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 biological networks are networks of genes and proteins induced in cells in response to different environmental conditions. The dependency structure of these networks provides important insight into how cells respond to healthy and stressful conditions.

Unfortunately, for most conditions the network structure is unknown and must be inferred from condition-specific measurements of the network nodes. My research goal is to develop a machine learning framework for inferring condition-specific networks by integrating static (e.g. interaction databases) and dynamic (e.g. gene expression) data. Machine learning approaches for condition-specific network inference must achieve two high-level goals: (a) infer different dependencies including pairwise, higher-order and cyclic, and (b) capture specific and generic subnetworks across different conditions.

I, with my advisors, Drs. Lane and Werner-Washburne, propose to model condition-specific networks using Markov random fields (MRFs), a class of undirected probabilistic graphical models. Unlike directed models such as Bayesian nets (Friedman 2004), which cannot explicitly represent cyclic dependencies, MRFs can represent cyclic as well as higher-order and pairwise dependencies. Unlike existing undirected model approaches, where learning is often restricted to structure refinements (Jaimovich et al. 2006), or to lower-order, often pairwise, dependencies (Margolin et al. 2005), we perform a complete search over MRF structures capturing different dependencies.

We describe an algorithm for learning the structure of MRFs, and methods to evaluate a structure learning algorithm's ability to capture different dependencies. We also propose to use the *physical network* (with edges corresponding to physical interactions among the network nodes), to bias our structure search. Because the physical network is largely incomplete, we employ classification methods to predict protein interactions absent from interaction databases.

*This work was supported by an HHMI-NIH/NIBIB grant (56005678), NSF grant (MCB0734918) to M.W.W., and NIMH grant (1R01MH076282-01) to T.L.

Existing approaches for identifying condition-specific behaviour of networks either use only gene expression (Segal et al. 2003), or do not learn generative models of the data (Oleg et al. 2007). Moreover, the generic and specific subnetworks across different conditions are usually identified *after* learning separate networks per condition. Our goal is to simultaneously learn specific and generic subnetworks using both gene and protein expressions.

Current work

Prediction of protein-protein interactions As a step to complete the physical network, we used a classification framework to predict protein interactions in yeast. We investigated the use of amino-acid composition (AAC), a simple amino acid count based feature for classification, and compared it with traditional features such as protein domains. A logistic regression classifier using only AAC was at par with classifiers trained on domain features; this trend was also noticed when using other classifiers like support vector machines. When using both domain and AAC, even higher classification accuracy was observed, indicating that AAC can boost performance of classifiers trained on other features. AAC's good performance, in spite of its strong independence assumptions, maybe due to its similarity to the *bag of words* model known to perform quite well in practice (Langley et al. 1992).

Coarse structure learning of undirected graphs We introduce *Markov Blanket Search* (MBS), a new algorithm for learning the structure of MRFs. MBS explicitly represents higher-order dependencies via a variable and its *Markov blanket* (MB). MBS learns the MRF structure by identifying the best MB per random variable (RV) that minimizes the conditional entropy of a variable given its MB. Our work is based on Abbeel *et al.*'s recent work of *MB canonical parameterization* (Abbeel *et al.* 2006). However, unlike Abbeel *et al.*, where MBs are learned for all subsets of n RVs up to size l, we learn MBs only per variable. We show that the MB independence assumptions for a set of RVs must hold for individual RVs, saving us $O(n^{l-1})$ computations during structure learning.

We compare MBS to Bayes nets (BNETs) and to an undirected pairwise model (PW) using scores measuring both pairwise and higher-order dependencies. Results for net-

works of known topology suggest that MBS dominates both BNET and PW, successfully capturing higher-order, pairwise, and cyclic dependencies. On real data, MBS identifed several higher-order structures that were not captured by PW or BNET.

Measuring higher-order dependencies Algorithms for network structure learning infer different networks depending upon the optimization criteria. For example, pairwise models optimize dependencies among exactly two variables, whereas higher-order models include dependencies among three or more variables. However, scores evaluating these algorithms typically measure accuracy of pairwise dependencies, matching only individual edges between the true and inferred networks (Margolin et al. 2005). As higher-order dependencies are not measured, the models truely capturing higher-order dependencies are not identified.

We introduce novel scores that measure higher-order dependencies in inferred networks. We assume meaningful higher-order dependencies to correspond to subgraphs of the true network, *true subgraphs*. We quantify the match using different higher-order scores measuring how well *edges* or *vertex degrees* of the true subgraphs are captured. Unlike pairwise scores, these scores compare *sets of edges* from the true subgraphs against the inferred network.

For example, we evaluate vertex degree match using a recall score: $R_{\text{VSG}} = \sum_{g \in \mathcal{S}_t} \bar{w}_g \frac{\sum_{v \in V_g} h_v}{\sum_{v \in V_g} h_v + m_v}$, where $\mathcal{S}_t$ denotes a set of subgraphs of a particular type such as cliques or cycles, $\bar{w}_g = \frac{|V_g|}{\sum_{f \in \mathcal{S}_t} |V_f|}$, V_g is the vertex set of subgraph g, h_v and m_v are the number of neighbours of v, matched and missed, respectively. These scores range in $[0, 1]$, and equal 1 when all subgraphs are matched perfectly. Comparison of different algorithms using both these and standard pairwise scores showed that higher-order models capture both higher-order and pairwise dependencies better than pairwise models.

Future plan

Incorporating biological prior for network structure learning Our algorithm currently infers the network structure *de-novo* without incorporating any biological knowledge. This makes structure learning slow and produces dependencies that may not all be biologically meaningful. We will guide our structure search by incorporating prior knowledge from existing protein interaction and ontology databases. These databases provide prior information about meaningful associations between RVs, which can be represented as edges of a *database-specific* graph of RVs. We can combine these database-specific graphs into a single multi-graph, with multi-edges representing evidence from multiple databases. The number of edges between two RVs can provide prior information of the strength of their interaction. The evidence multi-graphs also provides per-variable regularization allowing different RVs to have neighbourhoods of different sizes. In fact, our empirical results indicate that the neighbourhood size limit commonly used in network structure search often forces the graph topology to be regular, which is unrealistic for scale-free networks.

Inference of protein expression levels Our goal is to infer condition-specific networks not only using gene expression but also protein expression. Because protein expressions are not readily available we must infer them from observed gene expressions. The transcription factor (TF) protein expression can be modeled as a function of two components: one, determined by the expression of the gene coding the TF, and, the other, determined by expressions of the target genes regulated by the TF. Using a structual EM-like framework (Friedman 1997), we will allow both protein and gene expression to influence structure learning.

Learning condition-specific networks A simple approach for identifying condition-specific networks is to learn separate networks on data from separate conditions, and compare the individual networks. We hypothesize that the networks learned per condition share many components. Therefore, learning separate networks not only reduces the data available for structure learning but also makes the structure learning algorithm unaware of the information shared between the conditions.

Simultaneous learning of networks for multiple tasks has been addressed in both supervised (Geiger and Heckerman 1991) and unsupervised learning (Meila and Jordan 2000). Because we know the identity of the condition variable, condition-specific networks are similar to multi-nets in supervised learning. However, unlike multi-nets, we are interested in both discriminating (specific) and generic subnetworks across conditions. Furthermore, we require the graphs be undirected. We aim to extend existing multi-net frameworks for simultaneous identification of generic and condition-specific subnetworks.

References

Abbeel, P., *et al.* 2006. Learning factor graphs in polynomial time and sample complexity. *JMLR* 7:1743–1788.

Friedman, N. 1997. Learning belief networks in the presence of missing values and hidden variables. In *ICML*.

Friedman, N. 2004. Inferring cellular networks using probabilistic graphical models. *Science* 303:799–805.

Geiger, D., and Heckerman, D. 1991. Advances in probabilistic reasoning. In *UAI*.

Jaimovich, A., *et al.* 2006. Towards an integrated protein-protein interaction network: a relational markov network approach. *Journal of Comp. Biol.* 13(2):145–164.

Langley, P., *et al.* 1992. An analysis of bayesian classifiers. In *National Conference on Artificial Intelligence*, 223–228.

Margolin, A., *et al.* 2005. Aracne: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* (Suppl 1): S7.

Meila, M., and Jordan, M. I. 2000. Learning with mixtures of trees. *JMLR* 1:1–48.

Oleg, R., *et al.* 2007. Similarities and differences of gene expression in yeast stress conditions. *Bioinformatics* 23(2):e184–e190.

Segal, E. *et al.* 2003. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat Genet* 34(2):166–176.