生物ネットワークのデータマイニング

Datamining Biological Networks

京都大学化学研究所　バイオインフォマティクスセンター　Timothy Hancock

Abstract

The biological processes that occur within a cell are known to be organized in networks. There are many types of biological networks, each of which perform a specific function and are known to interact. However, as the network structures themselves are complex, the nature and structure of the interactions across networks are difficult to define. In this paper we assume that if a set of networks interact, the dynamics within each of these networks must share a common latent signal. To identify this signal across multiple networks we propose a Gaussian Process Latent Factor Model (GP-LFM). Our proposed model uses a GP-LFM to represent the observed time course at each network node based on the assumption that a diffusion process governs the dynamics within each network. From the assumption that diffusion of a common latent factor generates all observed node time courses we derive a latent force model. This latent force model explicitly predicts each node's time course based only on the neighborhood of that node and the strength of the diffusion process within that node's network. We then extend this idea to multiple networks by assuming that each network can be considered independent once the common latent function which generates all node time courses over all networks is known. Finally, we consider the transductive case, where some node time courses in some networks are unobserved and we wish to infer them. We show that by a re-parameterization of our proposed GP-LFM model we can estimate these missing node time courses without deviating from the standard GP-LFM optimization methods. We evaluate the performance of our GP-LFM on simulated and real data and clearly show that our GP-LFM approach is capable of identifying a common latent signal which is determines known dynamics across multiple networks.

1 Introduction

Cellular processes, such as regulation, signaling and metabolism, are often organized into network structures. It is common in bioinformatics and systems biology to analyze each of these networks separately in an effort to identify which sub-structures determine a observed phenomena. However, from a biological perspective it is known that these networks interact, but the nature and timing of this interaction on a network scale is unknown. Therefore, the creation of an observed phenomena in one network could be the result of entire network interaction, rather than any process generated from within a single network. Clearly a separate analysis of each network is insufficient to capture this level of complexity.

Unfortunately, methods to model interaction across multiple networks encounter representation problems; as each network possesses a different structure and has different numbers of nodes and the observational units across all networks may be of different types. For example, a regulatory network is a network of interacting genes or proteins whereas a metabolic network is a network of interacting chemical compounds. These differences in network architecture inhibit the ready encoding of any similarity between these networks. Another issue that must be considered is timing and sparsity of any network observations. For some networks types, such as those involving gene expression, high-throughput screening technology, such as microarray or next-generation sequencing, is available to give a snap-shot of the activity of all nodes present within the network at a specific time. For other network types such as PPI and metabolic, although high-throughput screening methods are available, such as Gas Chromatography Mass Spectroscopy (GC-MS), they do not provide a complete snap-shot of the state of all nodes within the network.

The methods in this paper are inspired by the need to infer interactions between multiple networks, which are related but not easily connected, and are known to possess a common signal. Our model does not require the networks to be connected, but treats each network as independent given knowledge of a common latent factor. We restrict ourselves to the analysis of time course data and assume that throughout all networks there is a common latent function which determines the time course profile at each node in each network. To estimate the latent function which is common across all networks and nodes, a Gaussian Process Latent Factor Model (GP-LFM) derived from a network diffusion process assumption. Nodes in each network are then weighted based on how much they resemble the latent function.

Our proposed approach is a Semiparametric Latent Factor Model (SLFM). SLFM's provide a general framework for sharing information from a single predictive function over multiple variables through the estimation of a mixing matrix. The SLFM approach provides the general framework for many Gaussian Process multi-task regression models. In these multi-task regression models the mixing matrix corresponds to estimated similarity of each task to an latent predictive function common to all tasks. SLFM approaches have the advantage of assuming no prior mixing matrix structure. In this paper we consider a variant of the SLFM approach where the mixing matrix is constrained to model diffusion of task similarity through a known network structure.

Specifically our approach falls into the latent force category of SLFM models. Latent force models are generative models of stochastic processes where a mechanistic model, in the form of differential equation, is assumed to account for a component of the observed stochastic variation. The parametric structure of this variation derived from the solution to the differential equation is used to define the SLFM approach which in turn generates a predictive latent function for the observed stochastic process. Latent force models have been shown to successfully model gene expression time courses, and inter latent protein profiles from mRNA expression. Our model falls into this category as we assume that a diffusion process on a network governs the time course profiles of each of the nodes.

Related work to our proposed idea is that of diffusion kernels and label propagation on a graph. Label propagation defines a latent function on a graph to be a categorical variable which assigns a label to each node. However, the node labeling is only observed on some partial subset of nodes of the graph, and the goal of label propagation is to learn the labels at the unobserved nodes. In the label propagation is performed as a random walk on a weighted graph Laplacian. The random walk style propagation has strong relationships to the diffusion processes on graphs. However, a diffusion kernel defines no labeling to the function being propagated but rather forms a general statement on how nodes on a graph share information.

The model we propose for identifying a latent function common to all nodes over multiple a networks is a stochastic Markov Random Field (MRF) model. Let $y_i(t)$ be the value of node i in graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ with nodes $\mathcal{V}$ and edges $\mathcal{E}$ observed at time t. Let each node "send" a fraction of its value, α_y to its neighbors, where $0 \leq \alpha_y \leq 1$. Therefore the value of any node, y_i at any future time $t+1$ can be assumed to be a stochastic process,

$$y_i(t+1) = y_i(t) + \alpha_y \sum_{j \in ne(i)} \left(y_j(t) - y_i(t)\right) \ . \tag{1}$$

The process is (1) defines a diffusion process over the graph $\mathcal{G}$ if the increments in time are taken to be infinitely small. However the nature of a diffusion process implies that there is a common latent function, $f(t)$ which is propagated through a connected subset of nodes in graph, $v \subset \mathcal{G}$. Therefore the nodes in subset v can be assumed to have a component which is a realization of this common latent process $f(t)$, the strength of which is dependent on the location within v and the strength of the network propagation α_y. Therefore, rather than take the limit of (1) as time increments tend to zero, we seek to explicitly identify the structure of the latent function, $f(t)$, and use it to predict the network dynamics at each node across multiple networks.

Furthermore, as we begin with a parametric statement on how the time course at any node in the entire network is generated, we also consider the case where time courses at some nodes are not observed. We show that by using transductive learning on the entire network structure we can estimate the weights of nodes where no time course are observed. Furthermore, we show that no additional optimization is required for this learning step and standard GP-LFM methods can be readily applied.

The results and conclusions section will present an application which seeks to identify a common function across regulatory, co-expression and metabolic networks using microarray and GC-MS data.

2 Results and Conclusions

Input Networks	Predictive r^2	Predictive correlation
REGULATORY	0.113	0.375
METABOLIC GENE	0.162	0.42
METABOLITE	0.142	0.442
REGULATORY + METABOLIC GENE	0.146	0.39
REGULATORY + METABOLITE	0.211	0.484
METABOLIC GENE + METABOLITE	0.268	0.537
REGULATORY + METABOLIC GENE + METABOLITE	0.244	0.515

Table 1: All combinations of real biological network network predictive performance results for control growth of E. coli.

The real analysis results for all combinations of biological networks are presented in Table 1. In Table 1 as there is no known latent time course the predictive correlation is the Pearson correlation coefficient between the predicted and observed time courses. The immediate result from this table is that both the r^2 and correlation can be improved when multiple networks are considered. This is most clearly observed with the regulatory network. The regulatory network alone offers no predictive performance ($r^2 = 0.113$). However when included with other networks the predictive performance rises to a maximum of $r^2 = 0.244$, corresponding to approximately 24% of time course variance modeled. This result clearly shows that there is a sharing of information between regulation, co-expression and metabolism, and including each of these elements can increase predictive performance. Additionally these results clearly show the success of our transductive model in overcoming the sparse observations within metabolite network with performances comparable with that of the densely observed expression networks.

We have proposed a GP-LFM based on diffusion kernel for prediction to time course data over multiple networks. Furthermore we extended our model to the transductive case, which allows for the reconstruction of time course data for nodes where no experimental data was available. Through extensive simulation experiments we clearly showed that our model can successfully predict a known latent function across multiple networks with large amounts of missing data. We further validated the performance of our model the systems biology setting and showed that our approach can successfully identify predictive latent functions which span regulatory, co-expression and metabolite networks.