Integrative Construction and Analysis of Condition-Specific Biological Networks

Sushmita Roy, Terran Lane
Department of Computer Science, University of New Mexico

Margaret Werner-Washburne
Department of Biology, University of New Mexico

Introduction

Condition-specific networks are networks of genes, proteins and metabolites that are triggered by changes in an organism’s environmental conditions such as nutrient availability, temperature etc. Inference and analysis of these condition-specific networks is critical to understanding how an organism survives in different environmental conditions. Condition-specific networks will provide insight into the mechanisms inducing different parts of the cellular machinery resulting in different diseased or healthy cellular states.

Machine learning algorithms for network structure inference can be broadly classified into two categories. In the first category, the structure of the network is inferred de-novo from the observed expression values of the network nodes (Friedman et al. 2000; Margolin et al. 2005). Although these approaches have been very successful in obtaining global properties of biological networks, they infer only statistical correlation between nodes with no interpretation of the physical interactions that drive an organism’s condition-specific response. The second category, based on undirected graphical models, incorporates detailed information of the physical attributes of the network but assume that the network structure is fixed (Yeang, Ideker, & Jaakkola 2004; Jaimovich et al. 2006). Although these approaches can build a detailed picture of the underlying physical mechanisms, the fixed network structure makes strong assumptions of the statistical dependency structure in the data. Neither of these approaches have accounted for the condition-specific properties of the biological networks.

We propose a multi-stage approach using a combination of classification, network structure and parameter learning, and graph theory algorithms to construct high resolution condition-specific networks. In the first stage, we predict the complete physical network by training classifiers on existing databases of known physical interactions. We use a classifier to predict interactions that are absent from the databases to build as complete a picture of the physical network. In the next stage, we use network structure learning and hidden variable inference algorithms to infer coarse and fine-grained network structures. The algorithms will leverage condition-specific information such as gene, protein and metabolite concentrations and will constrain the network structure using the predicted physical network from the first stage. Finally, we analyze the structure of different high resolution condition-specific networks using graph theoretic concepts.

In this abstract, we describe preliminary work regarding the first two stages of our approach. We describe novel classifier features, amino-acid compositions of proteins, to predict protein interactions. We also describe a novel score, pathwise score, that evaluates how well structure learning algorithms capture higher-order statistical dependencies.

Prediction of protein-protein interactions

Most classification algorithms for predicting protein-protein interactions incorporate protein domain information as features. We propose that amino-acid composition is also an important feature for predicting protein-protein interactions. We trained separate maximum-entropy classifiers on domains and on amino-acid composition of yeast proteins and compared their performance using the area under the ROC curve (AUC). We found that the classifier using amino-acid composition (AUC 0.71 ± 0.02) performed at par and sometimes better than the classifier using domains (AUC 0.66 ± 0.04). We also found that a classifier that combined domains and amino-acid composition had an improved performance (AUC 0.74 ± 0.03), indicating that amino-acid composition can boost performance of classifiers trained on other features. We had similar results with a support vector machine classifier, indicating that our results were not an artifact of a specific classifier. It is surprising and biologically interesting that a simple feature like amino-acid composition works at least as well as well-known but complex features such as protein domains. Preliminary analysis of our results indicates that amino-acid composition may influence structural and charge configurations of proteins that facilitate interactions.

Network learning algorithm comparison

Algorithms for learning biological networks must model higher-order dependencies, which are induced by the modular structure of such networks. Such modules represent biological pathways or protein complexes required for condition-specific behaviour in organisms. However, edge-wise precision and recall (Margolin et al. 2005), commonly...
used for comparing these algorithms, are based on matching individual edges between the true and predicted network. Such scores may not capture higher-order dependencies and may be too restrictive for algorithms capturing higher-order dependencies. Hence, we designed novel pathwise scores that account for higher-order dependencies.

The pathwise precision measures how well predicted edges are supported by paths in the true network. We define pathwise precision, \( P_p = \frac{1}{N} \sum_{i=1}^{N} S_i \), where \( N \) is the number of inferred edges and \( S_i \) is the support of the \( i^{th} \) predicted edge. \( S_i = \frac{1}{1 + l_i} \), where \( l_i \) is the length of the shortest path in the true network between the nodes of the \( i^{th} \) predicted edge. If no path exists, \( l_i = M + 1 \), where \( M \) is the number of nodes in the network. This definition of precision allows more flexibility when predicted edges correspond to higher-order structures, such as paths, in the true network. Recall is defined in a similar way where paths in the predicted network are compared with true edges.

We compare two network learning algorithms, Algorithm for the reconstruction of Accurate Cellular Networks (ARACNE) (Margolin et al. 2005) and Bayes networks (Friedman et al. 2000). The true network topology and node measurements are generated from a network simulation system. Bayes nets and ARACNE use the node values to infer the statistical correlation between the nodes. ARACNE, which optimizes mutual information, considers pairwise dependencies whereas Bayes networks, which optimize joint data likelihood, are not restricted to pairwise dependencies.

Our results with edgewise and pathwise scores are shown in figures 1 (a) and (b), respectively. The different precision and recall values correspond to different mutual information thresholds for ARACNE, and to maximum family size for Bayes nets. For edgewise scores, ARACNE performs better than Bayes networks. This is expected since ARACNE captures pairwise dependencies and edgewise scores evaluate exactly these dependencies. However, for pathwise scores, Bayes networks are better than ARACNE. This indicates that the statistical dependencies captured by Bayes networks have more support in the true network than those captured by ARACNE. These results highlight the importance of scores that evaluate different types of dependencies and therefore determine the suitability of structure learning algorithms to particular tasks.

**Future directions**

Our comparative analysis indicates that Bayes networks are well suited for capturing higher-order dependencies, which are characteristic of condition-specific networks. Because condition-specific networks are likely to have many cyclic dependencies, we prefer factor graphs over Bayes networks, since cyclic dependencies are not easily captured by Bayes networks. We intend to implement the structure and parameter learning algorithms for factor graphs proposed by Abbeel et al. (Abbeel, Koller, & Ng 2006) to learn coarse condition-specific networks. We will refine the coarse condition-specific networks by estimating values of physical attributes such as directionality and strength of interaction. Since most of these attributes are hidden, we will perform hidden variable inference and structure refinement by integrating information from different high-throughput assays. Finally, we will perform structural comparison of different condition-specific networks to obtain a wide spectrum of network level response of organisms under different environmental conditions.

**References**


---

1This work was supported by the Program in Interdisciplinary Biological and Bio-medical Science funded by HHMI, NIBIB and NIH.