CaSPIAN: A Causal Compressive Sensing Algorithm for Inference of Gene Interactions

Amin Emad, Mo Deng and Olgica Milenkovic

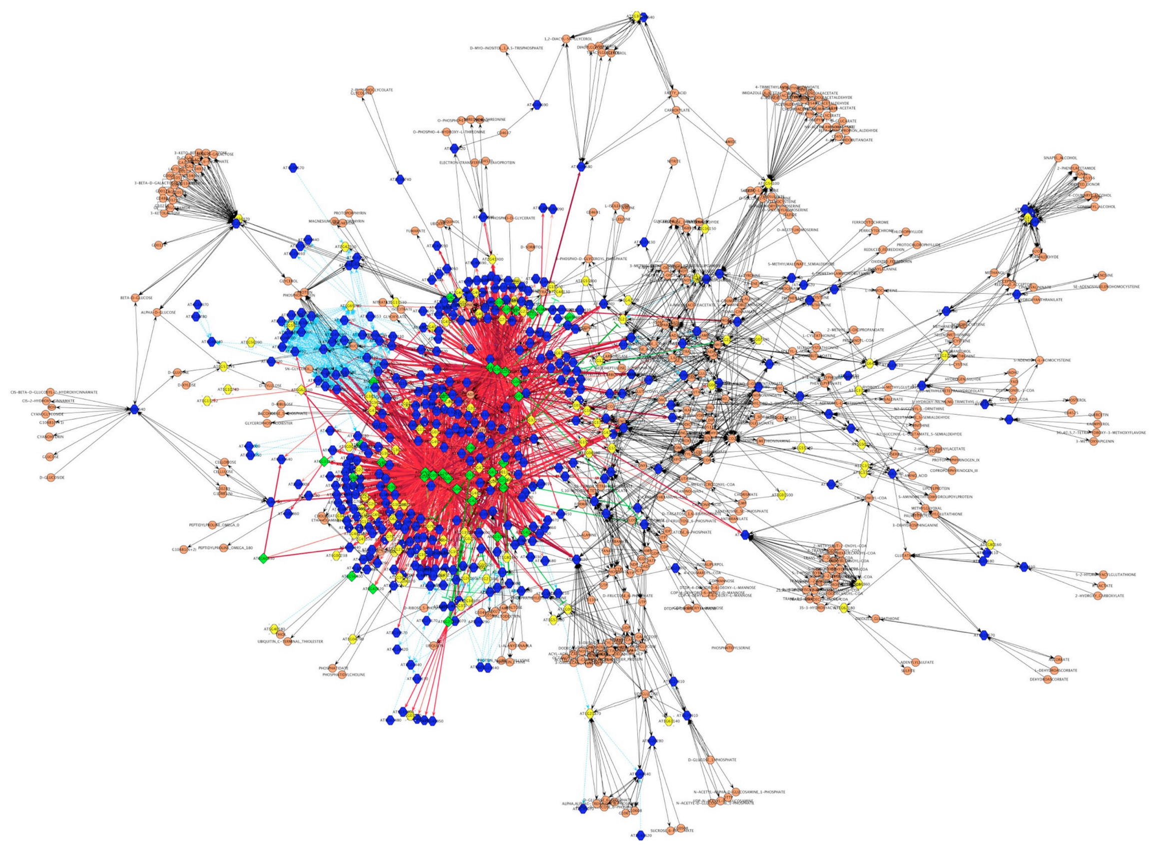
ECE, University of Illinois at Urbana-Champaign (UIUC)

emad2@illinois.edu

Motivation

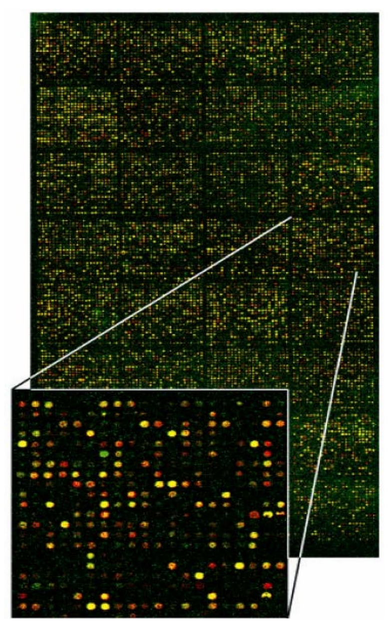
➢ Gene Regulatory Networks

- Genes "interact" with each other through their RNA or protein products to control the cell cycle
- Interactions occur via transcription factors which act as inhibitors or promoters



Gene Regulatory Network (source: www.bioquicknews.com)

- DNA microarrays measure the activity (expression level) of thousands of genes simultaneously
- Expression levels of genes are usually measured in terms of RNA product concentration



DNA Microarray (source: www.geneticsscience.blogspot.com)

➢ Sparsity in the network

- Most genes are affected by only "a few" other genes [Thieffry, 1998]

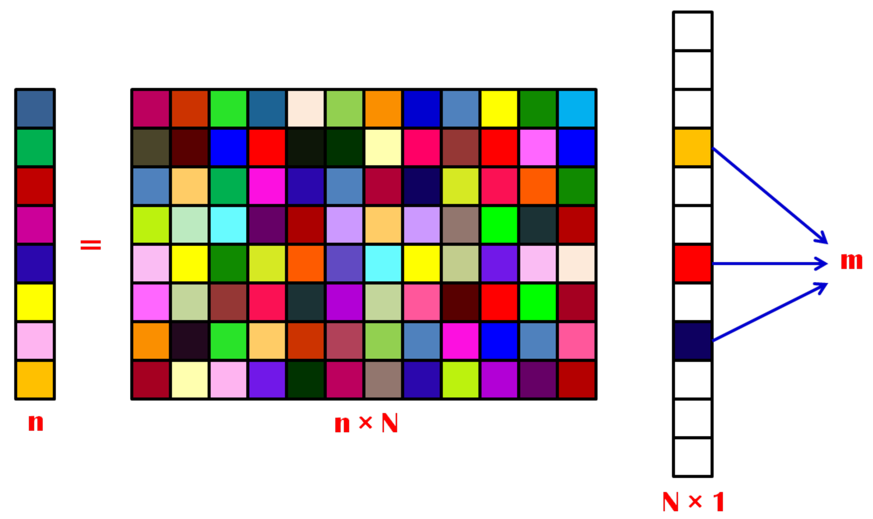
➢ Causal relationships among genes

- Genes affect each other causally
- Which are the genes that influence a specific gene?

Compressive Sensing (CS)

➢ Model

- In the absence of noise, vector $\mathbf{y} \in \mathbb{R}^n$ can be represented as a linear combination of m columns of $\mathbf{\Phi} \in \mathbb{R}^{n \times N}$, i.e. $\mathbf{y} = \mathbf{\Phi}\mathbf{x}$.



➢ Employs sparsity for recovery

- Sparsity: $m \ll N$
- Minimize the l_1 norm of $\mathbf{x} \in \mathbb{R}^N$

$$\min |\mathbf{x}|_1 \quad \text{s.t.} \quad |\mathbf{y} - \mathbf{\Phi}\mathbf{x}|_2^2 < \epsilon$$

- Recovery possible for $n = O(m \log(N/m))$
- $\mathbf{\Phi}$ should satisfy RIP

➢ Greedy Algorithms

- OMP [Tropp, 2004], SP [Dai & Milenkovic, 2008], IHT [Blumensath and Davis, 2009], etc.

Causality

➢ Granger Causality

- Let $y(t)$ be a stationary time series with a linear autoregressive (AR) prediction of the form

$$y(t) = \sum_{i=1}^{d} a_i y(t - iT) + r^{(1)}$$

- Let $x(t)$ be another stationary time series. The optimal linear AR prediction of $y(t)$ based on its past values and $x(t)$ is

$$y(t) = \sum_{i=1}^{d} a'_i y(t - iT) + \sum_{i=1}^{d} b_i x(t - iT) + r^{(2)}$$

- The time series $x(t)$ Granger-cause $y(t)$ if significantly

$$VAR\left(r^{(2)}\right) < VAR\left(r^{(1)}\right)$$

- We use an F-test do determine the significance

Method

Algorithm: CaSPIAN II

Input: $i \in \{1, 2, \cdots, N\}$, $\{\phi_{j,l}\}$, $k \in \mathbb{N}$, and P_F

Output: $\mathcal{S}_i \subset \mathcal{G}$

Initialization:

- Set $\mathbf{y} = \phi_{i,0}$, $\mathcal{S}_i^{(0)} = \varnothing$, $\mathcal{T}_i^{(0)} = \varnothing$, and form $\mathbf{\Phi}_{\mathcal{G}\setminus\{g_i\}}$

For $\kappa = 1, 2, \cdots, k$ **do**

- Run SP for vector $\mathbf{y}$, sensing matrix $\mathbf{\Phi}_{\mathcal{G}\setminus\{g_i\}}$, and sparsity κ
- $\mathcal{T}_i^{(\kappa)} = \mathcal{T}_i^{(\kappa-1)} \cup \{\kappa$ columns of $\mathbf{\Phi}_{\mathcal{G}\setminus\{g_i\}}$ recovered using SP$\}$

End

- Form $\mathcal{R}_i = \{g_{i_1}, g_{i_2}, \cdots\}$ as the set of genes corresponding to the columns in $\mathcal{T}_i^{(k)}$
- Form $\mathbf{\Phi}_{\mathcal{T}_i}$ using $\mathcal{T}_i^{(k)}$ and set $\mathbf{r}_i = \mathbf{y} - \mathbf{\Phi}_{\mathcal{T}_i}\mathbf{\Phi}_{\mathcal{T}_i}^{\dagger}\mathbf{y}$

For $j = 1, 2, \cdots, |\mathcal{R}_i|$ **do**

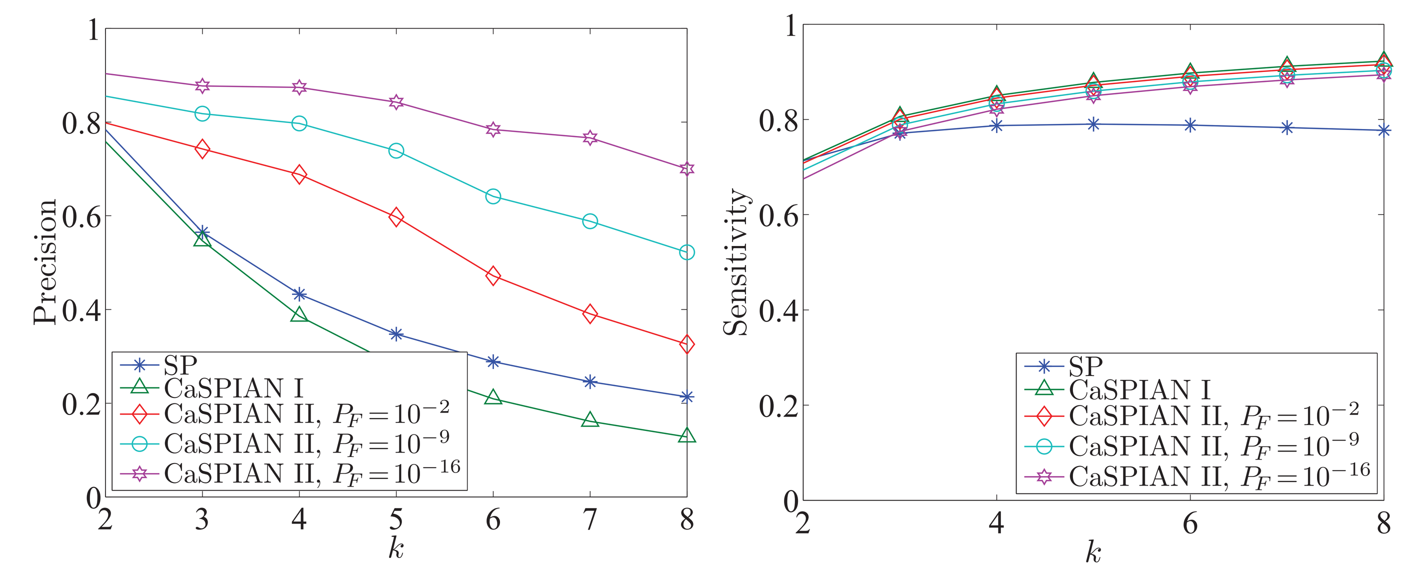
- Form $\mathbf{\Phi}_{\mathcal{T}_{i,j}}$ and calculate $\mathbf{r}_{i,j} = \mathbf{y} - \mathbf{\Phi}_{\mathcal{T}_{i,j}}\mathbf{\Phi}_{\mathcal{T}_{i,j}}^{\dagger}\mathbf{y}$
- Form the F-statistic using $||\mathbf{r}_i||$ and $||\mathbf{r}_{i,j}||$

  If F-statistic is greater than critical value corresponding to P_F
  Then $\mathcal{S}_i^{(j)} = \mathcal{S}_i^{(j-1)} \cup \{g_{i_j}\}$
  Else $\mathcal{S}_i^{(j)} = \mathcal{S}_i^{(j-1)}$

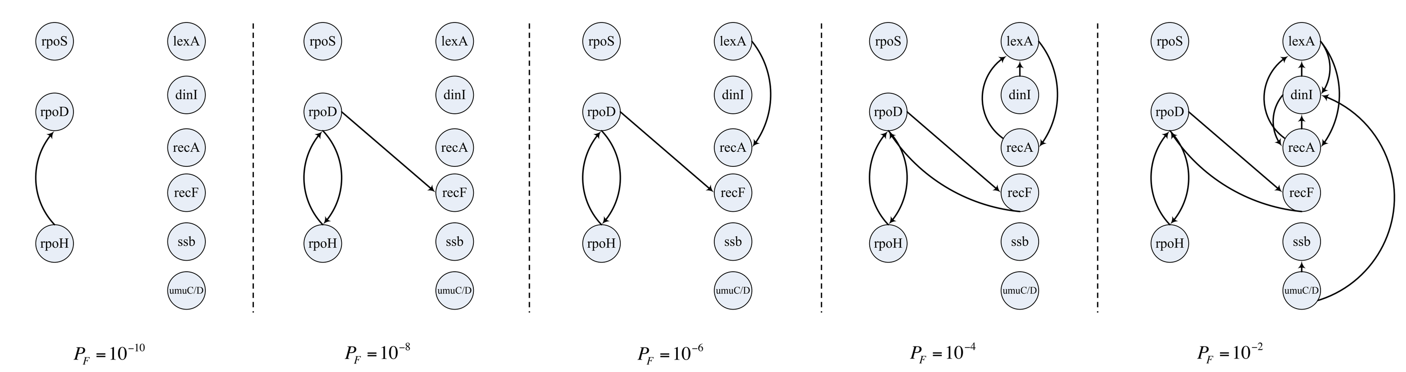
End

Return $\mathcal{S}_i = \mathcal{S}_i^{(|\mathcal{R}_i|)}$

Results



Synthetic network with n=72, N=100, max sparsity 5
Precision = TP/(TP+FP), Sensitivity = TP/(TP+FN)



E. Coli SOS network, 22 experiments with at least 3 time-points, M3D dataset
CaSPIAN II (12 TP and 0 FP), LASSO (0 TP, 0FP),
Truncated LASSO [Shojaei & Michailidis, 2010] (1 TP, 0FP)