

Selective Integration of Multiple Genomic Data for Biological Network Inference

Tsuyoshi Kato¹

Koji Tsuda^{1,2}

Kiyoshi Asai^{3,1}

kato-tsuyoshi@aist.go.jp

koji.tsuda@tuebingen.mpg.de

asai-cbrc@aist.go.jp

¹ AIST Computational Biology Research Center,
2-43, Aomi, Koto-ku, 135-0064 Tokyo, Japan

² Max Planck Institute for Biological Cybernetics, 72076 Tübingen, Germany

³ Graduate School of Frontier Sciences, The University of Tokyo,
5-1-5, Kashiwanoha, Kashiwa city, 277-8562, Japan

Keywords: supervised network inference, information geometry, multiple data integration

1 Introduction

In the field of computational biology, recently there has been a surge of interest in biological networks such as protein interaction networks, gene regulatory networks, or metabolic networks, which help us to understand the cellular machinery. Most of biological networks represent the relationships between genes or proteins. Namely the existence of edges means that the corresponding genes/proteins are related each other. In this paper, we are confronted with a problem to predict the interactions between genes/proteins. In our problem setting we are supposed to be given a part of the network and various genomic data such as gene expression data, localization data, and phylogenetic profile data. Our task is to predict the rest of the network. Yamanishi et al. [3] have already tackled the exactly same setting, and refer this problem to supervised network inference. Their approach uses canonical correlation analysis following the process of integration of multiple genomic data. The integration is done by linear combination of the multiple kernel matrices. We have developed another new method for this problem. Our method is based on information geometrical theory, and simultaneously infers the unknown subgraph and the weights of linear combination.

2 Method

Let us formulate the supervised network inference problem for ℓ proteins. We assume that the network is known for the first n proteins. We wish to predict the edges between the n proteins and the remaining $m := \ell - n$ proteins, and those among the remaining proteins. To this aim, we exploit n_K different kinds of data available for all ℓ proteins. Let us denote each of the correlation matrix by $P_k \in \mathfrak{R}^{\ell \times \ell}$. We combine them with weights $\mathbf{b} = \{b_1, \dots, b_{n_K}\}$ as $P(\mathbf{b}) = \sum_{k=1}^{n_K} b_k P_k + \sigma^2 I$, where $\sigma^2 I$ is a regularization term. To pose the network inference problem as statistical inference, the known part of the network is converted to an $n \times n$ correlation matrix K_I , for example, using diffusion kernels [1]. This matrix is made such that two proteins close to each other in the network have a high value. Thresholding the entries of K_I recovers the given network approximately, but we do not need that the known network is exactly recovered.

To predict edges, we estimate the rest of the network correlation matrix, say Q_{vh} and Q_{hh} , in $\ell \times \ell$ full network correlation matrix: $Q = \begin{bmatrix} K_I & Q_{vh} \\ Q_{vh}^\top & Q_{hh} \end{bmatrix}$, and threshold it. We determine the values of Q_{vh} and Q_{hh} and the weights \mathbf{b} which minimizes an distance between Q and $P(\mathbf{b})$ based on information geometry. The distance is defined by the Kullback-Leibler divergence between two zero-mean Gaussians with covariances, Q and $P(\mathbf{b})$, respectively.

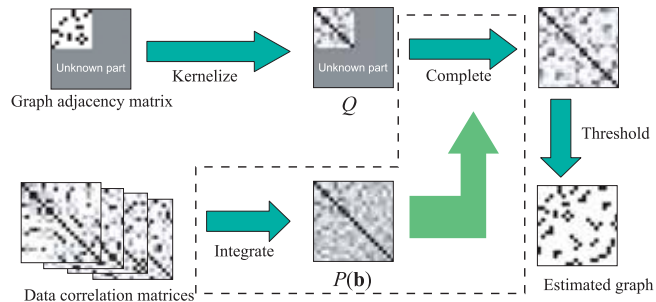


Figure 1: Flow of our method. First, the known part of adjacency matrix is kernelized. The rest of the resulting kernel matrix is inferred from the optimally integrated correlation matrix $P(\mathbf{b})$. Finally, we obtain the unknown part of adjacency matrix by thresholding at some value.

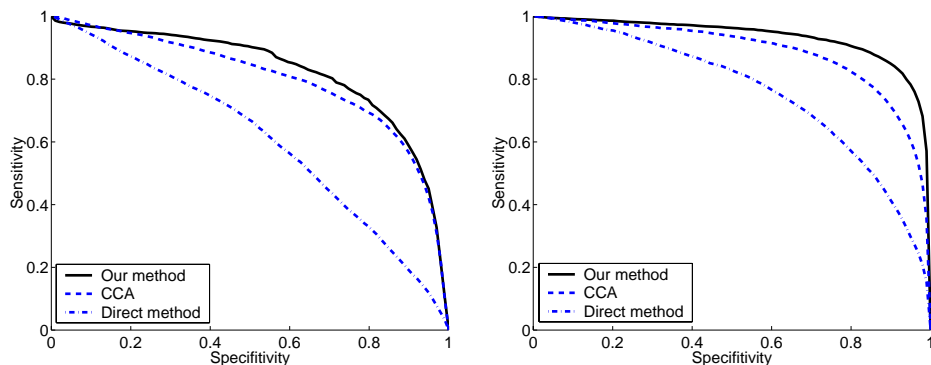


Figure 2: The ROC curves of prediction for a metabolic network (left) and a protein interaction network (right).

3 Experimental Results

To compare our method with CCA, we use the same data set as Yamanishi et al.’s report [3]. Those data set contains the following four different data: gene expression (‘exp’), yeast two-hybrid data (‘y2h’), protein localization (‘loc’), and phylogenetic profiles (‘phy’), which are readily converted to correlation matrices. Using them, we have carried out the prediction experiments for a metabolic network produced from KEGG/PATHWAY database. Additionally, we tested on a protein interaction network created by von Mering et al [2]. The data ‘y2h’ are also protein interactions, but the data is extremely noisy. Our method is compared with CCA as well as the *direct method* which pick up the entries of the integrated correlation matrix above some threshold. We have obtained encouraging results, as shown in Figure 2.

References

- [1] R. I. Kondor and J. Lafferty. Diffusion kernels on graphs and other discrete structures. In C. Sammut and A. G. Hoffmann, editors, *Machine Learning, Proceedings of the 19th International Conference (ICML 2002)*, pages 315–322. San Francisco, Morgan Kaufmann, 2002.
- [2] C. von Mering et al. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, 417:399–403, 2002.
- [3] Y. Yamanishi, J.P. Vert, and M. Kanehisa. Protein network inference from multiple genomic data: a supervised approach. *Bioinformatics*, 20(Suppl. 1):i363–i370, 2004.