Temporal Graphical Models for Cross-Species Gene Regulatory Network Discovery

Yan Liu* and Alexandru Niculescu-Mizil and Aurelie Lozano
IBM T.J. Watson Research Center,
Yorktown Heights, NY 10598, USA
Email: {liuya, anicule, aclozano}@us.ibm.com

Yong Lu*
Harvard Medical School, Harvard University,
Boston, MA 02115, USA
Email: Yong_Lu@hms.harvard.edu

Many genes and biological processes function in similar ways across different species. Cross-species gene expression analysis, as a powerful tool to characterize the dynamical properties of the cell, has found a number of applications, such as identifying a conserved core set of cell cycle genes. However, to the best of our knowledge, there is limited effort on developing appropriate techniques to capture the causal relations between genes from time-series microarray data across species. In this paper, we present hidden Markov random field regression with L_1 penalty to jointly uncover the regulatory networks for multiple species. The algorithm provides a framework for sharing information across species via hidden component graphs and can conveniently incorporate domain knowledge over evolution relationship between species. We demonstrate the effectiveness of our method on two synthetic datasets and one innate immune response microarray dataset.

1. INTRODUCTION

The activity of genes in a living cell is coordinated by a regulatory network that regulates gene expression conditioned on environmental stimuli. With genome-wide expression profiles, it is possible to reverse-engineer gene regulatory networks [1], which is essential for understanding how the cell functions. However, it remains a challenging task due to inherent and observational noise in expression data, the need to identify for each gene a small number of regulators among thousands of genes, and a very limited number of samples in each experiment.

Combining expression data from multiple species has been shown to help discover true associations between genes [2, 3]. The motivation is that many genes across species perform similar functions or share the same regulatory relations so that one can exploit information on related genes from multiple species. Similarly, expression data from multiple environmental conditions or cell types can be used to improve the prediction of gene functions [4], since many genes may share similar activities and regulatory patterns across various conditions and cell types.

In addition to improving prediction quality, cross-species expression analysis can identify conserved/common regulatory relations, which are more likely to play essential roles, as well as species-specific regulatory relations [5]. In the case of different environmental conditions and cell types, a combined analysis can identify common regulatory patterns as well as those specific to one cell type and/or one condition. With the exponential accumulation of microarray datasets, the benefits of cross-species analysis of expression data has become increasingly apparent [6].

A number of methods have been proposed for learning regulatory networks in a single species [7]. However, these methods do not take into account temporal patterns in time-series gene expression data. Other methods have been proposed to exploit information in temporal expression patterns. [8] applies auto-regression methods to causality inference on the expression data, which provides useful insights on the regulatory relationships between genes. More specifically, it combines Granger causality [9], an operational definition of causality well known in econometrics, and auto-regression algorithms with

*Corresponding author.

L_1 penalty to impose sparcity, for performing causality inference involving many variables. Similar methods have received considerable attention in other data mining problems [10, 11].

Computationally, inferring regulatory networks by cross-species analysis can be viewed as a multi-task learning problem. A multi-task learning method performs several related learning tasks simultaneously, borrowing information across tasks, instead of learning each task independently. In our application, one task refers to learning a regulatory network from time-series microarray data of one cell type, in a single species, and under some environmental condition. To the best of our knowledge, there is no systematic approach to jointly discover regulatory networks for several species by leveraging information across multiple species, cell types, and environmental conditions.

In this paper, we propose a novel probabilistic graphical model, *i.e.* hidden Markov random field with L_1 penalty, to solve this problem. It is based on the temporal causal models [10, 11, 8], but unlike [8], which can handle only one task (*i.e.* one species, cell type or environmental condition), our proposed method performs regulatory network discovery in a multi-task learning manner. Specifically, we assume the regulatory network for each task is generated from a mixture of hidden "shared component networks" (which are unobserved and to be inferred). Depending on the combination of species, cell type, and condition, the selection of component networks (which ultimately determines the regulatory network for each task) varies and its value can be learned from the data guided by the evolutionary distance between species. We also prefer sparse component networks by imposing L_1 penalty in the likelihood function. One major advantage of our model is the natural transfer of knowledge from one species, cell type or environment condition, to others. This is extremely important for time-series microarray data given the very limited number of samples (*i.e.* time points) available in each dataset. In addition, domain knowledge on the evolution distance of species can be naturally incorporated thanks to the graphical model framework.

In a related work, [12] proposes to use differential equations to infer regulatory networks by combining evolutionary cost and gene expression data across species. Unlike their work, which does not model time lag effect, our method explicitly takes into account the information from multiple previous time points when inferring causality, in order to better capture the properties of a biological system. There are other related work that address alignment of biological networks across species [5]. Network alignment methods take networks of the same type from several species as input, and the goal is to identify functionally conserved subnetworks. In contrast, our method takes gene expression time series from multiple species, and simultaneously infers the causal relationship between genes as well as similar subnetworks across species. Another related work is local alignment of network motifs [13], but it aims to address different goals, *i.e.* given the input of a single network in one species and a list of motifs, finding significant motifs present in the network. In [14, 4], a gene's dynamic property is summarized by computing an expression score from the time series, while our method uses all time points to infer causality, without first collapsing them into a single score.

The rest of the paper is organized as follows: we first review the temporal graphical modeling based on Granger causality in Section 2; then we motivate the challenges in cross-species analysis and describe the details of our proposed algorithms. We show experiment results on two synthetic datasets and on immune response expression data from human and mouse in Section 3. Finally, we summarize the paper and conclude with future work.

2. METHODOLOGY

Learning the graph structures of regulatory networks from microarray data have found great success. Recently, L_1-based auto-regression algorithms have been adapted and combined with Granger casuality [9] to discover the temporal "causal" networks between genes from time-series microarray data that reveals important dependency information between current observations and histories [8]. This approach serves as the foundation of our proposed algorithm. In what follows, we first review the temporal casual model, and next introduce the hidden Markov random field regression model (HMRF) for cross-species gene regulatory network discovery.

2.1. Notation

In this paper, we use the term "feature" to mean a time series (e.g. x) and use temporal variables to refer to the individual variables (e.g. x_t). In the context of microarray time series $\mathbf{x}$ with p number of genes over T number of time steps, a feature x_i denotes the time series of expression levels of a gene i, while a temporal variable $x_{i,t}$ refers to the expression level of a gene i at a given time point t. A lagged variable $x_{i,t-1} \dots x_{i,t-L}$ refers to concatenated histories of gene i from time $t-1$ to time $t-L$, where L is maximal time lag to be considered in the model.

2.2. Graphical Granger Modeling

"Granger Causality" [9] was introduced by the Nobel prize winning economist, Clive Granger, and has proven useful as an *operational* notion of causality in time series analysis in the area of econometrics. It is based on the intuition that a cause should necessarily precede its effect, and in particular that if a time series variable causally affects another, then the past values of the former should be helpful in predicting the future values of the latter. More specifically, let $\{x_{1,t}\}_{t=1}^{T}$ denote the time series variables for x_1 and $\{x_{2,t}\}_{t=1}^{T}$ the same for x_2. A time series x_1 is said to "Granger cause" another time series x_2, if given the following two regressions:

$$x_{2,t} \approx \sum_{j=1}^{L} a_j x_{2,t-j} + \sum_{j=1}^{L} b_j x_{1,t-j} \ , \qquad (1)$$

$$x_{2,t} \approx \sum_{j=1}^{L} a'_j x_{2,t-j} \qquad (2)$$

where L is the maximum "lag" allowed in past observations, eq (1) is more accurate than eq (2) with a statistically significant advantage, such as F-test [a].

The notion of Granger causality was defined only for a pair of time series. Recently, several graphical modeling approaches have been developed to determine the causal relationships between *multiple* time series variables [19, 20, 8]. These approaches are based on L_1 regularized regression (e.g. lasso), a more convenient and effective alternative to the exhaustive pairwise Granger tests among all the time series. Taking three time series x_1, x_2, x_3 as an example, for all i, these approaches regress $x_{i,t}$ in terms of the previous d values of all the time series, applying an L_1 penalty on the coefficients:

$$\hat{\beta} = \arg\min_{\beta} \sum_{t} (x_{1,t} - \sum_{i=1}^{3} \sum_{j=1}^{L} \beta_{i,j} x_{i,t-j})^2 + \lambda \|\beta\|_1 \qquad (3)$$

where λ is the parameter that controls the number of non-zero values in β. L_1 regularization is well known for variable selection, *i.e.* variables that are not significantly improving the accuracy of the model will have their values set to 0. This can be readily used to determine causality in the Granger sense: if any of the coefficients corresponding to a past value of x_j is non-zero, it means that it helps significantly to improve the accuracy of modeling the current value of x_i, and thus x_j is a cause of x_i in a Granger sense. We can represent the causal relationships between variables via a feature graph (see Figure 1(a) for an example).

2.3. Cross-species Regulatory Network Discovery

In Introduction, we identified our task of cross-species microarray analysis as an application of multi-task learning. Multi-task learning is a machine learning approach that learns a problem together with other related problems at the same time, using a shared representation [21]. One of the dominant approaches in multi-task learning is to model the tasks as generated from a linear combination of a set of base components (classifiers or networks). In other words, the relationship between multiple tasks can be explained by the fact that they share a certain number of hidden components [21]. Borrowing the idea, one simple approach to cross-species learning is to assume that the networks are generated from a mixture of hidden component networks. Mapping to the temporal causal models, we can think of the gene expression level $x_{i,t}$ of gene i at time t as generated from a mixture of regressions over lagged variables $x_{t-1} \dots x_{t-L}$. Depending on the species, cell type and environment condition, the mixture assignment may be different.

[a] Notice that the Granger Causality is not meant to be equivalent to true causality, but is merely intended to provide useful information regarding causation.

One major difference between our application and most previous multi-task learning settings is: we are also given rich prior knowledge on the relations between these tasks (for example, the evolutionary distance between species represented by the phylogenetic tree). This information can be abstracted as a relational graph G, in which a node corresponds to a species, cell type or condition, and there is an edge between two nodes if they share the same cell type, species, or condition [b]. The relational graph G provides essential guidance for inferring the hidden component networks and mixture assignments. The computational challenge is how to incorporate the graph in the modeling framework.

2.3.1. *Data processing and relational graph construction*

To conduct cross-species microarray analysis, the first step is to decide on a common universe of genes for study. Among the many possible ways, we choose to select the subset of genes or orthologous genes that are shared across all the datasets. More specifically, we first choose a benchmark species that is close in evolution to all the concerned species. It may or may not appear in our collection of microarray datasets. Next, for each microarray dataset, we map the genes to their orthologous genes in the benchmark species. It is possible that one gene might map to multiple orthologs. In this case, we will keep the mapping as a set of orthologs. Finally, we select the subset of genes (from the benchmark species) that are shared by all the datasets. In this way, we get a common universe of genes for cross-species microarray analysis. Notice that when outputting the gene regulatory networks, we map the benchmark genes back to their corresponding genes from the original species.

The evolution paths between species provide important guidance on how regulatory networks from different species can be similar (or dissimilar). We represent this qualitative information via the relational graph G, in which a node represents one species (or a cell type under some condition). There is an edge between two nodes if the corresponding microarray experiments are on the same species but under different cell type/condition because we expect many genes would exhibit similar regulatory relations; there is also an edge between two nodes representing microarray data of the same cell type and condition but from different species if two species are evolutionarily related. The motivation is that some genes may share similar regulated functions as their orthologs.

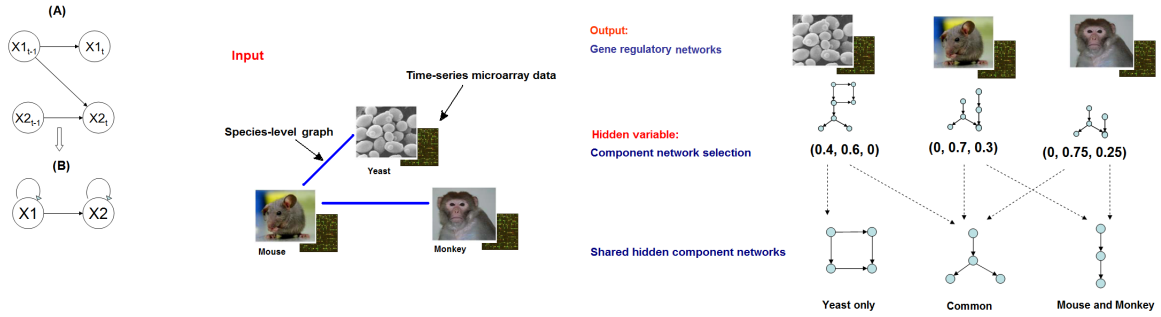
Figure 1 shows an example of the species-level relational graph G for three species, *i.e.* yeast, mouse and monkey. Essentially it is a reduced phylogenetic tree by removing all the unconcerned species. We will show an example of the relational graph in Section 3 for two species with different cell types and conditions. Notice that our proposed model only needs qualitative domain knowledge (*i.e.* the relational graph) as input and uses them as guidance for multi-task learning. As discussed later, the model automatically infers the degree of similarity between species or cell types/conditions from the data, which is also one of the major advantages of our model.

2.3.2. *Hidden Markov Random Field Regression*

A hidden Markov random field (HMRF) [22] is a generalization of hidden Markov model (HMM). It is a stochastic process generated by a Markov random field whose state (which in our application refers to the selection of component networks for each species) cannot be observed directly but through other related observations. One important feature of the HMRF model is the encoded contextual constraints between the states of neighboring nodes in the relational graph. HMRF has been successfully applied to many applications with relational information, such as image segmentation, genetics, and disease mapping.

In order to integrate the species-level constraints from domain knowledge with multi-task learning (*i.e.* cross-species regulatory network discovery), we extend HMRF to regression. The basic assumption is that the time-series are generated from a stochastic process, where the current observation of node i (*i.e.* species i) $x_t^{(i)}$ is from a mixture of regressions over lagged variables $x_{t-1}^{(i)}, \ldots, x_{t-L}^{(i)}$. The hidden states $s^{(i)}$ associated with species i determines the selec-

[b] Notice that this relational graph G is different from the output regulatory networks and the shared component networks.



(a) (b) (c)

Fig. 1. (a) Demonstration to convert a temporal graph (A) to feature graph (B); (b,c) Demonstration of HMRF-regression for cross-species gene regulatory network discovery. Input: microarray data from multiple species (c-Input 2) and domain knowledge on the evolution path between species (cell-type or environment)(b-Input 1). Output: the regulatory networks for each species (cell-type or environment). The algorithm assumes the regulatory networks of each species are generated from a linear mixture of common component networks. By maximizing regularized likelihood of the data, we can infer the shared hidden component networks as well as the value of hidden variable, *i.e.* the selection of hidden component networks, for each species.

tion of regression coefficients (and ultimately determines which component networks contribute to the output regulatory networks). More specifically, given M number of time-series observations, where node i corresponds to $x^{(i)} = [x_1^{(i)}, \ldots, x_N^{(i)}]^T$, we can define the joint probability of time-series observations and hidden states as a product of node potentials and pairwise edge potentials, *i.e.*

$$P(x^{(1)}, \ldots, x^{(M)}, s^{(1)}, \ldots, s^{(M)} | \beta, \Sigma, w) = \frac{1}{Z} \prod_{i=1}^{M} \Phi(x^{(i)}, s^{(i)} | \beta, \Sigma) \prod_{(i,j) \in \text{edge}} \Phi(s^{(i)}, s^{(j)} | w) \quad (4)$$

where the node potential $\Phi(x^{(i)}, s^{(i)} | \beta, \Sigma)$ is a product of multivariate Gaussian distributions, *i.e.*

$$\Phi(x^{(i)}, s^{(i)} | \beta, \Sigma) = \prod_{t=1}^{N_{(i)}} \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} \times \exp(-\frac{1}{2}(x_t^{(i)} - o_t^{(i)} \beta_{s^{(i)}})^T \Sigma_{s^{(i)}}^{-1} (x_t^{(i)} - o_t^{(i)} \beta_{s^{(i)}})) \quad (5)$$

$o_t^{(i)}$ is a concatenated matrix of lagged observations $[x_{t-1}^{(i)}, \ldots, x_{t-L}^{(i)}]^T$, and p is the dimension of $x_t^{(i)}$; Σ is the covariance matrix and β is the coefficient, whose value determines the edges of the networks. The edge potential $\Phi(s^{(i)}, s^{(j)} | w)$ is defined as

$$\Phi(s^{(i)}, s^{(j)} | w) = \exp(\sum_k w_k \delta_k(s^{(i)}, s^{(j)})) \quad (6)$$

where δ is the indicator function, *i.e.* $\delta_{k=(s,s')}(s^{(i)}, s^{(j)}) = 1$ if $s^{(i)} = s$ and $s^{(j)} = s'$, and 0 otherwise; $w_{(s,s')}$ is the parameter to evaluate the similarity between state s and s', similar to the transition probability in HMM; Z is the normalization constant. By our definition of node potentials, the value of Z will only be affected by the edge potentials, *i.e.*

$$Z = \sum_{s^{(1)}, \ldots, s^{(M)}} \exp(\sum_{(i,j) \in \text{edge}} w_{s^{(i)}, s^{(j)}} \delta(s^{(i)}, s^{(j)})) \quad (7)$$

In summary, the model aims to infer the hidden component networks (captured by the regression coefficients β_s) via mixture of regressions [23] (*i.e.* the node potential); in addition, the assignment of hidden states, *i.e.* the mixture selection, is constrained by species-level graph from domain knowledge (*i.e.* the edge potential).

There are three sets of parameters in the model, namely β, Σ and w. Since the value of the state variables $s^{(1)}, \ldots, s^{(M)}$ is not known, EM algorithm [24] can be applied to estimate the parameters. We skip the details of the derivation due to limited space, but note two observations: (1) it turns out that the solution to β_s can be achieved as a normal linear regression by reweighting the observed variables $o_t^{(i)}$ and response variables $x_t^{(i)}$ with weights $\widetilde{P}^{(i)}(s)$, *i.e.* the posterior probability of node i with state s; (2) the exact estimation of $\widetilde{P}^{(i)}(s)$ is infeasible, and approximate inference algorithm, such as loopy belief propagation [25], can be applied.

2.3.3. *Extending HMRF regression with L_1 penalty*

Next we examine how to extend HMRF regression to L_1 penalty so that the learned component network are sparse. The model is referred to as HMRF-L_1 in later discussion. Following the idea in [26], we add a Laplacian prior for β as follows:

$$P(\beta|\lambda) = (\lambda/2)^N \exp(-\lambda\|\beta\|_1), \tag{8}$$

where λ is the hyperparameter and determines the number of non-zero values in coefficients β. As a result, the auxiliary objective function Q relevant to β in the EM-algorithm has the following form:

$$Q_\lambda = -\sum_{i=1}^{M}\sum_{t=1}^{N^{(i)}}\sum_{s^{(i)}} P(s^{(i)}|x^{(1)},\ldots,x^{(M)},\widetilde{\beta},\widetilde{\Sigma},\widetilde{w})\times \tag{9}$$
$$(x_t^{(i)} - x_{t-1..t-L}^{(i)}\beta_{s^{(i)}})^T\Sigma_{s^{(i)}}^{-1}(x_t^{(i)} - x_{t-1..t-L}^{(i)}\beta_{s^{(i)}}) - \lambda\|\beta\|_1$$

Recent reexamination of gradient-based optimization algorithms, such as the coordinate descent, has shown that they are very effective for solving lasso-type regressions [27]. We compute the first derivative of Q_λ with respect to β_s and then apply stochastic gradient algorithms to get the solution of β. Notice that other regression algorithms with L_1 penalty, such as elastic net and group lasso, can also be extended similarly. We refer readers to [27] for details.

2.4. HMRF-L_1 for Regulatory Network Discovery

After applying the HMRF regression to cross-species microarray data, we need to output the regulatory network for each combination of species, cell type and condition by selecting a subset of the shared component graphs. In this paper, we use a heuristic weighted average approach: for each gene i, reweighting the base graph of state s (represented by coefficient β_s) with its mixing proportion $\widetilde{P}^{(i)}(s)$. Then we decide that there is an edge between two nodes if and only if the corresponding coefficients in the weighted average matrix $\sum_s \widetilde{P}^{(i)}(s)\beta_s$ are above some threshold. In our experiment, the threshold is set to 0.05. A summary of the workflow is demonstrated in the following algorithm.

Algorithm: HMRF-L_1 for temporal graph structure learning

1. **Input:** For each gene i, we are given time series data $x^{(i)} = \{x_1^{(i)}\ldots x_{N_{(i)}}^{(i)}\}$ where $x_t^{(i)}$ is a p-dimensional vector;
   Parameters: (1) time lag L; (2) number of hidden states K; (3) threshold θ
   Function input: regression function f
2. Run HMRF-L_1 and output coefficients β_s for each state s, the mixing of hidden states $\widetilde{P}^{(i)}(s)$
3. For each gene i, iterate the following steps:

3.1 Initialize the adjacency matrix for the p features, *i.e.* $G = \langle V, E\rangle$ where V is the set of p features.

3.2 For each feature $x_u \in V$ place an edge $x_u \to x_v$ into E, if and only if at least one of the corresponding coefficients for x_u in $\sum_s \widetilde{P}^{(i)}(s)\beta_s$ is above threshold θ.

---

3. EXPERIMENT RESULTS

The goal of the experiments is to demonstrate multi-task learning (*i.e.* cross-species regulatory networks) by our proposed model is able to achieve better results than learning each single task independently or naively concatenating the observations from different tasks and yielding one output for all tasks. Therefore we compare the performance of HMRF-L_1 with two other baselines: one is ALL, namely aggregating all the observations from different tasks (or microarray datasets for different species or cell type/condition) and learning one single graph; the other is SUB, namely learning a graph from the observations of individual task without considering those from other tasks. Both ALL and SUB use the auto-regression algorithm with L_1 penalty as discussed in Section 2.2. We conduct experiments on two simulation datasets and then apply our model to cross-species innate immune response analysis.

3.1. Simulation data

The two simulation datasets are both generated from a 2-state MRF, whose graph structure is a 10×10 grid (notice that this corresponds to species-level graph in the application of cross-species regulatory network discovery), and the coefficients are defined as follows: $w(i,i) = 1$ and $w(i,i') = 0.5$ for $i \neq i'$.

The observations of each node (*i.e.* each task) are generated from Gaussian distributions using examples of the AR models used in [17, 18] (notice that this corresponds to the gene-regulatory networks of individual species in the application of cross-species regulatory network discovery). More specifically:

Simulation Data I assume that state 1 corresponds to a AR(1) model with the inverse of the covariance matrix as follows: $(\sum^{-1})_{ii} = 1$, $(\sum^{-1})_{i,i-1} = (\sum^{-1})_{i-1,i} = 0.5$, and state 2 corresponds to sparse scenario, $(\sum^{-1})_{ii} = i$. The goal of conducting experiments on this dataset is to verify whether our algorithm is able to recover the sparse component graph from data mixed with dense component graph.

Simulation Data II contain data generated from Gaussian distributions of inverse covariance with similar graph structures: state 1 corresponds to the same distribution as state 1 in Simulation Data I, and state 2 corresponds $(\sum^{-1})_{ii} = 1$, $(\sum^{-1})_{i,i-1} = (\sum^{-1})_{i-1,i} = 0.5$, $(\sum^{-1})_{i,i-2} = (\sum^{-1})_{i-2,i} = 0.25$. Our goal is to examine whether the algorithm can recover the true graphs when the underlying two component graphs are similar, which better mimics our application data on cross-species gene regulatory networks.

In the experiment, we sample the values of underlying hidden states for all the nodes using Gibbs sampling; then for each node, we generate 20 samples from the underlying distributions determined by the value of hidden states. The penalty terms λ are selected by cross-validation. We evaluate the performance of structure learning methods using the F1-measure, *i.e.* viewing the causal modeling problem as that of predicting the inclusion of the edges in the true graph, or the corresponding adjacency matrix. Recall that, given precision P and recall R, the F1-measure is defined as $F1 = 2PR/(P+R)$, and hence strikes a balance in the trade-off between the two measures (see [28] for example of using these metrics in evaluation of structural learning methods).

We repeated the experiments 30 times and report the average on Table 1. As we can see, HMRF-L_1 achieves better performance than competing methods on both Simulation Data I and II.

Table 1. Comparison Results of Structure Learning on Simulation Data

Algorithm	Simulation I (F_1)		Simulation II (F_1)	
	State 1	State 2	State 1	State 2
HMRF-L_1	**0.8674**	**0.6093**	**0.8396**	**0.5853**
ALL	0.8214	0.3763	0.8352	0.5115
SUB	0.5356	0.4508	0.6258	0.4879

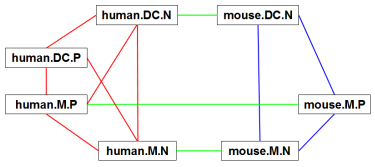
3.2. Applications to Cross-Species Gene Regulatory Network Discovery

Most multicellular organisms rely on their immune system to defend against the infection from a multitude of pathogens. There have been many studies using microarray data to compare immune gene expression programs under different conditions [29–31]. To understand the roles and possible interplays between different types of immune cells, it is important to identify both regulatory relations common to different immune cells and those specific to a certain cell type. While each of these subsets of experiments (macrophages vs. dendritic cells, human vs. mouse etc.) can be analyzed separately and then compared to each other, the learned biological networks become much less reliable due to the noise and limited samples in gene expression data. It is therefore desirable to combine microarray gene expression data from different studies to overcome these challenges and jointly infer regulatory networks involved in immune response.

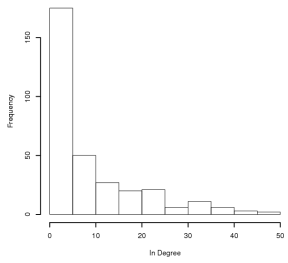
We applied our algorithm to learn the causal networks between genes for immune response system. Specifically, we collected time-series microarray datasets on innate immune response of human and mouse from the supporting websites of [30, 32–34, 31, 29, 35–37]. The gene expression experiments were done on macrophages (M) and dendritic cells (DC) in humans and mice, under the infection of two types of bacteria, Gram-positive (P) and Gram-negative (N). The only exception is mouse dendritic cells, where we only found data on Gram-negative bacteria. The 39 microarrayexperiments are grouped into seven datasets, and referred to as "human.DC.N", "human.DC.P", "human.M.N", "human.M.P", "mouse.DC.N", "mouse.M.N" and "mouse.M.P" respectively (see [14] for full details of the data).

In order to exploit information shared across species/cell types, we process the data as follows:

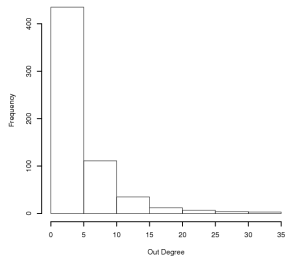
for those experiments on the same species, we only select the genes that appear in all the experiments. This results in 3869 genes for mouse and 1651 genes for human; next we obtain the human and mouse orthologs from Mouse Genome Database [38], and select the common candidate regulatory genes where either themselves or their orthologs can be found in our dataset. This results in a set of 789 common genes across species. We construct the species-level graph as follows: there is an edge between two experiments on the same species if they share the same cell type or the same infection type; there is also an edge between the same cell type and infection type across different species because we expect that some of the genes may share the similar regulated functions as their orthologs. This results in the species-level graph as Figure 2(a).



(a) Species-level Graph



(b) In-degree Distribution



(c) Out-degree Distribution

Fig. 2. (a) Species-level Graph. Red/blue edges: dependency due to same species (*i.e.* human/mouse); green edges: dependency due to same experiments; (conveniently generated from domain knowledge) (b) Distribution of in-degree counts (c) Distribution of out-degree counts

We varied the number of hidden component graphs from 2 to 7 and choose 4 by Bayesian information criterion (BIC) score. We ran experiments for a maximum lags of 2. There is an edge in the component graph if and only if the absolute value of its corresponding coefficients are larger than 0.05. In the end, we have around 2000 edges in each component graph. Generally, the degree of the nodes in the component graph roughly follows the power law (Figure 2(b, c)).

3.2.1. *Component-Independent Regulations*

In order to better examine the results, we divided the genes in a component graph into three classes based on their connectivity: genes with only out-going edges, genes with only incoming edges, and genes with both types of edges. As we show later, genes in each class have demonstrated different characteristics.

We identified the top ten well-connected genes that have only out-going edges and common to all component graphs (Table 2). The list contains a number of chemokines and receptors, which are consistent with the hypothesis that genes in this class serve to sense environment and inter-cellular communication. E.g. IL1R2 is a receptor for pro-inflammatory interleukin 1 (IL-1) and related to cell migration [39]. NFkB is a transcription factor that can be activated by intra-/extra-cellular stimuli including cytokines and bacterial products. CXCL10 is a chemokine which can trigger many effects including stimulation of immune cells.

We identified the top ten well-connected genes with only incoming edges (Table 3). These genes are involved in various cellular processes. E.g. CCT5 is a member of TCP1 ring complex that folds various proteins including actin and tubulin. CYP2E1 is an enzyme that catalyzes many reactions involved in drug metabolism. CD1 mediates the presentation of primarily lipid and glycolipid antigens of self or microbial origin to T cells.

Next, we look at densely connected subgraphs in the learned component graphs. To further enforce sparcity, we apply a more stringent threshold (0.2) on the absolute value of the edge weights. Here we show one example of the subgraphs (Figure 3). The genes with only out-going edges in this subgraph include a number of genes located on the membrane, e.g. CD14 [40], and HLA class II histocompatibility antigen (HLA-DRA), which are expressed in antigen presenting cells and play a central role in the immune system [41]. The middle layer of the subgraph includes GTP binding protein (GTPBP1), and CDKN1A, a cell cycle regulator, and VIM, which is involved in attachment, migration, and cell signaling [42]. The

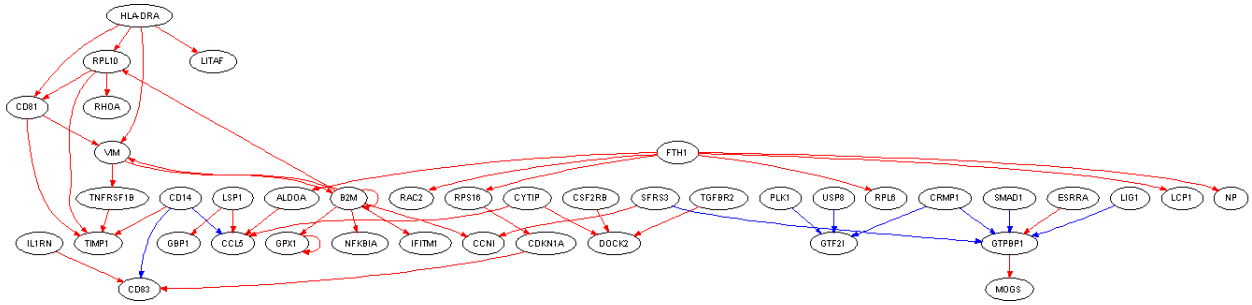


Fig. 3. An example of densely connected subgraphs in the learned component graphs. The regulation relation can be either positive (red edges) or negative (blue edges).

bottom level of the graph includes genes mediate signal transduction (CD83), important chemokines (CCL5), and interferon-induced GTPase (GBP1).

Table 2. Top 10 Well-connected Genes by Out-going Edges

Out-degree	Symbol	Description
43	FTH1	ferritin, heavy polypeptide 1
25	RPL37	ribosomal protein L37
20	IL1R2	interleukin 1 receptor, type II
18	NFKB1	nuclear factor of kappa light polypeptide gene enhancer in B-cells 1
18	CXCL10	chemokine (C-X-C motif) ligand 10
17	CYTIP	cytohesin 1 interacting protein
14	DUSP2	dual specificity phosphatase 2
12	PTGS2	prostaglandin-endoperoxide synthase 2
12	MMP12	matrix metallopeptidase 12
12	LSP1	lymphocyte-specific protein 1

Table 3. Top 10 Well-connected Genes by In-coming Edges

In-degree	Symbol	Description
28	CCT5	chaperonin TCP1 subunit 5
28	PCNA	proliferating cell nuclear antigen
21	CYP2E1	cytochrome P450, family 2, subfamily E, polypeptide 1
20	NEDD4	neural precursor developmentally down-regulated 4
16	ZFHX3	zinc finger homeobox 3
15	EXT2	exostoses (multiple) 2
13	CLK3	CDC-like kinase 3
12	NMT1	N-myristoyltransferase 1
10	CBX5	chromobox homolog 5
10	CD1D	T-cell surface glycoprotein CD1d

Table 4. Example of Component-specific Hubs

Graph Name	Genes
Component 1	HLA-DRA, ID1, CTSB, ELK1, CDKN2A
Component 2	TSC22D3, ACVR2A, EPHA5, NFE2, PCTK3
Component 3	PIK3R1, TK2, IL1R1
Component 4	ASNS, MAP4K1, KCNH2, INPPL1, COL9A2

3.2.2. *Component-specific Regulations*

Next we compare the component graphs and identify characteristics specific to each graph. First, we compare the hub genes in each component graph, which are defined as genes with at least 5 outgoing edges and no incoming edges. We identify a total of 40 component specific hub genes. For component graph 1, the list includes genes involved in cell cycle control (E2F1, CDKN2A) and wound repair (MMP3). In addition, ITGA7 is involved in cell-cell interaction, and MAP3K8 can induce the production of NFkB. For component graph 2, the list includes TSC22D3, which plays a key role in the anti-inflammatory process. Hub genes in component 3 include Il1R1, interleukin 1 receptor, and PIK3R1, which is involved in metabolism of insulin. For component 4, hub genes include IRF9, a regulatory factor of interferons (proteins released by cells in response to pathogens), and MYH9, which has a function in the maintenance of cell shape. Some of the component-specific hub genes are listed in Table 4.

To characterize the genes with high incoming edges in each component graph, we examine the genes with at least 20 incoming edges and confirmed enriched GO categories [43]. For example, some of the top enriched categories include "Regulation of Glucose Transport" (component 1; corrected $pval = 0.002$), "Leukocyte Homeostasis" (component 2 and 3; corrected $pval < 0.001$), "Locomotory Behavior" (component 2 and 3; corrected $pval < 0.006$), and "Double-strand break repair" (component 4; corrected $pval = 0.034$).

Table 5. Top Five Component-Specific Enriched Biological Processes

Component 1	Component 2	Component 3	Component 4
regulation of glucose import	response to organic substance	response to organic substance	double-strand break repair
glucose import	leukocyte homeostasis	leukocyte homeostasisstimulus	response to abiotic stimulus
regulation of glucose transport	homeostasis of number of cells	cellular response to stimulus	cellular response to stimulus
response to organic substance	cellular response to stimulus	response to chemical stimulus	response to heat
response to peptide hormone stimulus	positive regulation of cellular process	positive regulation of cellular process	positive regulation of catabolic process

3.2.3. *Comparison with Other Approaches*

We also compare the learned networks generated by HMRF-L_1 with those by two other baselines: one is aggregating the samples from all microarray datasets and learn one network ("ALL"), and the other is to learning a network from individual dataset only ("SUB"). Compared with SUB, our method has major advantages since some datasets, for example, Human.DC.P and Mouse.M.N, have very limited number of time-series observations (1-2), and no reasonable graph can be generated by SUB. For fair comparison (in favor of SUB method), we choose the dataset with the largest number of time-series observations (Human.M.N), to compare the results of different methods. One general observation is that the networks by ALL (31,218 edges) and SUB (14,346 edges) are much denser than that by HMRF-L_1 (7458 edges) while the three graphs share 4,071 edges in common. Sparse graphs do not necessarily suggest better performance, but among all the edges uncovered by HMRF-L_1, around 54.6% of them are also found in other methods, which seems to suggest higher precisions. Figure 6 lists an example of 10 genes with the highest number of out-degrees in the learned networks. From the results, we can see that HMRF-L_1 not only shares some top-ranked genes with the other two algorithms, such as CXCL10, but also uniquely identifies important immune genes, such as IL1R2, HLA-DRA, and CD14, as well as B2M (Beta-2-microglobulin), which is a serum protein found in association with the major histocompatibility complex (MHC) class I heavy chain on the surface of nearly all nucleated cells; MSN (Moesin), which is localized to filopodia and other membranous protrusions that are important for cell-cell recognition and functions as cross-linkers between plasma membranes and actin-based cytoskeletons.

Table 6. Top 10 Genes by Out-degress in the Learned Networks by Different Methods

HMRF-L_1		ALL		SUB	
EntrezID	Count	EntrezID	Count	EntrezID	Count
FTH1	182	PTGS2	170	ACVR2A	224
IL1R2	110	ACVR2A	157	VPS45	179
B2M	104	CXCL10	154	PTGS2	175
VIM	75	DUSP2	145	NFE2	172
CXCL10	74	PPIB	140	FTH1	168
RPL37	71	FMO1	136	FOS	167
LSP1	70	PECAM1	135	PECAM1	162
HLA-DRA	68	NR4A1	132	FPR1	160
MSN	66	MCM4	131	CDC6	157
CD14	60	IL7R	128	LSP1	140

3.2.4. *Bootstrap Evaluation*

In addition, we also evaluate the performance of our method by applying the Bootstrap procedure, which is a technique widely used in statistics for evaluating statistical accuracy (see [44] for a review). More precisely, given the original lagged data matrix, we randomly draw B datasets by sampling with replacement the rows of the original data matrix, so that each dataset has the same number of rows as the original lagged data matrix. We then apply our method to each of the B bootstrap datasets. Comparing the original network (*i.e.* the network obtained by using the original dataset) with the bootstrap networks (*i.e.* those obtained using the bootstrap datasets) allows us to get a measure of confidence in the causal relationships identified in the original network. In particular, for each causal relationship identified in the original network, we can get confidence in that relationship by counting the number of times it appears in the bootstrap networks. As shown in Table 7, the causal relationships identified by our method in the original network appear on the average 75.2% of the time in the bootstrap networks, which demonstrates that HMRF-L_1 produces stable networks.

Table 7. Percentage of Overlap between Bootstrap Networks and Original Networks

Species Type	% of Overlap
human.DC.N	0.7572
human.DC.P	0.7569
human.M.N	0.7541
human.M.P	0.7575
mouse.DC.N	0.7713
mouse.M.N	0.7510
mouse.M.P	0.7527

4. CONCLUSION

In this paper we examine the problem of discovering regulatory networks from multi-species time-series microarray data by leveraging the common regulation information across species. We develop hidden Markov random field regression with L_1 penalty to extend temporal Granger modeling to multi-task learning. We show that our method is able to uncover causal relations on two synthetic datasets, as well as conserved regulatory network common to two types of cells in humans and mice and shared between response to different types of bacteria. For future work, we are interested in more systematic evaluation of the experiment results. We also plan to apply our model for other types of cross-species regulatory network discovery, such as antifungal drug resistance.

Acknowledgments

We sincerely thank Naoki Abe, Hongfei Li, Jonathan Hosking, Rick Lawrence and Piotr Mirowski for discussing the ideas in the paper. We thank anonymous reviewers for their valuable suggestions.

References

1. P. D'haeseleer, S. Liang, and R. Somogyi. Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics*, 16(8):707, 2000.
2. S. Bergmann, J. Ihmels, and N. Barkai. Similarities and differences in genome-wide expression data of six organisms. *PLoS Biol*, 2(1):e9, 2003.
3. J.M. Stuart, E. Segal, D. Koller, and S.K. Kim. A gene-coexpression network for global discovery of conserved genetic modules. *Science*, 302(5643):249, 2003.
4. Y. Lu, R. Rosenfeld, G. J. Nau, and Z. Bar-Joseph. Cross species expression analysis of innate immune response. In *Journal of Computational Biology*, 17(3):253–268, 2010.
5. R. Sharan and T. Ideker. Modeling cellular machinery through biological network comparison. *Nature Biotechnology*, 24:427–433, 2006.
6. Y. Lu, P. Huggins, and Z. Bar-Joseph. Cross species analysis of microarray expression data. *Bioinformatics*, 25(12):1476, 2009.
7. N. Friedman. Inferring cellular networks using probabilistic graphical models. *Science*, 303(5659):799, 2004.
8. A. Lozano, N. Abe, Y. Liu, and S. Rosset. Grouped graphical granger modeling for gene expression regulatory networks discovery. In *Proceedings of International Conference on Intelligent Systems for Molecular Biology (ISMB-09)*, 2009.
9. C. Granger. Testing for causlity: A personal viewpoint. *Journal of Economic Dynamics and Control*, 2:329–352, 1980.
10. R. Dahlhaus and M. Eichler. Causality and graphical models in time series analysis. *Highly structured stochastic systems*, 2003.
11. N.D. Mukhopadhyay and S. Chatterjee. Causality and pathway search in microarray time series experiment. *Bioinformatics*, 23(4), 2007.
12. G. Bourque and D. Sankoff. Improving gene network inference by comparing expression time-series across species, developmental stages or tissues. *J Bioinform Comput Biol.*, 2(4):765–83, 2004.
13. J. Berg and M. Lässig. Cross-species analysis of biological networks by bayesian alignment. *PNAS*, 103(29):10967–10972, 2004.
14. Y. Lu, S. Mahony, P. Benos, R. Rosenfeld, I. Simon, L. Breeden, and Z. Bar-Joseph. Combined analysis reveals a core set of cycling genes. *Genome Biology*, 8(7):R146, 2010.
15. N. Meinshausen and P. Buhlmann. High dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 34(6):1436–1462, 2006.
16. M. J. Wainwright, P. Ravikumar, and J. D. Lafferty. High-dimensional graphical model selection using ℓ_1-regularized logistic regression. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1465–1472. 2007.
17. M. Yuan and Y. Lin. Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.
18. J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostat*, 9(3):432–441, 2008.
19. A. Arnold, Y. Liu, and N. Abe. Temporal causal modeling with graphical granger methods. In *Proceedings of International Conference on Knowledge Discovery and Data Mining (SIGKDD-07)*, 2007.
20. A. Lozano, N. Abe, Y. Liu, and S. Rosset. Grouped graphical granger modeling methods for temporal causal modeling. In *Proceedings of International Conference on Knowledge Discovery and Data Mining (SIGKDD-09)*, 2009.
21. R. Caruana. Multitask learning. *Ph.D. Thesis,*

School of Computer Science, Carnegie Mellon University. 1997.
22. H. Kunsch, S. Geman, and A. Kehagias. Hidden markov random fields. *Ann. Appl. Probab.*, 5(3):577–602, 1995.
23. C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2007.
24. J. A. Bilmes. A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. Technical report TR-97-021, ICSI, 1997.
25. K. P. Murphy, Y. Weiss, and M. I. Jordan. Loopy belief propagation for approximate inference: An empirical study. In *Proceedings of Uncertainty in AI*, pages 467–475, 1999.
26. S. I. Lee, H. Lee, P. Abbeel, and A. Y. Ng. Efficient l1 regularized logistic regression. In *Proceedings of AAAI*, 2006.
27. J. Friedman, T. Hastie, and R. Tibshirani. Regularized paths for generalized linear models via coordinate descent. Journal of Statistical Software, 33(1), 2010.
28. R. Silva, R. Scheine, C. Glymour, and P. Spirtes. Learning the structure of linear latent variable models. *J. Mach. Learn. Res.*, 7:191–246, 2006.
29. Q. Huang, D. Liu, P. Majewski, L.C. Schulte, J.M. Korn, R.A. Young, E.S. Lander, and N. Hacohen. The plasticity of dendritic cell responses to pathogens and their components. *Science*, 294(5543):870–875, 2001.
30. D Chaussabel, R.T Semnani, M.A. McDowell, D. Sacks, A. Sher, and T.B. Nutman. Unique gene expression profiles of human macrophages and dendritic cells to phylogenetically distinct parasites. *Blood*, 202:672–681, 2003.
31. R. Hoffmann, K. van Erp, K. Trulzsch, and J. Heesemann. Transcriptional responses of murine macrophages to infection with Yersinia enterocolitica. *Cellular Microbiology*, 6(4):377–390, 2004.
32. C.S. Detweiler, D.B. Cunanan, and S. Falkow. Host microarray analysis reveals a role for the salmonella response regulator phop in human macrophage cell death. *Proc. Natl. Acad. Sci.*, 98:5850–855, 2001.
33. D.W. Draper, H.N. Bethea, and Y.W. He. Toll-like receptor 2-dependent and-independent activation of macrophages by group B streptococci. *Immunology letters*, 102(2):202–214, 2006.
34. F. Granucci, C. Vizzardelli, N. Pavelka, S. Feau, M. Persico, E. Virzi, M. Rescigno, G. Moro, and P. Ricciardi-Castagnoli. Inducible IL-2 production by dendritic cells revealed by global gene expression analysis. *Nat immunol*, 2(9):882–888, 2001.
35. R. Lang, D. Patel, J.J. Morris, R.L. Rutschman, and P.J. Murray. Shaping gene expression in activated and resting primary macrophages by IL-10. *J of Immunol*, 169(5):2253–2263, 2002.
36. R.L. McCaffrey, P. Fawcett, M. O'Riordan, K.D. Lee, E.A. Havell, P.O. Brown, and D.A. Portnoy. A specific gene expression program triggered by Gram-positive bacteria in the cytosol. *Proceedings of the National Academy of Sciences*, 101(31):11386–11391, 2004.
37. K. van Erp, K. Dach, I. Koch, J. Heesemann, and R. Hoffmann. Role of strain differences on host resistance and the transcriptional response of macrophages to infection with Yersinia enterocolitica. *Physiological Genomics*, 25(1):75, 2006.
38. J.T. Eppig, C.J. Bult, J.A. Kadin, J.E. Richardson, and J.A. Blake. The Mouse Genome Database (MGD): from genes to mice–a community resource for mouse biology. *Nucl Acids Res*, 33(Database Issue):D471, 2005.
39. S.Y. Chang, P.F. Su, and T.C. Lee. Ectopic expression of interleukin-1 receptor type ii enhances cell migration through activation of the pre-interleukin 1alpha pathway. *Cytokine*, 45(1):32–8, 2009.
40. D.L. Simmons, S. Tan, D.G. Tenen, A. Nicholson-Weller, and Seed B. Monocyte antigen cd14 is a phospholipid anchored membrane protein. *Blood*, 73(1):284–9, 1989.
41. P. Stumptner-Cuvelette, S. Morchoisne, M. Dugast, S. Le Gall, G. Raposo, O. Schwartz, and P. Benaroch. Hiv-1 nef impairs mhc class ii antigen presentation and surface expression. *Proc Natl Acad Sci U S A*, 98(21):12144–9, 2001.
42. D.C. Phua, P.O. Humbert, and W. Hunziker. Vimentin regulates scribble activity by protecting it from proteasomal degradation. *Mol Biol Cell*, 20(12):2841–55, 2009.
43. J. Ernst and Z. Bar-Joseph. Stem: a tool for the analysis of short time series gene expression data. *BMC Bioinformatics*, 7(191), 2006.
44. A. C. Davison and D. V. Hinkley. *Bootstrap Methods and Their Application (Cambridge Series in Statistical and Probabilistic Mathematics , No 1)*. Cambridge University Press, 1997.