

Inferring Protein Interaction Network by Boosting Algorithm

Yong Wang

Feng Bao

Jiadong Zhang

Luonan Chen

chen@elec.osaka-sandai.ac.jp

Osaka Sangyo University, Nakagaito 3-1-1, Daito, Osaka 574-8530, Japan

Keywords: protein interaction, protein network, boosting algorithm

1 Introduction

One of major goals of functional genomics is to elucidate protein interaction networks for whole organisms. Determining protein interactions provides not only detailed functional insights on characterized proteins, but also an information base for identifying biological complexes and metabolic or signal transduction pathways [1]. The recent emergence of high-throughput proteomics techniques has opened new prospects to systematically characterize physical interactions between proteins. Based on experimental dataset, many computational algorithms have been developed to infer the protein-protein or domain-domain interactions. For instance, for inferring protein interactions, there are the gene fusion (Rosetta Stone) method, the phylogenetic profile method, the interaction domain pair profile method, the probabilistic method, the SVM-based method, and the LP-based approach, whereas for inferring domain interactions, there are the association method, the EM algorithm. Despite the relative success, there is much room for improvement of protein interaction inference in terms of prediction quality and computational efficiency. Based on the association method, we propose an association probabilistic method (APM) to infer protein interactions directly from the experimental data, and then further improve the accuracy of APM [1] by adopting boosting algorithm. By the numerical simulation, we show that the proposed method achieves the highest accuracy among the existing approaches for the measures of root mean square error and the Pearson correlation coefficient with the efficiency.

2 Method and Results

In this work, we propose a new algorithm to improve the quality of protein-protein interaction inference by using boosting algorithm. Boosting is a method of finding a highly accurate rule (or hypothesis) by combining many weak rules. The basic idea is to repeatedly apply a simple learning algorithm or weak (base) learner, to different weightings of the same training set. In this paper, we take APM as a base learner in the boosting algorithm due to its simple computation procedure, and achieve an accurate and efficient prediction of protein-protein interactions from the experiment data by proposing a boosting learning algorithm. It can be proven that if the rules generated in the iterations are all slightly correlated with the label or confident ratio, then the strong rule will have a very high correlation with the label or confident ratio. In other words, the strong rule can predict the protein interactions very accurately. Although we introduce the distribution for training set, it is not necessary to generate samples from the training set according to the

algorithm because the ratio can be analytically expressed by samples. In addition to a high accuracy, such a feature implies that the computation may also be very efficient in boosting learning. In addition, both binary and confident ratio experimental data of protein interactions can be considered in the proposed algorithm, which significantly extends the available data set thereby improving the accuracy of the interaction prediction.

References

- [1] Chen, L., We, L.Y., Wang, Y., and Zhang, X.S., Inferring protein interactions from experimental data by association probabilistic method, *PROTEINS: Structure, Function, and Bioinformatics*. 2005.