of dataset kernels to be fitted to the gold standard kernel, as described in Table 1. In order to assess the accuracy of the method, we carried out a 10-fold cross-validation experiment as follows. In each out of 10 iterations, the set of 769 proteins in the gold standard is split into a training set and a test set in the proportion $9/1$. The feature space is trained on the training set, and the inference of interaction is performed on the possible interactions involving the proteins in the test set (the gray part in Fig. 1). Once again a graph is progressively built and we record the number of true positive interactions as a function of false positives. The resulting ROC curves averaged over 10 iterations are plotted in Figure 5. As opposed to the direct and spectral approaches, we observe here that the supervised approach is able to catch information about the protein network and make interesting prediction. Among all single datasets, expression and phylogenetic profiles seem to provide similar amount of information, followed by localization data and noisy protein interactions. The supervised method applied in conjunction with the integration of all four datasets gives the overall best results. The comparison of these experimental results highlights the accuracy improvements resulting from both the integration of multiple dataset, and the use of a supervised approach.

Finally we investigated the effect of the number of features L on the performance of the spectral and supervised approaches. In both cases, we used the integrated kernel that represents all genomic dataset, and varied the number of features L from 10 to 400. Figures 6 shows the area under the ROC curves obtained by both approaches for varying L, where triangles and squares indicate spectral and supervised approaches, respectively. The supervised approach seems to be sensitive to the number of features, with a maximum reached for $L = 40$. To the contrary, the spectral approach seems to have little variability when

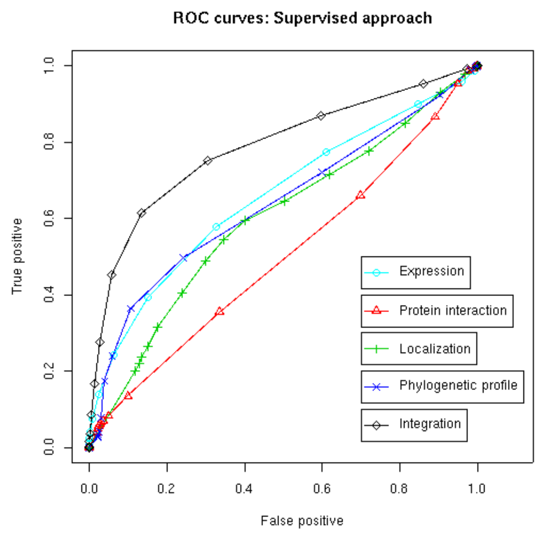

Fig. 5. ROC curves: Supervised approach

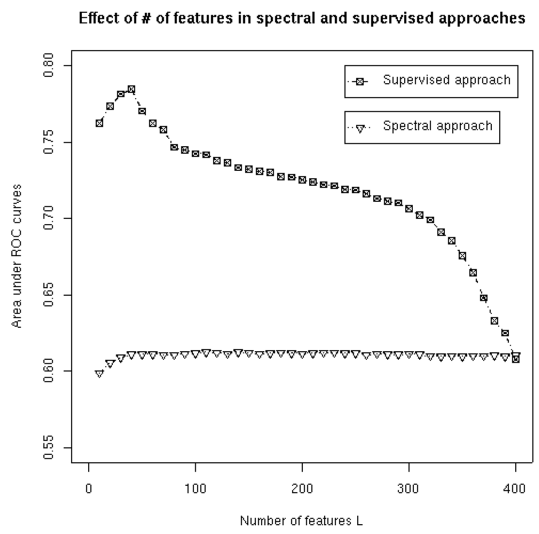

Fig. 6. Effect of number of features L in spectral and supervised approaches.

the number of features varies. This result suggests that we need to choose an appropriate number of features in actual applications of the supervised approach.

The validity of the supervised method being confirmed by these experiments, we then conducted a comprehensive prediction of protein network for all the proteins (6059 ORFs in this study) of the yeast. The predicted network enabled us to make new biological inferences about unknown protein interactions, but also about missing enzymes in biochemical pathways. As an example, there is a missing enzyme (EC:2.4.1.141) between EC:2.7.8.15 and EC:2.4.1.142 in the N-Glycans biosynthesis pathway (see Figure 7). From the predicted protein network with a threshold set to 0.6, we regard YPL207W and YGL010W as candidates for the missing enzyme, because they have high scores with both EC:2.7.8.15 and EC:2.4.1.142. According to the annotation, they are hypothetical proteins, and the other high scoring proteins are glycosyltransferase. Therefore, we can guess that they might work as an enzyme catalyzing the chemical reaction. Of course, such inference can be applied to missing enzymes in other pathways. The results of the whole protein network predicted can be obtained from the author's website (http://web.kuicr.kyoto-u.ac.jp/~yoshi/ismb04/).

DISCUSSION AND CONCLUSION

In this paper we proposed an approach for predicting the protein network from multiple genomic data using a supervised learning approach. The resulting algorithm borrows ideas from the theory of spectral clustering, and involves the kernel CCA algorithm as a pre-processing step. Cross-validated experiments show that this method predicts the protein network more accurately than several other competing techniques. The predicted pathway network of all proteins enables us to make new biological inferences for unknown protein-protein interactions.

This method is a *supervised* approach, while most methods which have been proposed so far are *unsupervised*. The motivation to use a supervised approach is to explicitly learn the correlation between known networks and genomic data in the algorithm. It should be pointed out that in this supervised framework, different networks can be inferred from the same data, by changing the partial network used in the learning step. Another strength of this method is the possibility to naturally integrate heterogeneous data. Experimental results confirmed that this integration is beneficial for the prediction accuracy of the method. Moreover, other sorts of genomic data can be integrated, as long as kernels can be derived from them. As the list of kernels for genomic data keeps increasing fast (Schölkopf *et al.*, 2004), new opportunities might be worth investigating.

A drawback of our method is that in its current form, it is limited to the prediction of undirected interactions between proteins, which might be insufficient for example in the case of gene regulatory networks. The incorporation of directional information is a topic we are currently

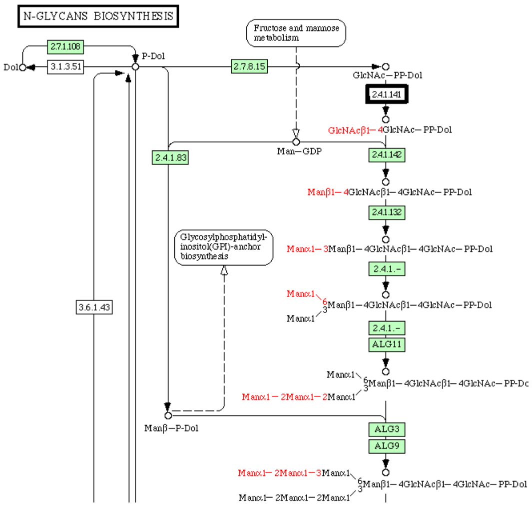

Fig. 7. N-Glycan biosynthesis pathway: EC:2.4.1.141 is a missing enzyme.

investigating, which we expect to bring about more biologically interesting findings.

ACKNOWLEDGMENTS

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, the Japan Science and Technology Corporation, and by a French-Japanese Sakura grant. The computational resource was provided by the Bioinformatics Center, Institute for Chemical Research, Kyoto University.

REFERENCES

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

Akutsu,T., Miyano,S. and Kuhara,S. (2000) Algorithms for identifying Boolean networks and related biological networks based on matrix multiplication and fingerprint function. *J. Comput. Biol.*, **7(3-4)**, 331–343.

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

Bengio,Y., Vincent,P., Paiement,J.-F., Delalleau,O., Ouimet,M. and Le Roux,N. (2003) Spectral clustering and kernel PCA are learning eigenfunctions. *Technical Report 1239, Département d'informatique et recherche opérationnelle, Université de Montreal*.

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.

Friedman,N., Linial,M., Nachman,I. and Pe'er,D. (2000) Using Bayesian networks to analyze expression data. *J. Comput. Biol.*, **7(3-4)**, 601–620.

Huh,W.K., Falvo,J.V., Gerke,C., Carroll,A.S., Howson,R.W., Weissman,J.S. and O'Shea,E.K. (2003) Global analysis of protein localization in budding yeast. *Nature*, **425**, 686–691.

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.

Kanehisa,M., Goto,S., Kawashima,S., Okuno,Y. and Hattori,M. (2004) The KEGG resources for deciphering the genome. *Nucleic Acids Res.*, **32**, D277–D280.

Kondor,R.I. and Lafferty,J. (2002) Diffusion kernels on graphs and other discrete input. *Proc. Int. Conf. Machine Learning (ICML 2002)*, 315–322.

Marcotte,E.M., Pellegrini,M., Thompson,M.J., Yeates,T.O. and Eisenberg,D. (1999) A combined algorithm for genome-wide prediction of protein function. *Nature*, **402**, 83–86.

Ng,A.Y., Jordan,M.I. and Weiss,Y. (2002) On Spectral Clustering: Analysis and an algorithm. *Advances in Neural Information Processing Systems*, **14**.

Pavlidis,P., Weston,J., Cai,J. and Grundy,W.N. (2001) Gene functional classification from heterogeneous data. *RECOMB 2001*, 249–255.

Pazos,F. and Valencia,A. (2001) Similarity of phylogenetic trees as indicator of protein-protein interaction. *Protein Engineering*, **14**, 609–614.

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.

Schölkopf,B. and Smola,A.J. (2002) *Learning with Kernels*. MIT Press, Cambridge, MA.

Schölkopf,B., Smola,A.J. and Müller,K.-R. (1998) Nonlinear component analysis as a kernel eigenvalue problem.. *Neural Computation*, **10**, 1299–1319.

Schölkopf,B., Tsuda,K. and Vert,J.-P. (2004) *Kernel methods in computational biology*. MIT Press, Cambridge, MA.

Spellman,P.T., Sherlock,G., Zhang,M.Q., Iyer,V.R., Anders,K., Eisen,M.B. and et al., (1998) Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol. Biol. Cell.*, **9(12)**, 3273–3297.

Uetz,P., Giot,L., Cagney,G., Mansfield,T.A., Judson,R.S., Knight,J.R., Lockshon,D., Narayan,V. and et al., (2000) A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature*, **10;403(6770)**, 601–603.

Vert,J.-P. and Kanehisa,M. (2003a) Graph-driven features extraction from microarray data using diffusion kernels and kernel CCA. *Advances in Neural Information Processing Systems*, **15**, 1425–1432.

Vert,J.-P. and Kanehisa,M. (2003b) Extracting active pathways from gene expression data. *Bioinformatics (in ECCB 2003)*, **19**, 238ii–244ii.

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 (in ISMB 2003)*, **19**, i323–i330.