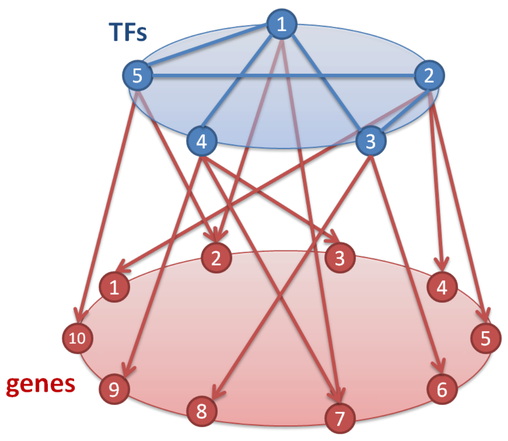
Predicting Gene Expression from TF Expression Reveals TF-TF Interactions in *E. coli*

Francisco Sánchez-Vega, Laurent Younes and Donald Geman

Applied Mathematics and Statistics Department and Center for Imaging Sciences.
Johns Hopkins University, Baltimore, MD.

Introduction

- TF-gene interactions can be used to learn TF-TF interactions.
- **Assumption:** if TF 'A' regulates TF 'B', then it is more likely that TF 'A' and TF 'B' will regulate similar sets of non-TF genes.
- **Idea:** Represent each TF as a vector of TF-gene interactions.



	TF_1	TF_2	TF_3	TF_4	TF_5
g_1	0	1	0	0	0
g_2	1	0	0	0	1
g_3	0	0	0	1	0
g_4	0	1	0	0	0
g_5	0	1	0	0	0
g_6	0	0	1	0	0
g_7	1	0	0	1	0
g_8	0	0	1	0	0
g_9	0	0	0	1	0
g_{10}	0	0	0	0	1

- When TF-gene edges are not known in advance, these vectors can be *estimated* by regressing gene expression on TF expression.

Learning TF-TF Interactions from TF-Gene Interactions

- Many algorithms for learning gene regulatory networks (such as relevance networks [1], ARACNE [2] and CLR [3]) compute pairwise measurements of similarity between random variables (typically, mutual information).

d variables (TFs)

X_1	X_2	...	X_d
$x_1^{(1)}$	$x_2^{(1)}$	...	$x_d^{(1)}$
$x_1^{(2)}$	$x_2^{(2)}$	...	$x_d^{(2)}$
...	...	...	...
$x_1^{(n)}$	$x_2^{(n)}$	...	$x_d^{(n)}$

n samples

- The usual approach consists in working with a matrix of microarray data where columns correspond to TFs and rows correspond to different samples.

$$\mathcal{L} = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}\}, \forall i, \mathbf{x}^{(i)} \in \mathbb{R}^d$$

d variables (TFs)

Y_1	Y_2	...	Y_d
$y_1^{(1)}$	$y_2^{(1)}$	...	$y_d^{(1)}$
$y_1^{(2)}$	$y_2^{(2)}$	...	$y_d^{(2)}$
...	...	...	...
$y_1^{(m)}$	$y_2^{(m)}$	...	$y_d^{(m)}$

m target genes

- When TF-gene interactions are known, we propose to use an alternative data matrix where $\boldsymbol{Y_i^{(j)} = 1}$ if TF $\boldsymbol{i}$ regulates gene $\boldsymbol{j}$, and $\boldsymbol{Y_i^{(j)} = 0}$ otherwise.

$$\mathcal{L}^* = \{\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(m)}\}, \forall i, \mathbf{y}^{(i)} \in \{0,1\}^d$$

d variables (TFs)

$\hat{Y}_1$	$\hat{Y}_2$	...	$\hat{Y}_d$
$\lvert\beta_1^{(1)}\rvert$	$\lvert\beta_2^{(1)}\rvert$	...	$\lvert\beta_d^{(1)}\rvert$
$\lvert\beta_1^{(2)}\rvert$	$\lvert\beta_2^{(2)}\rvert$	...	$\lvert\beta_d^{(2)}\rvert$
...	...	...	...
$\lvert\beta_1^{(m)}\rvert$	$\lvert\beta_2^{(m)}\rvert$	...	$\lvert\beta_d^{(m)}\rvert$

m target genes

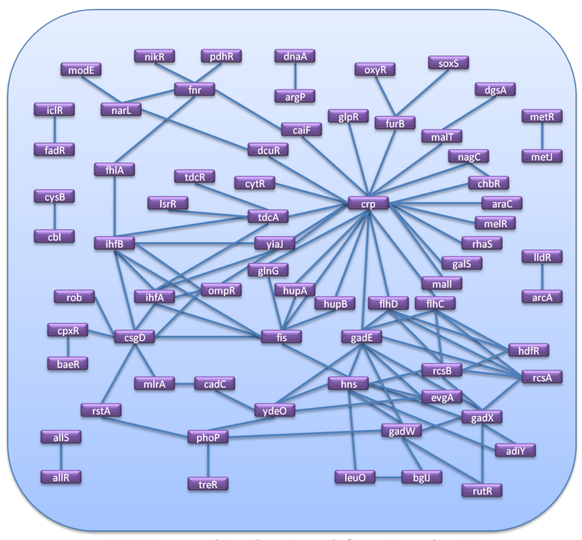
- When TF-gene interactions are unknown, they can be estimated from microarray data. For each gene target $\boldsymbol{t}$, solve

$$\min_{\boldsymbol{\beta}^{(t)}} ||\boldsymbol{\mathcal{X}} \cdot \boldsymbol{\beta}^{(t)} - \mathbf{x}^{(t)}||_2$$

where $\mathbf{x}^{(t)} \in \mathbb{R}^n$ is the expression data for gene $\boldsymbol{t}$, $\boldsymbol{\mathcal{X}} \in \mathbb{M}_{n \times d}$ is the matrix of TF expression and $\boldsymbol{\beta}^{(t)} \in \mathbb{R}^d$.

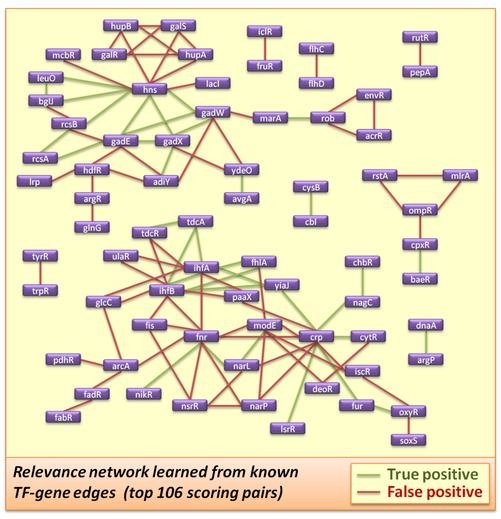
Transcriptional Networks in *E. coli*

- Ground truth network from RegulonDB database [4], which contains 106 TF-TF interactions and 2,109 TF-gene interactions involving $\boldsymbol{d = 126}$ TFs and $\boldsymbol{m = 984}$ genes.
- Expression data from the Many Microbe Microarrays Database (M3D) [5], which contains $\boldsymbol{n = 466}$ microarray samples.



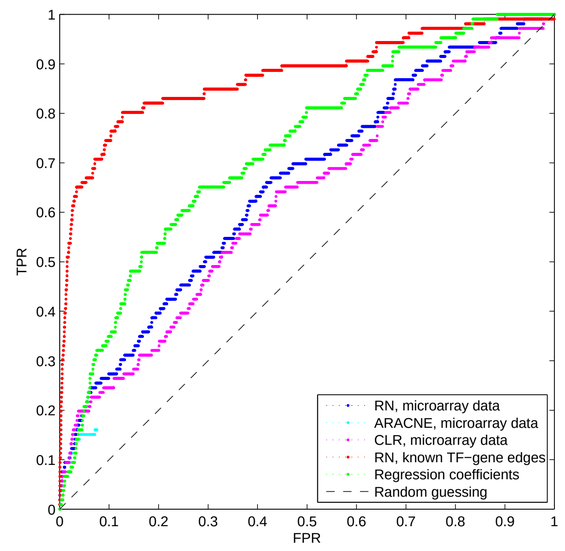
TF-TF ground truth network from RegulonDB

- We measured ROC performance for edgewise network reconstruction accuracy using the three types of data matrix representation.
- For the regression coefficients case, we simply ranked all pairwise Euclidean distances between columns.



Relevance network learned from known TF-gene edges (top 106 scoring pairs)

True positive
False positive



Discussion

- New approach intended to improve (not replace) existing TF-TF network reconstruction algorithms.
- By estimating TF-gene edges, we look jointly at microarray data for TFs and non-TF target genes (as opposed to alternatives that learn TF-TF networks using TF expression alone).
- Non-penalized linear regression was used only for illustration purposes. Sparse regression techniques may lead to better results, possibly closing the gap between the green and red ROC curves.
- Basis for two-phase network learning strategy: first, learn TF-gene edges using regression and then learn TF-TF edges as graphical models using the vectors of regression coefficients.

References

[1] Butte, A. J. and Kohane, I. S.: "Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements," *Pac. Symp. Biocomputing*, 2000.

[2] Margolin, A. et al.: "ARACNE: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context," *BMC Bioinformatics*, vol. 7, 2006.

[3] Faith, J. J. et al.: "Large-Scale mapping and validation of *Escherichia coli* transcriptional regulation from a compendium of expression profiles," *PLoS Biol.*, vol. 5, n. 1, 2007.

[4] Gama-Castro S. et al.: "RegulonDB (version 7.0): Transcriptional regulation of *Escherichia coli* K-12 integrated within genetic sensory response units (gensor units)," *Nucleic Acids Research*, 2010.

[5] Faith, J. J. et al.: "Many Microbe Microarrays Database: uniformly normalized Affymetrix compendia with structured experimental metadata," *Nucleic Acids Research*, 2008.

Please send correspondence to Francisco Sánchez-Vega (sanchez@cis.jhu.edu) at the Center for Imaging Sciences, 307C Clark Hall. The Johns Hopkins University. 3400 N. Charles Street, Baltimore, MD, 21218-2686, USA. http://www.cis.jhu.edu

Acknowledgments: Francisco Sánchez-Vega was supported by a travel grant from the Acheson J. Duncan Fund for the Advancement of Research in Statistics at Johns Hopkins. The work of Donald Geman is partially supported by NIH-NCRR Grant UL1 RR 025005.