Learning Gene Regulatory Networks via Globally Regularized Risk Minimization

Yuhong Guo

Joint work with Dale Schuurmans

Department of Computing Science
University of Alberta

September 16, 2007

Gene Regulatory Networks

Genes do not work independently

- Gene expressions are regulated (control the amount and timing of appearance of their functional products) to achieve proper cell function
- The regulation mechanism forms a network—the gene regulatory network

Identifying gene regulatory networks helps gain insight into biological function

Given availability of high-throughput microarray data

- mRNA expression levels of thousands of genes are measured simultaneously

Raises an important, challenging task in computational biology

- Learn gene regulatory networks from time-series gene expression data

Modeling Gene Regulatory Networks

Approaches proposed in the literature

- Linear models: linear differential equations [De Jong et al., 2004; Chen et al., 2005]; sparse linear modeling [De Hoon et al., 2003; Li et al., 2004]
- Boolean network models
- (Dynamic) Bayesian networks

- Prototype approach [Van Someren et al., 2000]

Motivation

Difficulty:

- a few time-points for a large number of genes
- identifying regulators for each gene separately is error prone

Biological assumption:

- genes with similar expression patterns are likely to be co-regulated

Idea

Idea:

- Identify common regulators for groups of genes with similar expression profiles while still permitting individual differences

Method: based on linear regression

- First, after rescaling the expression data into values between 0 and 1, cluster the genes using k-means
- For each cluster, identify the regulatory relationships using a novel combination of local and global feature selection (regularization)

Global Regularized Approach

Introduce the base linear regression model

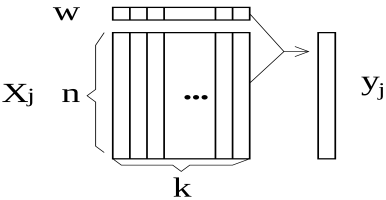
Then address the following three issues

- Coping with Time Lags
- Local feature selection – permit individual differences
- Global feature selection – regulation Sharing

Linear Modeling

- Given the time-series expression vector $\mathbf{y}_j : n \times 1$ for the jth target gene and the expression matrix $X_j : n \times k$ for its k candidate regulators
- How well that $\mathbf{y}_j$ can be predicted from X_j can be determined by solving a linear regression

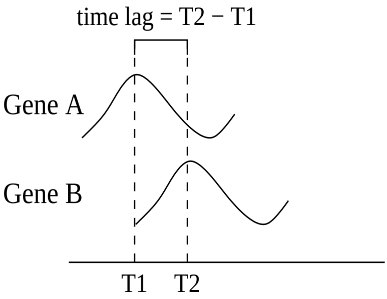
$$\min_{\mathbf{w}_j} \; \|X_j \mathbf{w}_j - \mathbf{y}_j\|_2^2$$



Time Lags

Problem

- Regulation does not occur instantaneously. There are potential time lags between the expression of a regulator and its downstream target genes.



Time Lags

Solution: time shifting

- For a target gene j, the optimal shift between $\mathbf{y}_j$ and the profile $\mathbf{x}_{ij}$ of its ith candidate regulator can be computed by aligning $\mathbf{x}_{ij}$ with $\mathbf{y}_j$

$$s_{ij}^* = \arg \min_{s \in \{0,1,2,3\}} \|\mathbf{x}_{ij}(1,...,n-s) - \mathbf{y}_j(s+1,...,n)\|_2^2$$

Coping with Time Lags

- Compute the maximum shift: $s_{max} = \max_i s_{ij}$
- Truncate $\mathbf{y}_j$ to obtain $\tilde{\mathbf{y}}_j = \mathbf{y}_j(s_{max}, ..., n)$
- Apply the optimal shift to each column of X_j, and truncate the columns to a common length based on s_{max}. Finally obtain a $(n - s_{max}) \times k$ time-lag aligned matrix Φ_j.
- The linear regression can then be written as

$$\min_{\mathbf{w}_j} \|\Phi_j \mathbf{w}_j - \tilde{\mathbf{y}}_j\|_2^2$$

Feature Selection

- **Issue:** the set of candidate regulators for a target gene is much larger than its true regulator set
- **Feature selection** need to be conducted to discard the irrelevant candidate regulators

$$\min_{\mathbf{w}_j} \quad \|\Phi_j \mathbf{w}_j - \tilde{\mathbf{y}}_j\|_2^2 + \alpha \|\mathbf{w}_j\|_1$$

Using L1 norm for regularization, many weights **w** (corresponding to irrelevant candidate regulators) would be set to 0

Key contribution

Tackle the problem of the lack of time points by sharing regulatory information across genes with similar expression profiles

- Introduce a set of 0-1 valued global feature selection variables $\boldsymbol{\eta} = \{\eta_1, ..., \eta_I\}^\top$, corresponding to the common candidate regulator set $X = \{\mathbf{x}_1, ..., \mathbf{x}_I\}$
- Globally regularized risk minimization:

$$\min_{\boldsymbol{\eta}\in\{0,1\}^I} \min_{\mathbf{w}} \sum_j \left(\|\Phi diag(\boldsymbol{\eta})\mathbf{w}_j - \tilde{\mathbf{y}}_j\|_2^2 + \alpha\|\mathbf{w}_j\|_1\right) + \lambda \mathbf{u}^\top \boldsymbol{\eta} \qquad (1)$$

where Φ is the aligned expression matrix for the candidate regulators of the genes in the considered cluster

Note that (1) has both global and local regularization terms

- the global regularization term $\lambda\mathbf{u}^\top\boldsymbol{\eta}$ is an L0 norm regularizer, aims to identify the common regulators for the cluster genes by sharing regulatory information (thus with more time points)
- the local L1 norm regularizer, $\alpha\|\mathbf{w}_j\|_1$, makes individual choices of regulators

Hope to achieve more accurate regulator identification

Optimization Procedure

- The min-min integer optimization problem (1) can be relaxed into

$$\begin{aligned} \min_{\boldsymbol{\eta}} \min_{\mathbf{w}} \quad & \sum_j \left(\|\Phi diag(\boldsymbol{\eta}) \mathbf{w}_j - \tilde{\mathbf{y}}_j\|_2^2 + \alpha \|\mathbf{w}_j\|_1 \right) + \lambda \mathbf{u}^\top \boldsymbol{\eta} \\ \text{subject to} \quad & 0 \leq \boldsymbol{\eta} \leq 1 \end{aligned}$$

- Conduct the optimization in two alternating steps:
  - $min_{\mathbf{w}}$: using quadratic programming or a fast grafting algorithm
  - $min_{\boldsymbol{\eta}}$: use a quasi-Newton BFGS method

Experiments

Conduct experiments to identify cell cycle regulation networks where the cell cycle genes are regulated by a set of transcription activators

Experimental design

- Compare the proposed global regularization approach to two extremes based on linear regression models:
  - local regularization approach: use only the local L1 norm regularizer to determine the regulators for each gene separately
  - prototype method: use only the global regularizer to identify the common regulators for the whole cluster

Synthetic Experiments

Goal

Simulate a cell cycle process controlled by a small number of critical transcription factors (TFs) to gauge the potential effectiveness of the proposed approach when the ground truth is known

Setup

Define a 4-phase cell cycle where 10 TFs regulate the expression levels of 212 genes (53 genes in each phase); 10 TFs are associated with the 4 phases with (3, 2, 3, 2) in each phase; each gene/TF is regulated by one TF or the combination of 2 TFs randomly selected from the TFs from the previous phase in the cycle

Synthetic Profile Generation

Data generation procedure

- Simulate the expression level for the TFs in a selected phase for two complete cell cycles (16 time steps)
- Generate the expression profiles for the genes/TFs in the next phase by a 2 time step delayed response (with Gaussian noise) from the profiles of randomly selected one or two TFs in the current phase
- Repeat this generating procedure for all phases in turn

Synthetic Results

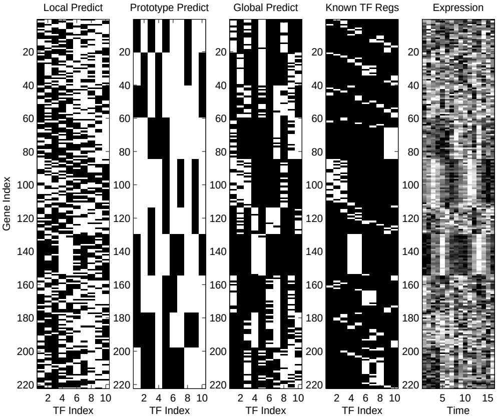


Figure: Rows denote target genes in the synthetic experiment. Columns denote candidate regulators (transcription factors).

Synthetic Results

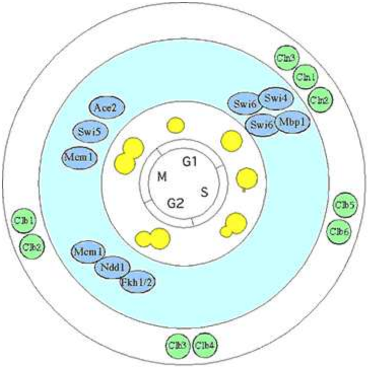
Results obtained using 10 clusters

Table: Results on synthetic data, predicting TF-based regulations.

Performance comparison	Local regularization	Prototype method	Global regularization
accuracy (%)	57.6	47.2	73.0
precision (%)	21.4	18.1	30.0
recall (%)	71.5	75.0	63.8
F-measure	33.0	29.2	40.8

Yeast Cell Cycle Gene Regulation

Yeast cell cycle gene expression is known to be regulated by nine cell cycle transcriptional activators: Mbp1, Swi4, Swi6, Mcm1, Fkh1, Fkh2, Ndd1, Swi5 and Ace2 [Simon et al., 2001]



from http://web.wi.mit.edu/young/cellcycle/

Yeast Cell Cycle Gene Regulation

- **Experiments:** Identify the 9 TFs based cell cycle regulatory network
- Conduct experiments on a subset of 267 cell cycle genes from Cho et al.'s data [Cho et al., 1998]
- Evaluate the performance on a subset of 127 genes for which
    - the confirmed regulatory information can be obtained from previous literature [Simon et al., 2001; Iyer et al., 2001]
    - or potential regulation relationships can be inferred from the existing binding data [Iyer et al., 2001]

Yeast Cell Cycle Results

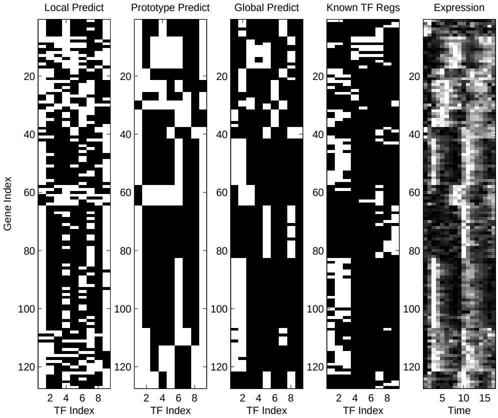


Figure: Rows denote target genes. Columns denote candidate regulators (transcription factors).

Yeast Cell Cycle Results

Results obtained using 15 clusters

Table: Results on the subset of the real yeast cell cycle gene expression data, restricted to genes where TF-based regulation information is known or can be inferred from other sources.

Performance comparison	Local regularization	Prototype method	Global regularization
accuracy (%)	57.8	55.4	73.9
precision (%)	22.3	21.2	35.7
recall (%)	47.5	48.0	43.4
F-measure	30.4	29.4	39.2

Conclusion & Future Work

Conclusion

By sharing regulation information across genes with similar expression profiles, more time points can be used to predict the common regulators, which leads to improved prediction quality

Future Work

- Consider incorporating other sources of biologically relevant data, or other prior knowledge into network induction
- Extend this feature selection strategy to solve other feature selection problems in bioinformatics

Thanks!