BIOINFORMATICS

Vol. 21 Suppl. 1 2005, pages i468–i477
doi:10.1093/bioinformatics/bti1012

Supervised enzyme network inference from the integration of genomic data and chemical information

Yoshihiro Yamanishi[1,], Jean-Philippe Vert[2] and Minoru 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

Received on January 15, 2005; accepted on March 27, 2005

ABSTRACT

Motivation: The metabolic network is an important biological network which relates enzyme proteins and chemical compounds. A large number of metabolic pathways remain unknown nowadays, and many enzymes are missing even in known metabolic pathways. There is, therefore, an incentive to develop methods to reconstruct the unknown parts of the metabolic network and to identify genes coding for missing enzymes.

Results: This paper presents new methods to infer enzyme networks from the integration of multiple genomic data and chemical information, in the framework of supervised graph inference. The originality of the methods is the introduction of chemical compatibility as a constraint for refining the network predicted by the network inference engine. The chemical compatibility between two enzymes is obtained automatically from the information encoded by their Enzyme Commission (EC) numbers. The proposed methods are tested and compared on their ability to infer the enzyme network of the yeast *Saccharomyces cerevisiae* from four datasets for enzymes with assigned EC numbers: gene expression data, protein localization data, phylogenetic profiles and chemical compatibility information. It is shown that the prediction accuracy of the network reconstruction consistently improves owing to the introduction of chemical constraints, the use of a supervised approach and the weighted integration of multiple datasets. Finally, we conduct a comprehensive prediction of a global enzyme network consisting of all enzyme candidate proteins of the yeast to obtain new biological findings.

Availability: Softwares are available upon request.

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

1 INTRODUCTION

Most biological functions involve interactions between many proteins in the cell. To uncover systemic functional behaviors of the cell, an important issue in recent computational biology is the prediction of protein network using all available genomic datasets, which are produced from high throughput experimental technologies, such as DNA sequences, gene expression profiles (Eisen *et al*., 1998), protein–protein interactions (Ito *et al*., 2001), protein intracellular localizations (Huh *et al*., 2003) and phylogenetic profiles (Pellegrini *et al*., 1999). Toward this goal, a variety of computational approaches have been proposed so far. For protein interaction network, examples include the joint graph method to detect strong interacting protein pairs by overlapping the graphs representing similarities with respect to multiple sources of genomic information (Marcotte *et al*., 1999) and an extension of the Bayesian network approach to deal with multiple genomic data (Jansen *et al*., 2003).

Metabolic networks are an important class of protein networks and a typical biochemical network which consists of enzymes and chemical compounds. If we focus on the relationships between enzymes, the metabolic network can be regarded as a graph with enzymes as nodes and enzyme–enzyme relations as edges. By enzyme–enzyme relations, we mean that two enzymes catalyze two successive chemical reactions in the biochemical pathway. In other words, chemical compounds can be considered as edges mediating between enzymes.

Recent developments of pathway databases, such as the KEGG/PATHWAY database (Kanehisa *et al*., 2004), enable us to collect the current knowledge about known metabolic networks. Unfortunately, most of the metabolic networks remain largely unknown, and many enzymes are missing even in known metabolic pathways. Since the experimental determination of metabolic networks remains very challenging nowadays, even for the most basic organisms, there is an incentive to develop methods to infer the unknown parts of the metabolic network and to identify genes coding for missing enzymes in known metabolic pathways.

Almost all previous research on the problem of missing enzyme prediction has focused on the use of genomic data only

*To whom correspondence should be addressed.

(Yamanishi *et al*., 2004) or chemical data only (Goto *et al*., 1996), depending on the choice of the viewpoint to describe the metabolic network: either as a graph of enzymes or as a graph of compounds. However, it is more natural to think of using both genomic and chemical information simultaneously for predicting the enzyme network, rather than using each of them independently.

This paper presents an attempt to infer enzyme networks from the integration of multiple genomic data and chemical information, in the framework of supervised graph inference. The originality of the proposed methods is the integration of both genomic and chemical informations describing enzymes in order to predict more biologically reliable networks. This is made possible by the introduction of chemical compatibility constraints, representing the possibility of successive chemical reactions involving candidate enzymes. These constraints are built from the chemical information encoded in the Enzyme Commission (EC) number assigned to the enzymes. In general, these constraints enable the elimination of incompatible predicted enzyme–enzyme relations from the network predicted by the supervised network inference approach proposed in Yamanishi *et al*. (2004) and Vert and Yamanishi (2005). Moreover, we propose a new and simple weighting scheme based on the estimated relevance of each information source in order to integrate heterogeneous datasets and highlight its positive impact on the inference prediction.

The approach proposed in this paper therefore, only attempts to infer the metabolic networks involving 'genes with known EC number'. The goal is therefore, not to find new enzymes in the genome but to predict in which known or still unknown pathway each known enzyme is involved. This problem is important because it is usually easier to assign EC numbers to genes, using traditional annotation tools based, for instance, on comparative genomics, than to assign them a precise role in a given pathway. As an example, in the well-studied yeast *Saccharomyces cerevisiae* genome, 1120 genes have currently been assigned at least one EC number, but only 668 of them have been assigned at least one precise role in a metabolic pathway.

The proposed methods are tested on their ability to reconstruct the enzyme network of the yeast *S.cerevisiae* from four datasets for enzymes or their genes: gene-expression data obtained from DNA microarrays, protein localization data, sequence data encoded into phylogenetic profiles and chemical compatibility information. The experimental results show that the new methods consistently improve the prediction accuracy over other methods, due to the use of chemical constraint, the use of supervised approach and the weighted integration of multiple datasets. Finally, we conducted a comprehensive prediction of a global enzyme network consisting all possible enzyme candidate proteins (1120 enzymes in this study) of the yeast in order to obtain new biological findings.

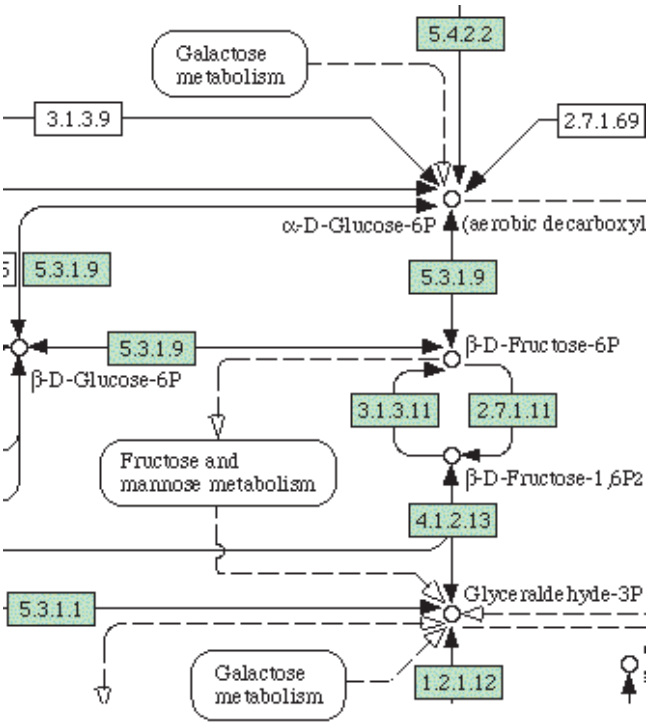

Fig. 1. An example of the metabolic pathway in the KEGG/PATHWAY database. A box indicates an enzyme protein and the number in the box corresponds to the EC number. A circle indicates a chemical compound.

2 MATERIALS AND METHODS

2.1 Gold standard metabolic network

In this study, we focus on the metabolic pathways of the yeast *S.cerevisiae*. As a gold standard for part of the enzyme network, we take the KEGG/PATHWAY database (Kanehisa *et al*., 2004) which is a graph with proteins as vertices, and with edges as enzyme–enzyme relations when two proteins are enzymes that catalyze successive reactions in known pathways. Figure 1 shows an example of the metabolic network in the KEGG. The resulting enzyme network, that contains 668 nodes and 2782 edges, is regarded as a reliable part of the global metabolic network to be inferred below. This network is based on biological phenomena, representing known molecular interaction networks in various cellular process.

2.2 Genomic data

2.2.1 Gene expression data The expression data corresponding to 157 experiments, 77 from (Spellman *et al*., 1998) and 80 from (Eisen *et al*., 1998), are used. Therefore a vector of dimension 157 is associated to each gene coding for an enzyme protein.

2.2.2 Localization data The localization data were obtained from the large-scale budding yeast localization

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

2.2.3 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 orthologuous protein is coded as 1 and 0, respectively, across the 145 organisms.

2.3 Chemical compatibility

We obtained the information about enzyme genes from the KEGG/GENES database, in which EC numbers are assigned to enzyme genes. By the time of writing this paper, the number of genes to which at least one EC number is assigned is 1120.

We obtained the chemical information for the enzymes, such as chemical reactions, substrates and products from their EC numbers, by using the KEGG/LIGAND database (Goto *et al*., 2002), which stores 11 817 compounds and 6349 reactions as of writing this paper. We collected organic chemical compounds that are involved in the enzymes catalyzing chemical reactions. For example, we do not take inorganic compounds (e.g. water, oxygen and phosphate) into consideration, because such compounds tend to appear in too many reactions. Following the definition of EC numbers, we focus on the first three digits in the EC number, because the fourth digit in the EC number is just a serial number. If the first three digits in the EC numbers are the same between two enzymes, we merge all compounds involved with the EC numbers into a list of compounds. If two enzymes share at least one compound across their compound lists, there is a possibility that the two enzymes catalyze successive chemical reactions in metabolic networks. We refer to this property as chemical compatibility in this study.

It should be pointed out that keeping only the first three digits of EC numbers is also a protection against annotation errors resulting from homology-based EC number annotation. Indeed, the prediction of the chemical reaction type (corresponding to the first three numbers) has a higher accuracy than the prediction of the fourth number, which anyway is not used in our approach.

We regard the chemical compatibility between two enzymes as a possible enzyme–enzyme relation, or successive chemical reactions in metabolic networks. Using the chemical compatibilities among all enzymes, we constructed a graph with enzymes as nodes and chemical compatibilities as edges. The numbers of nodes and edges of the resulting graph are 1120 and 404 853, respectively. Obviously, most enzyme–enzyme relations in this network are not likely to correspond to biologically meaningful enzyme–enzyme relations (or biological phenomena). However, this simple constraint already enables to disregard roughly one-third of the 627 760 possible edges between 1120 nodes. We refer to this network as the chemical compatibility network.

3 DATA REPRESENTATION AND INTEGRATION BY KERNEL

3.1 Kernel representation

To deal with the heterogeneity of the datasets, we propose to transform all the datasets into kernel similarity matrices (Schölkop *et al*., 2004). This operation enables us to work in a unified mathematical framework across different types of datasets. All the kernel matrices are supposed to be normalized such that the diagonal elements are all ones and centered in the feature space (Schölkop and Smola, 2002).

The gold standard metabolic network consists of a graph with enzymes as nodes and compounds as edges, so a natural candidate is the diffusion kernel (Kondor and Lafferty, 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). In this study, the diffusion kernel with parameter $\beta = 1$ is applied to the gold standard network data, and the resulting kernel is denoted as K_{gold}.

The expression data, localization data and phylogenetic profiles are sets of numerical vectors, so the Gaussian RBF kernel $k(\mathbf{x}, \mathbf{y}) = \exp(-\|\mathbf{x} - \mathbf{y}\|^2/2\sigma^2)$ or the linear kernel $k(\mathbf{x}, \mathbf{y}) = \mathbf{x} \cdot \mathbf{y}$ are natural candidates. In this study, the Gaussian RBF kernels with parameter $\sigma = 8$ and $\sigma = 3$ are applied to the expression data and phylogenetic profiles, and the resulting kernels are denoted as K_{exp} and K_{phy}, respectively. The linear kernel is applied to the localization data, and the resulting kernel is denoted as K_{loc}.

Since the data structure of the chemical compatibility network is a graph, a candidate of the kernel is the diffusion kernel (Kondor and Lafferty, 2002). In this study, the diffusion kernel with parameter $\beta = 0.01$ is applied to the chemical compatibility network, and the resulting kernel is denoted as K_{che}.

3.2 Data integration

Let us suppose that we use $P \geq 1$ sorts of heterogeneous data (genomic and chemical data) as predictors to infer the enzyme network, and that they are represented by P kernels $K_1, \ldots, K_P$. The function K_p measures the similarity of enzymes with respect to the p-th dataset. A simple data integration is obtained by creating a new kernel as the sum of

the kernels as $K = \sum_{p=1}^{P} K_p$. The usefulness of this procedure has already been proved (Yamanishi *et al.*, 2003, 2004). In this paper, we go further in this strategy by considering weighted sums of kernels, of the form $K = \sum_{p=1}^{P} w_p K_p$, where w_p represents the weight associated to the p-th dataset for predicting metabolic networks. Intuitively, the weight of a dataset should be related to the relevance of the dataset for predicting metabolic networks.

Therefore, the essential problem is how to determine the weight w_p in the integration process. In this study, we propose to take the weights $(w_1, \ldots, w_p)$ proportional to an estimation of prediction accuracy of the corresponding dataset [e.g. receiver operating curve (ROC) scores -0.5], obtained from experiments on each individual datasets. While more complex algorithm can be imagined to automatically determine the weight in the integration of heterogeneous data through kernel operation (e.g. convex optimization), this problem is out of the scope of the paper and is not considered here owing to space limitation.

4 METHODS FOR NETWORK INFERENCE

We attempt to infer enzyme networks using either individual kernels or integrated kernels representing the above genomic and chemical datasets.

4.1 Direct approach for network inference

The most direct approach to network reconstruction is a similarity-based approach, assuming that functionally related enzyme pairs are likely to share high similarity with respect to a given dataset. Intuitively, the kernel value $k(\mathbf{x}, \mathbf{y})$ can often be considered as a measure of similarity between enzyme $\mathbf{x}$ and enzyme $\mathbf{y}$. A direct strategy is therefore to predict an edge between two enzymes whenever the kernel value between these enzymes is above a threshold to be determined. We refer to this approach as direct approach. By varying the threshold we can obtain different rates of true positives and true negatives.

4.2 Supervised approach for network inference

We tested two recently proposed algorithms to perform a supervised inference of the metabolic network (Yamanishi *et al.*, 2004; Vert and Yamanishi, 2005). As opposed to the direct approach described in Section 4.1, these methods require a partial knowledge of the metabolic network, in order to infer unknown parts.

As illustrated in Figures 2 and 3 both methods involve a training process, where a mapping of all genes to a low-dimensional space is learned by exploiting the partial knowledge of the network, and a test process where new edges are inferred. The test process is simply the direct approach performed after genes are mapped to the low-dimensional Euclidean space, i.e. pairs of enzymes with short interdistance are connected. The learning process must therefore produce a mapping after which the similarity of genes, in terms of

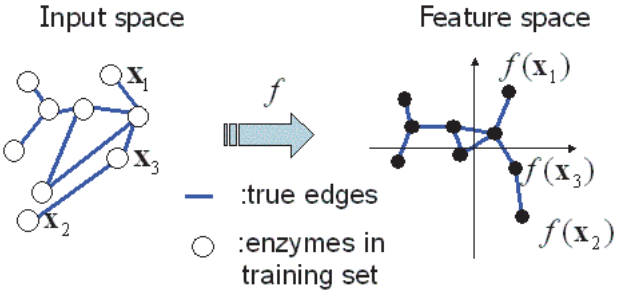

Fig. 2. Supervised network inference: training process. Enzymes in the training set are mapped onto a feature space, where interacting enzymes are close to each other. The projection f is learned by kernel canonical correlation analysis (KCCA).

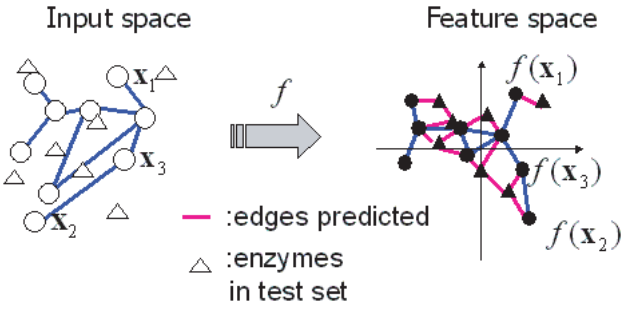

Fig. 3. Supervised network inference: test process. Enzymes in the test set are mapped onto the feature space constructed by KCCA. Then, interacting enzyme pairs are predicted using the direct approach.

Euclidean distance, is a good indicator of the presence or absence of edges.

Both methods differ slightly in the training phase. In Yamanishi *et al.* (2004), the genes are mapped onto the first few components obtained by kernel canonical correlation analysis (KCCA) (Akaho, 2001) between the data kernel and the kernel describing the known part of the network. In Vert and Yamanishi (2005) the genes are mapped onto components obtained by minimizing a functional that explicitly seeks to move known connected pairs close to each other in the feature space. Both approaches are solved by a generalized eigenvalue problem.

Following the experimental results presented in Yamanishi *et al.* (2004) and Vert and Yamanishi (2005), we set the number of features (dimension of the feature space) to 30, the λ parameter of the method described in Yamanishi *et al.* (2004) to 0.01, and the λ parameter of the method described in Vert and Yamanishi (2005) to 2. Although the influence of these parameters could be studied in more detail, we chose in this paper to vary as few parameters as possible in order focus on the influence of the kernel used and the introduction of chemical constraints.

Table 1. List of experiments for the direct approach, the direct approach with chemical compatibility, the supervised approaches and the supervised approaches with chemical compatibility

Approach	Data	Kernel	Chemical constraint
Direct	Expression	K_{exp}	
Direct	Localization	K_{loc}	
Direct	Phylogenetic profile	K_{phy}	
Direct	Chemical compatibility	K_{che}	
Direct	Integration	$K_{\text{exp}} + K_{\text{loc}} + K_{\text{phy}}$	
Direct	Integration + Chemical compatibility	$K_{\text{exp}} + K_{\text{loc}} + K_{\text{phy}} + K_{\text{che}}$	
Direct	Weighted Integration	$w_1 K_{\text{exp}} + w_2 K_{\text{loc}} + w_3 K_{\text{phy}} + w_4 K_{\text{che}}$	
Direct	Expression	K_{exp}	Post-integration
Direct	Localization	K_{loc}	Post-integration
Direct	Phylogenetic profile	K_{phy}	Post-integration
Direct	Integration	$K_{\text{exp}} + K_{\text{loc}} + K_{\text{phy}}$	Post-integration
Direct	Weighted Integration	$w_1 K_{\text{exp}} + w_2 K_{\text{loc}} + w_3 K_{\text{phy}}$	Post-integration
Supervised	Expression	K_{exp}	
Supervised	Localization	K_{loc}	
Supervised	Phylogenetic profile	K_{phy}	
Supervised	Chemical compatibility	K_{che}	
Supervised	Integration	$K_{\text{exp}}+K_{\text{loc}}+K_{\text{phy}}$	
Supervised	Integration + Chemical compatibility	$K_{\text{exp}} + K_{\text{loc}} + K_{\text{phy}} + K_{\text{che}}$	
Supervised	Weighted Integration	$w_1 K_{\text{exp}} + w_2 K_{\text{loc}} + w_3 K_{\text{phy}} + w_4 K_{\text{che}}$	
Supervised	Expression	K_{exp}	Post-integration
Supervised	Localization	K_{loc}	Post-integration
Supervised	Phylogenetic profile	K_{phy}	Post-integration
Supervised	Integration	$K_{\text{exp}} + K_{\text{loc}} + K_{\text{phy}}$	Post-integration
Supervised	Weighted Integration	$w_1 K_{\text{exp}} + w_2 K_{\text{loc}} + w_3 K_{\text{phy}}$	Post-integration

4.3 Incorporating chemical constraint

For each network inference strategy, we tested two approaches to integrate the chemical information contained in the chemical compatibility network, which we refer to as pre-integration and post-integration.

The pre-integration strategy consists in considering the chemical compatibility network as an additional source of information about the enzymes, encode it into a kernel similarity measure (as explained in Section 3.1) and use it as an additional component when kernels are integrated through sum or weighted convex combination (Section 3.2). The rationale behind this approach is to try to enforce some chemical constraint in the kernel itself, without strictly enforcing all constraints. Indeed, the absence of an edge between two genes in the chemical compatibility network can be due to the absence of an EC number assignment, in which case, a lack of edge in the chemical compatibility should not be strictly enforced as strong evidences for the presence of an edge stem from other sources of data.

To the contrary, the post-integration strategy gives a dominant role to the chemical constraints by strictly enforcing them. It consists in first performing the normal graph inference without chemical information, followed by the selection among predicted edges of those that fulfill the chemical constrain, namely the ones that are present in the chemical compatibility network. With this method, all edges of the resulting graph necessarily fulfill the chemical constraints.

5 RESULTS

We performed a series of experiments to test the performance of the methods and their differences on the problem of reconstructing the gold standard metabolic network. The evaluation of each method relies on the area under the ROC (Gribskov and Robinson, 1996), i.e. the area under the plot of true positives as a function of false positives, normalized to 1 for a perfect inference. For each method, the ROC curve is obtained by varying the threshold below which edges are predicted. A true positive is a correctly predicted edge, a false positive is a predicted edge that is not present in the gold standard metabolic network.

The curves can be computed for the direct approach by just inferring the global gold standard network. The supervised methods require the knowledge of part of the network in the training process. We therefore evaluated them with the following 10-fold cross-validation procedure: the set of all nodes is split into 10 subsets of roughly equal size, each subset is taken apart in turn to perform the training with 90% of the nodes, and the prediction concerns the edges that involve the nodes in the subset taken apart during training.

Table 1 summarizes the experiments performed. For each unsupervised and supervised method, we tested both the

Table 2. Result summary: area under the ROC curves for different methods

Approach	Direct	Direct	Supervised (KCCA)	Supervised (KCCA)	Supervised (DML)	Supervised (DML)
Chemical post-processing	No	Yes	No	Yes	No	Yes
Expression	0.502	0.571	0.639	0.688	0.706	0.741
Localization	0.561	0.624	0.567	0.626	0.577	0.640
Phylogenetic profile	0.567	0.629	0.747	0.779	0.707	0.764
Chemical compatibility	0.539	—	0.750	—	0.592	—
Integration	0.574	0.628	0.804	0.829	0.778	0.807
Integration with chem.	0.586	—	0.801	—	0.770	—
Weighted Integration	0.595	0.642	0.829	0.836	0.777	0.817

Two supervised graph inference methods are tested: the approach based on KCCA described in Yamanishi *et al.* (2004), and the methods based on distance metric learning described (DML) described in Vert and Yamanishi (2005).

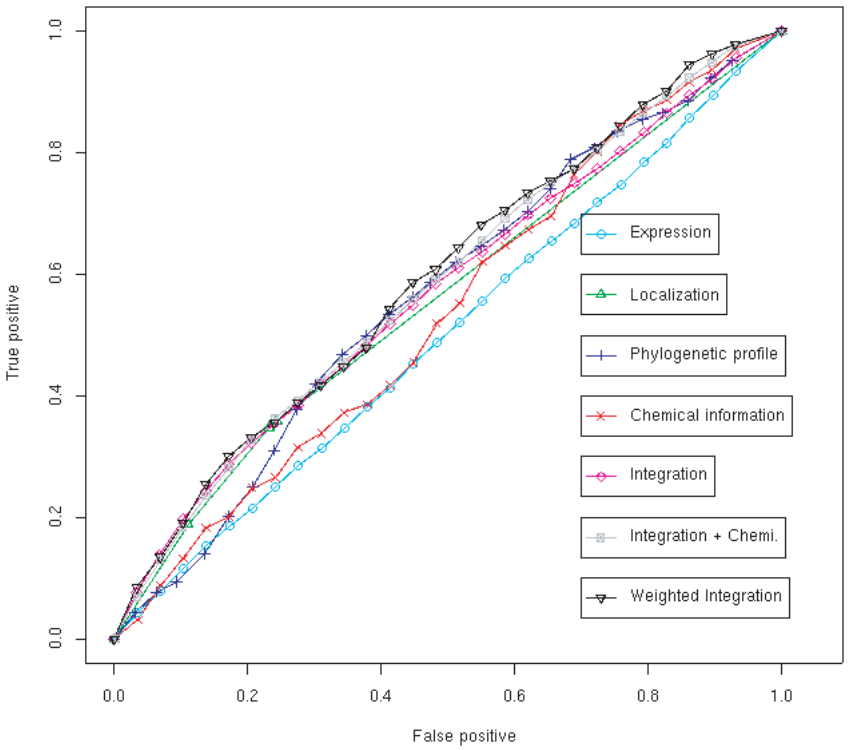

Fig. 4. ROC curves: direct approach.

pre-integration and the post-integration strategies to take into account the chemical constraints. Results are reported in Table 2.

Figures 4 and 5 show the ROC curves for the direct approach and direct approach with chemical constraint, respectively. The ROC scores (area under the ROC curves) for the two methods are computed and summarized in the columns 2 and 3 in Table 2. Both direct methods seem to catch little information to recover the metabolic network from the datasets. The use of the constraint of chemical compatibility seems to have effects of refining the network predicted by the direct approaches in all the cases. However, these methods are impractical in actual applications because of their high false positive rate against true positive rate at any threshold.

Next, we tested both supervised approaches with the pre-integration and the post-integration strategy for chemical constraints. We denote by 'supervised-KCCA' the method based on KCCA (Yamanishi *et al.*, 2004), and by 'supervised-MDL' the method based on metric distance learning (Vert and Yamanishi, 2005). While both methods give slightly different

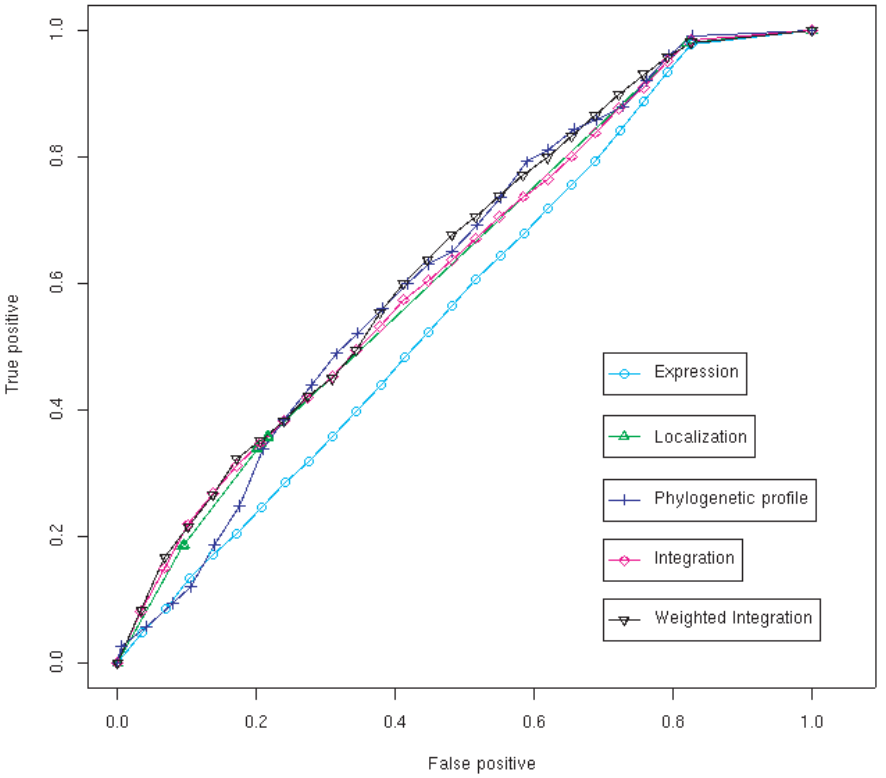

Fig. 5. ROC curves: direct approach with chemical constraint.

results, their overall behavior is similar. Figures 6 and 7 show the ROC curves for the supervised-KCCA approaches, with a pre-integration and a post-integration strategy, respectively. The ROC scores (area under the ROC curves) for the two methods and two strategies are reported in the columns 4–7 of Table 2. We observe that the shift from unsupervised to supervised learning significantly improves the prediction accuracy in all cases. We observe slight differences between the two supervised methods when individual kernels are used. The phylogenetic profile kernel significantly outperforms both the expression and the localization kernel with the supervised-KCCA method, while it is roughly at the level of the expression kernel and above the localization kernel with the supervised-MDL method. More importantly, we observe that the pre-integration of chemical constraint has little influence on the accuracy, while the post-integration strategy consistently improves the ROC score in all cases. Finally, we observe a significant improvement when kernels are combined with weights.

The overall best result (83.6%) is obtained by the supervised-KCCA approach in conjunction with a weighted integration of all genomic datasets, combined with the chemical constraints using the post-integration strategy. The comparison of these experimental results highlights the accuracy improvements resulting from the use of supervised approach, the weighted integration of multiple datasets and the use of the chemical constraint.

Since we confirmed the validity of the supervised method by the cross-validation experiments, we finally conducted a comprehensive prediction of a global network for all enzyme candidate proteins (1120 enzymes in this study) of the yeast. The predicted network enabled us not only to make new biological inferences about unknown enzyme–enzyme relations but also to identify genes coding for missing enzymes in known metabolic pathways. We take the protein YJR137C as a target protein, for example. The detailed function of this protein was not clear as of starting this work, although the first two digits of EC number was known as EC:1.8.-,-. In the predicted network, this protein is connected to the enzyme proteins YPR167C (EC:1.8.4.8) and YGR012W (EC:2.5.1.47), from which we can guess that the target protein might be functionally related to these enzymes. Recently, there has been a report that this protein is annotated as EC:1.8.1.2 according to the MIPS database, where EC:1.8.1.2 is known to have successive reactions with EC:1.8.4.8 and EC:2.5.1.47, for example, according to the sulfur metabolism in the KEGG/PATHWAY database. Of course, such inference can be applied to other enzyme proteins. The results of the predicted enzyme network can be obtained from the author's website (http://web.kuicr.kyoto-u.ac.jp/~yoshi/ismb05/).

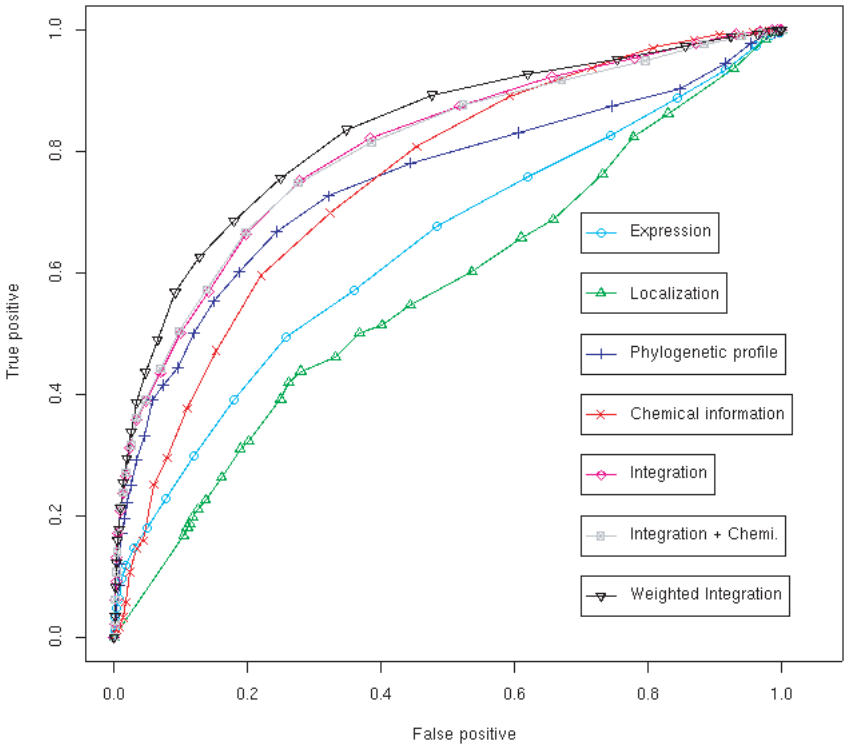

Fig. 6. ROC curves: supervised approach based on KCCA.

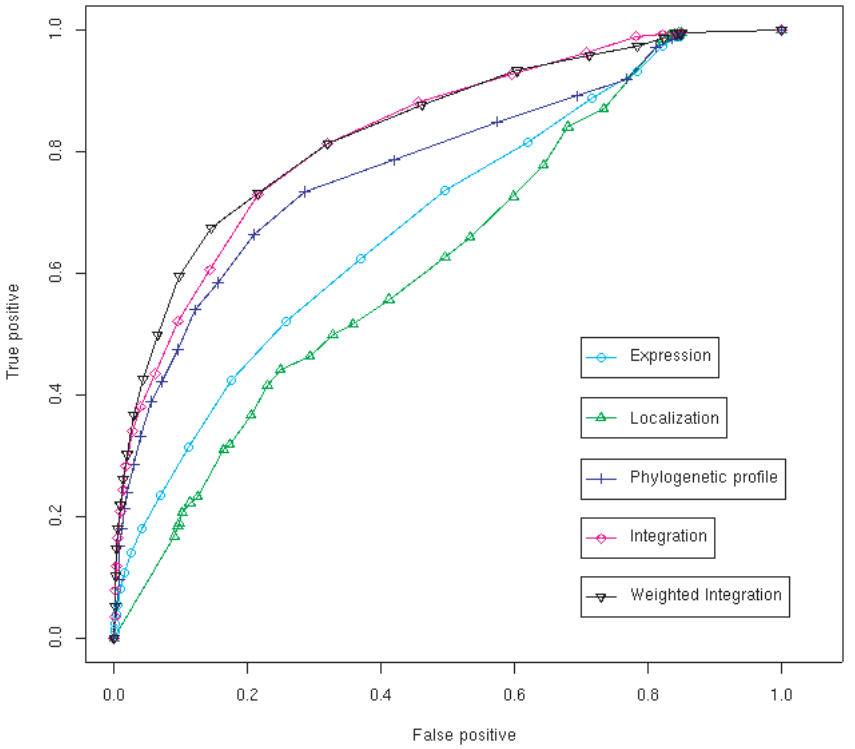

Fig. 7. ROC curves: supervised approach with chemical constraint based on KCCA.

6 DISCUSSION

In this paper, we proposed two strategies to infer enzyme networks from the integration of multiple genomic data and chemical information in the framework of supervised graph inference. One strategy (pre-integration) amounts to considering the chemical constraints as another genomic dataset, while the second strategy operates as a filter on the predicted edges to strictly enforce chemical constraints. The originality of our method is in the integration of heterogeneous datasets representing enzymes in genomes (sequence data encoded into phylogenetic profiles), transcriptome (gene expression), proteome (protein localization) and metabolome (chemical compounds involved), using a simple weighting scheme. The cross-validation experiments showed that the methods, in particular the post-integration strategy combined with supervised learning on weighted linear combinations of kernels, improved the prediction accuracy to a large extent. Finally, we predicted a global enzyme network for obtaining new biological findings.

The biological motivation of using chemical information is related to the path computation problem (Goto *et al*., 1996), which is the problem of computing possible paths between compounds to resolve the missing enzyme issue in metabolic pathways. When enzymes supposed to catalyze the reaction between two compounds are missing in known pathways, the path computation method enables the search of possible paths between the two compounds based on the network with the compounds as nodes and the enzymes as edges. However, it has been pointed out (Goto *et al*., 1996) that the system tends to show too many path candidates between the compounds in the path computation method, as we have observed in our chemical compatibility network.

In this study, we focused on the first three digits in the EC numbers assigned to the enzyme candidate genes, and we created a chemical compatibility network to look for biochemically possible edges. However, this process tends to produce too many possible enzyme–enzyme relations, similar to the path computation method, so we might need to develop a method to reduce the candidates of possible enzyme–enzyme relations. One possibility is to eliminate cofactors (coenzymes), such as $NAD(P)^+$ and ATP, because they tend to appear in too many reactions, although the definition of cofactor is a very difficult problem.

From the viewpoint of algorithms, our method first starts by transforming each dataset into a kernel matrix whose elements represent similarities between proteins. This process enables us to deal with different data types elegantly and in a unified framework. However, the performance of kernel methods often depends on the definition of the kernel function and its parameters, as well as regularization parameters in the supervised network inference. The performance of our method might therefore be improved by using more specific kernel functions and more appropriate parameters in actual application. Therefore, the development of appropriate parameter optimization methods is one of our future works. As a straightforward approach, one way to optimize the parameters should be, for example, to apply cross-validation with the ROC area as a target criterion.

Future work includes the experimental validation of the predicted metabolic networks, both in yeast and in other organisms when sufficient data are available. In case of bacterial genome, because many sets of proteins involved in the same biological process are encoded into operons in the genome, an additional dataset to be considered is certainly the position of the genes on the chromosome, which can also be encoded into a kernel (Yamanishi *et al*., 2003).

ACKNOWLEDGEMENTS

Y.Y. and M.K. are 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. J.P.V. acknowledges the support of NIH grant R33HG003070-01. The computational resource was provided by the Bioinformatics Center, Institute for Chemical Research, Kyoto University. This collaboration was also supported by the French–Japanese Sakura grant.

REFERENCES

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

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*, **2**, 175–186.

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

Gribskov,M. and Robinson,N.L. (1996) Use of receiver operating characteristic (ROC) analysis to evaluate sequence matching. *Comput. Chem.*, **20**, 25–33.

Huh,W.K., Falvo,J.V., Gerke,L.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 interactome. *Proc. Natl Acad. Sci. USA*, **98**, 4569–4574.

Jansen,R., Yu,H., Greenbaum,D., Kluger,Y., Krogan,N.J., Chung,S., Emili,A., Snyder,M., Greenblatt,J.F. and Gerstein,M. (2003) A Bayesian networks approach for predicting protein–protein interactions from genomic data. *Science*, **302**, 449–453.

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. In Proceedings of the 19th International Conference on Machine Learning, Morgan Kaufmann, University of South Wales, Sydney, Australia, pp. 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.

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.

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

Schölkop,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., 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.

Vert,J.-P. and Yamanishi,Y. (2005) Supervised graph inference. In Saul,L.K., Weiss,Y. and Bottou,L. (eds) *Proceeding of the Conference on Advances in Neural Information and Processing System.* MIT Press, Cambridge, MA, pp. 1433–1440.

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*, **19**, i323–i330.

Yamanishi,Y., Vert,J.-P. and Kanehisa,M. (2004) Protein network inference from multiple genomic data: a supervised approach. *Bioinformatics*, **20**, i363–i370.