A Transfer Learning Approach and Selective Integration of Multiple Types of Assays for Biological Network Inference

Tsuyoshi Kato, AIST Computational Biology Research Center, and Ochanomizu University, Japan
Kinya Okada, KO Institute for Medical Bioinformatics, Japan
Hisashi Kashima, IBM Research, Japan
Masashi Sugiyama, Tokyo Institute of Technology, Japan

ABSTRACT

Inferring the relationship among proteins is a central issue of computational biology and a diversity of biological assays are utilized to predict the relationship. However, as experiments are usually expensive to perform, automatic data selection is employed to reduce the data collection cost. Although data useful for link prediction are different in each local sub-network, existing methods cannot select different data for different processes. This article presents a new algorithm for inferring biological networks from multiple types of assays. The proposed algorithm is based on transfer learning and can exploit local information effectively. Each assay is automatically weighted through learning and the weights can be adaptively different in each local part. The authors’ algorithm was favorably examined on two kinds of biological networks: a metabolic network and a protein interaction network. A statistical test confirmed that the weight that our algorithm assigned to each assay was meaningful.

Keywords: Local Model, Multiple Kernel Learning, Supervised Network Inference, Support Vector Machine, Transfer Learning

BACKGROUND

A thorough understanding of cellular processes is the central goal of molecular biology. For this purpose, it is required to understand not only individual functions but also relationships among their components. Nowadays, a huge amount of data on molecular relationships can be generated through high throughput assays and analyzed as molecular networks. Although each of the assays contains useful information, molecular networks reconstructed from
a single assay often have too much noise. To cope with this problem, several groups have tried to integrate data obtained from multiple types of assays for reliable network reconstruction (Pavlidis et al., 2002; Kato et al., 2005; Yamanishi et al., 2005).

How can we integrate data from multiple types of assays? A primitive method is to average data from the multiple types of assays and reconstruct a network from the averaged data (Yamanishi et al., 2005; Vert & Yamanishi, 2005). This method in a sense treats all types of assays equivalently, even if they may not be equivalent; that is, the assays may differ from each other in quality and resolution. To take non-equivalence among multiple types of assays into account in data integration, several groups have proposed methods to optimize the weight assigned to each type of assay (Kato et al., 2005; Lanckriet et al., 2004).

Although the existing methods allow each type of assay to be weighted, the weights are global in the entire network. Namely, every edge is predicted using weights common to the whole data. Since the mechanisms underlying cellular processes are complicated and heterogeneous, a type of assay may help shed light on some local cellular processes even if a low weight is assigned to the type of assay. Consequently, due to the heterogeneity of the relationship between assays and cellular mechanisms, the global weighting is too coarse, and finer modeling is desired.

Bleakley et al. (2007) have proposed a method that uses local models to cope with the heterogeneity issue. Their method builds a local model for each node (they call it a target node) and trains it with only its local (i.e., neighboring) information, resulting in scoring functions that are not corrupted by the irrelevant effects of distant parts of its molecular network. One shortcoming of their method is that the amount of local information is often limited, because most of the nodes have a few edges due to the fact that the node connectivity in molecular networks follows a power law distribution (Caldarelli, 2007). Insufficient information for training adversely affects the generalization ability of the method and restricts our ability to automatically acquire the genuine weights of the assays.

We present herein a novel algorithm for edge prediction. Our algorithm is based on Bleakley et al.’s approach, but our algorithm is extended to enjoy two remarkable features. First, in order to address the loss of generalization ability due to the scarcity of local information, we apply transfer learning to building local models. Transfer learning is an approach to machine learning that learns a task together with other related tasks simultaneously. This often leads to a better model for the target task, because it allows the learner to share appropriate information across the tasks. The use of this approach is motivated by the observation that node A is likely to be linked with node B when they share common neighbors in molecular networks (Chua et al., 2006; Samanta & Liang, 2003; Okada et al., 2005). According to this observation, the task of building the local model of a target node is likely to be related to that of its neighboring nodes. Transfer learning builds the local model of a target node with the help of its neighboring nodes. Furthermore, transfer learning optimizes weights assigned to the types of assays when building local models, offering selected assays to each local part of a molecular network. In this study, we reconstruct molecular networks using this method and validate the efficiency of our method.

The rest of this section summarizes abbreviations which are used frequently in this article. SVM is the abbreviation of the support vector machine (Schölkopf & Smola, 2002). RBF is the radial basis function (Schölkopf & Smola, 2002). BBVK, TPPK, and MLPK are the BBV kernel (Bleakley et al., 2007), the tensor product pairwise kernel (Ben-Hur & Noble, 2005), and the metric learning pairwise kernel (Vert et al., 2007), as detailed in the section of Methods. PIF is the protein interaction profile defined in the section of Discussion. GO is the gene ontology. We will show the full name again at the first occurrence except this paragraph.
Hierarchical Density-Based Clustering of White Matter Tracts in the Human Brain
Junming Shao, Klaus Hahn, Qinli Yang, Afra Wohlschläger, Christian Boehm, Nicholas Myers and Claudia Plant (2010). International Journal of Knowledge Discovery in Bioinformatics (pp. 1-25).
www.igi-global.com/article/hierarchical-density-based-clustering-white/49547?camid=4v1a

Clustering Genes Using Heterogeneous Data Sources
Erliang Zeng, Chengyong Yang, Tao Li and Giri Narasimhan (2012). Computational Knowledge Discovery for Bioinformatics Research (pp. 67-83).
www.igi-global.com/chapter/clustering-genes-using-heterogeneous-data/66705?camid=4v1a