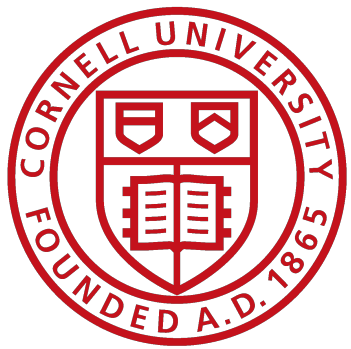
A Bayesian Dynamical Systems Approach to Clustering Gene Expression Time Series Data

Sara Venkatraman

Cornell University, Statistics and Data Science

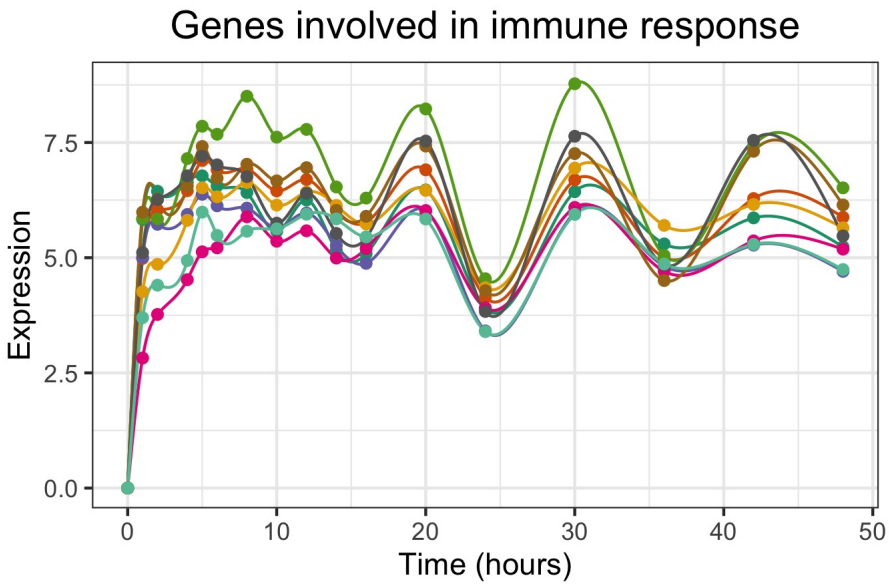


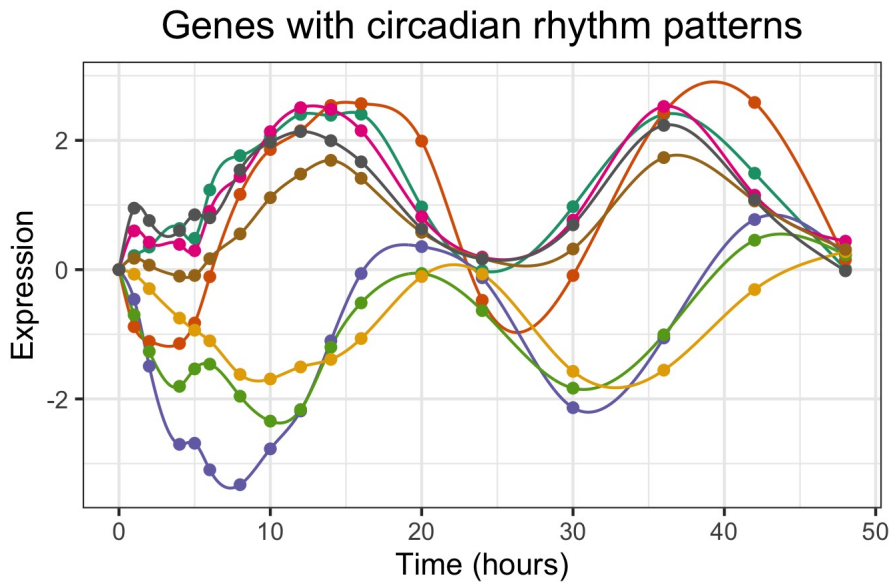
Joint work with Sumanta Basu, Andrew G. Clark,
Sofie Delbare, Myung Hee Lee, Martin T. Wells

Introduction

Time-course gene expression datasets measure expressions of thousands of genes at a few time points.

Statistical task: want to find clusters/networks of genes with similar time dynamics (either co-moving or lead-lag)





Challenges: complex time dynamics, data is high-dimensional

Challenges in cluster analysis of gene expression

How to measure "similarity" in two genes' expressions?

Idea: Derive similarity metrics from ODEs that model co-movement/lagged relationships in gene expression over time

How to find similar gene pairs within thousands of genes?

Idea: Encourage high similarity scores between genes that are known to be associated, according to prior biological information (obtained from public databases)

Gene expression as a dynamical system

How does a gene's expression vary over time?

Let $m_A(t)$ = expression of gene A at time t. Possible model:

$$\frac{\mathrm{d}m_A(t)}{\mathrm{d}t} = p(t) - \kappa_A m_A(t),$$

where $p(t)$ = some regulatory signal, κ_A = degradation rate.

[Farina et al., 2007]

How do two associated genes A and B vary over time?

$$\frac{\mathrm{d}m_A(t)}{\mathrm{d}t} = (\alpha_A\, p(t) + \beta_A) - \kappa_A m_A(t),$$
$$\frac{\mathrm{d}m_B(t)}{\mathrm{d}t} = (\alpha_B\, p(t) + \beta_B) - \kappa_B m_B(t).$$

Gene expression as a dynamical system

Rearrange/integrate ODEs to get gene A's expression in terms of B's:

$$m_A(t) = c_1 m_B(t) + c_2 \int_0^t m_B(s)\,\mathrm{d}s + c_3 \int_0^t m_A(s)\,\mathrm{d}s + c_4 t + c_5.$$

This is linear in the coefficients $c_1, \ldots, c_5$

(which are composed from parameters $\alpha_A, \alpha_B, \beta_A, \beta_B, \kappa_A, \kappa_B$).

Therefore:

- We can fit this model to time-series data $\{m_A(t_i)\}_{i=1}^n$, $\{m_B(t_i)\}_{i=1}^n$ using **linear regression**
- Then, we can use the R^2 to measure association between the temporal expressions of genes A, B

Fitting dynamical models to data

Given time-series data $\{m_A(t_i)\}_{i=1}^n$, $\{m_B(t_i)\}_{i=1}^n$, we express our model

$$m_A(t) = c_1 m_B(t) + c_2 \int_0^t m_B(s)\,\mathrm{d}s + c_3 \int_0^t m_A(s)\,\mathrm{d}s + c_4 t + c_5$$

as $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\beta} = [c_1, ..., c_5]^T$ and $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$, with:

$$\mathbf{Y} = \begin{bmatrix} m_A(t_1) \\ ... \\ m_A(t_n) \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} m_B(t_1) & \int_0^{t_1} m_B(s) & \int_0^{t_1} m_A(s) & t_1 & 1 \\ ... & ... & ... & ... & ... \\ m_B(t_n) & \int_0^{t_n} m_B(s) & \int_0^{t_n} m_A(s) & t_n & 1 \end{bmatrix}$$

Then calculate: $R^2 = \underset{\text{explained by model above}}{\text{Fraction of variance in } m_A(t)} = \dfrac{\|\mathbf{X}\hat{\boldsymbol{\beta}} - \bar{Y}\mathbb{1}_n\|^2}{\|\mathbf{Y} - \bar{Y}\mathbb{1}_n\|^2}$

where $\hat{\boldsymbol{\beta}}$ = least-squares estimate of $\boldsymbol{\beta}$, and $\bar{Y}$ = mean of $\mathbf{Y}$.

Measuring similarity in time dynamics of two genes

We'll call this R^2 the lead-lag R^2.

- Measures association in temporal patterns of genes A, B
- But: does not account for prior knowledge about their relationship

Our contribution: Use empirical Bayesian regression to incorporate prior biological information into lead-lag R^2

("empirical" because hyperparameters will be chosen in a data-driven way).

Sources of biological information: pathway databases (e.g., GO, KEGG, STRING), protein-protein interaction networks

Background on Bayesian regression

Consider the linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$:

- $\mathbf{X} \in \mathbb{R}^{n \times p}$ and $\mathbf{Y} \in \mathbb{R}^{n \times 1}$ are observed, $\boldsymbol{\beta} \in \mathbb{R}^{p}$ is unknown
- Assume $\boldsymbol{\varepsilon}$ are i.i.d. normal errors: $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$

Approaches to estimating $\boldsymbol{\beta}$:

- Frequentist approach: Use the ordinary least-squares estimate $\hat{\boldsymbol{\beta}}_{\text{OLS}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$
- Bayesian approach: Choose prior probability distributions $p(\sigma^2)$ and $p(\boldsymbol{\beta}|\sigma^2)$.
  - Combine $p(\mathbf{Y}|\boldsymbol{\beta}, \sigma^2)$, $p(\boldsymbol{\beta}|\sigma^2)$, and $p(\sigma^2)$ via Bayes' theorem to get posterior distribution of $\boldsymbol{\beta}$
  - Can use mean of posterior distribution as an estimate of $\boldsymbol{\beta}$

Which prior distributions should we use?

The "normal-inverse gamma" prior is a common conjugate prior:

- Choose $p(\boldsymbol{\beta}|\sigma^2)$ to be the $N(\boldsymbol{\beta}_0, \sigma^2\mathbf{V}_0)$ distribution for some $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ and p.s.d. matrix $\mathbf{V}_0$
- Choose $p(\sigma^2)$ to be the $\Gamma^{-1}(a, b)$ distribution for $a,\ b > 0$

If we choose $\mathbf{V}_0 = g(\mathbf{X}^T\mathbf{X})^{-1}$, for some $g > 0$. Then:

$$\mathbb{E}(\boldsymbol{\beta}|\mathbf{Y}) = \frac{1}{1+g}\boldsymbol{\beta}_0 + \frac{g}{1+g}\hat{\boldsymbol{\beta}}_{\text{OLS}}$$

This is "Zellner's g-prior".

Soon we'll see how to choose $\boldsymbol{\beta}_0$ for our gene clustering problem.

(Hint: this will be where we can incorporate prior information about the genes!)

Our Bayesian regression methodology

Given a dataset of N genes measured at T time points,

1. Define a $N \times N$ prior "adjacency matrix" $\mathbf{W}$:

$$\mathbf{W}_{ij} = \begin{cases} 1 & \text{if genes } i, j \text{ have known association} \\ \text{NA} & \text{if genes } i, j \text{ have unknown relationship} \\ 0 & \text{if genes } i, j \text{ are unlikely to be associated} \end{cases}$$

2. For each gene pair, use Bayesian regression to fit the model $m_A(t) = c_1 m_B(t) + c_2 \int_0^t m_B(s) + c_3 \int_0^t m_A(s) + c_4 t + c_5$:
   - Use $\mathbf{W}$ to set mean of prior distribution on $\beta = [c_1, ..., c_5]$: $\beta_0 = [1, 1, 0, 0, 0]$ if $\mathbf{W}_{ij} = 1$, or all 0 otherwise.

     Why: first two parameters of β link expressions of genes A, B.
   - Compute posterior mean of β, and then the lead-lag R^2.

Data-driven tuning parameter selection

Recall the posterior mean of $\boldsymbol{\beta}$ was: $\dfrac{1}{1+g}\boldsymbol{\beta}_0 + \dfrac{g}{1+g}\hat{\boldsymbol{\beta}}_{\text{OLS}}$.

How do we choose g?

- No solutions to $g_* = \text{argmin}_{g>0} \|\mathbf{Y} - \hat{\mathbf{Y}}\|^2$ (sum of squared residuals), where $\hat{\mathbf{Y}} = \mathbf{X}\boldsymbol{\beta}_*$ and $\boldsymbol{\beta}_* = \mathbb{E}(\boldsymbol{\beta}|\mathbf{Y})$
- Instead, choose g to minimize Stein's unbiased risk estimate (unbiased estimate of $\|\hat{\mathbf{Y}} - \mathbf{X}\boldsymbol{\beta}\|^2$).

Theorem

Stein's unbiased risk estimate is minimized by:

$$g_* = \frac{\|\hat{\mathbf{Y}}_{\text{OLS}} - \mathbf{X}\boldsymbol{\beta}_0\|^2 - p\hat{\sigma}^2}{p\hat{\sigma}^2},$$

where $\hat{\mathbf{Y}}_{\text{OLS}} = \mathbf{X}\hat{\boldsymbol{\beta}}_{\text{OLS}}$, $\hat{\sigma}^2 = \frac{\|\mathbf{Y} - \hat{\mathbf{Y}}_{\text{OLS}}\|^2}{n-p}$, and n, p are dims. of $\mathbf{X}$.

R^2 for Bayesian regression models

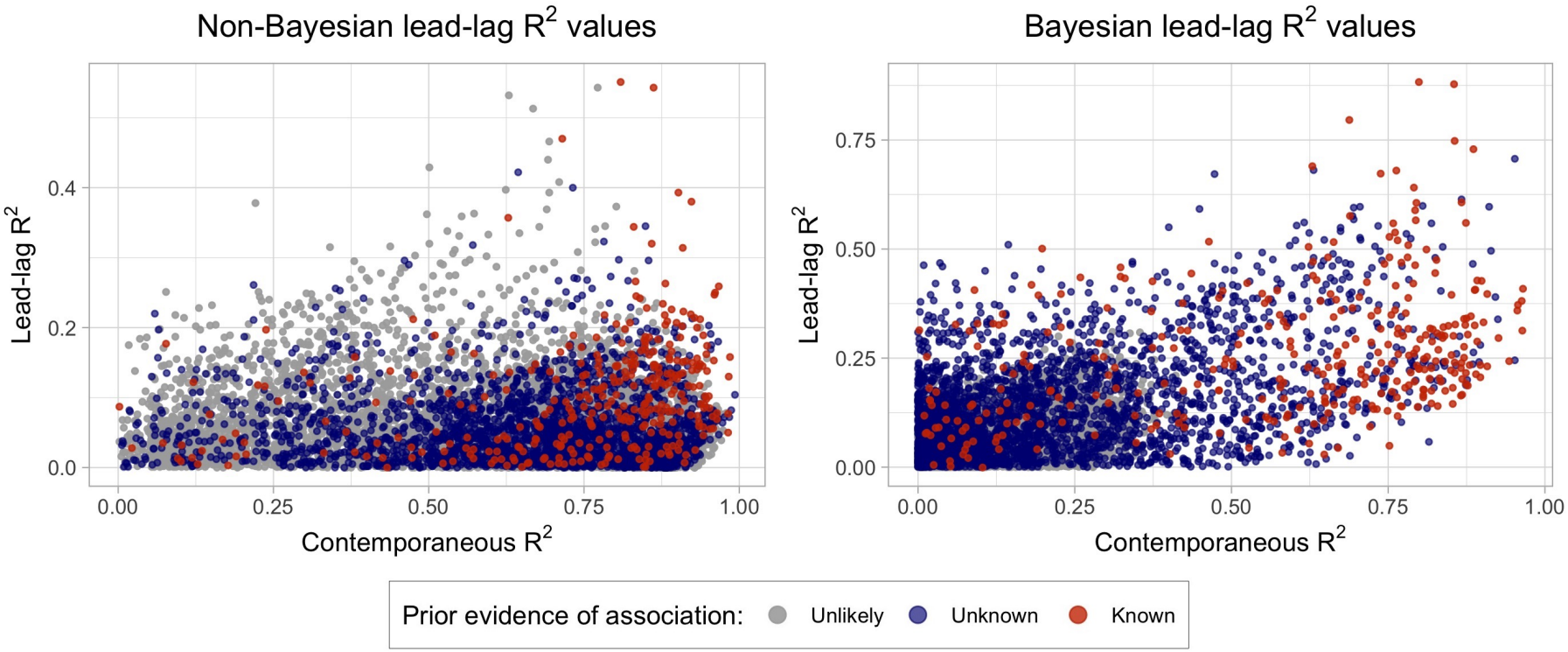
Classical definition of R^2 for ordinary least-squares may yield $R^2 > 1$ for Bayesian regression.

Instead, we define:

$$R^2 = \frac{\widehat{\mathrm{Var}}(\mathbf{X}\boldsymbol{\beta}_*)}{\widehat{\mathrm{Var}}(\mathbf{X}\boldsymbol{\beta}_*) + \widehat{\mathrm{Var}}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}_*)},$$

which we call the Bayesian lead-lag R^2 between genes A and B, where $\boldsymbol{\beta}_* = \frac{1}{1+g}\boldsymbol{\beta}_0 + \frac{g}{1+g}\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}$ is the posterior mean of $\boldsymbol{\beta}$.

Outline of empirical results



Dataset: expressions of 1735 genes in fruit flies at 21 time points, immediately following an induced immune response.

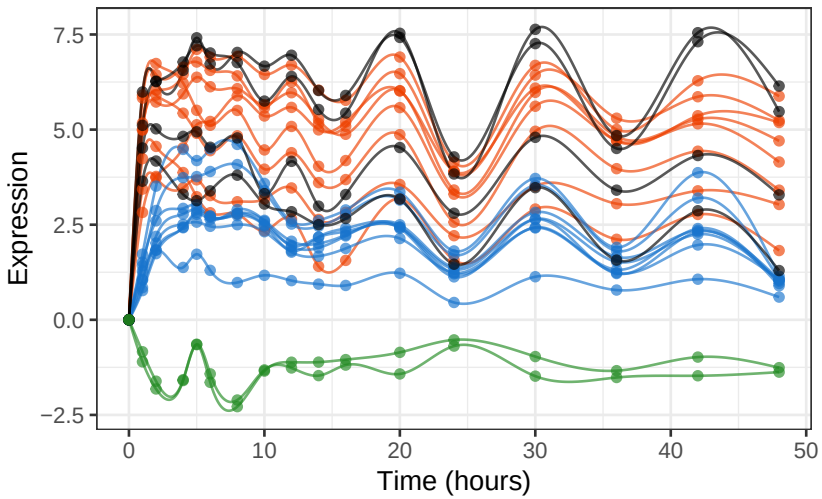
Method successfully identifies:

- Metabolism-immunity tradeoff found in previous studies
- Known groups of circadian rhythm, metabolic, immune response genes
- Novel interactions between orphan genes and known pathways

Hierarchical clustering on Bayesian lead-lag R^2 similarity matrix

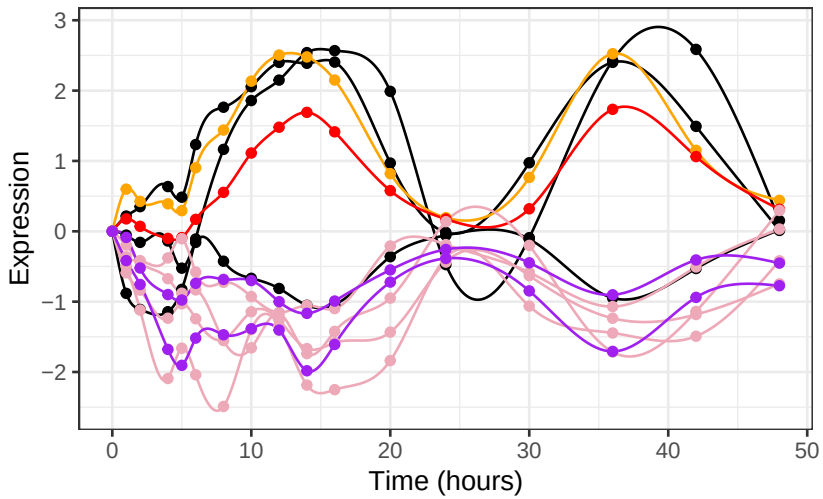
Genes involved in immune response

(Imd−regulated genes; Toll−regulated genes; cuticle proteins; genes potentially associated with Imd signaling)



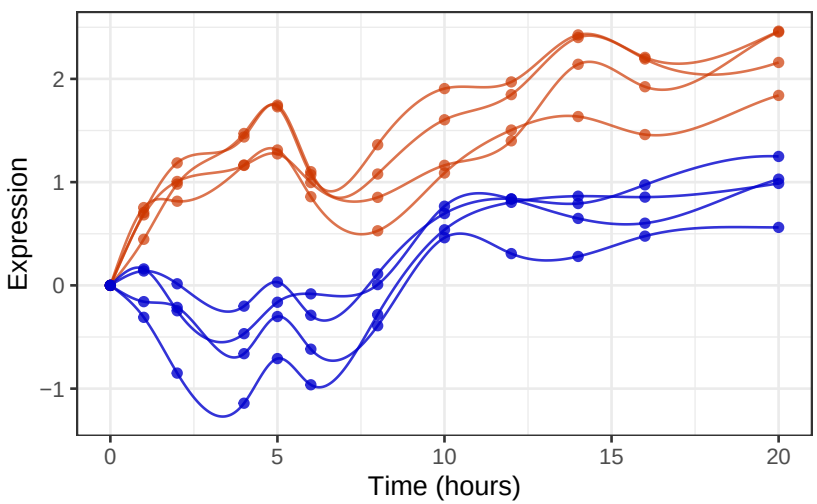
Genes exhibiting circadian rhythms

(Regulators of circadian clock: *tim, per, Clk, vri, Pdp1*; cuticle proteins; genes involved in dopamine synthesis)



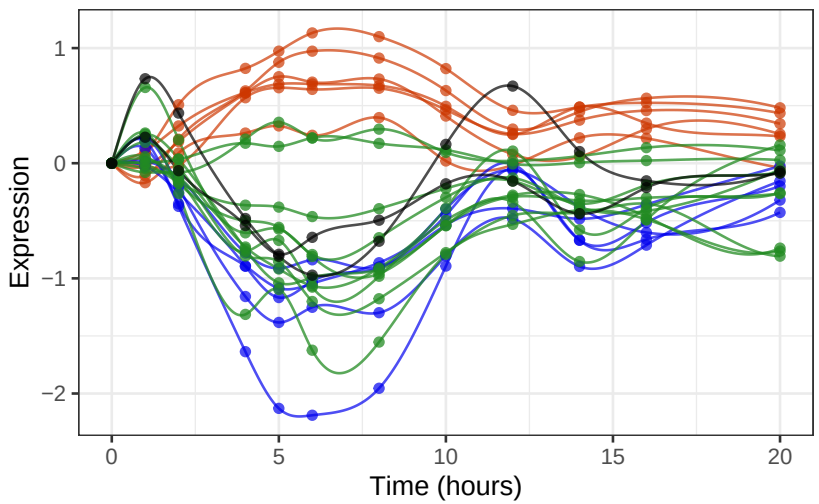
Genes involved in metabolism and immune response

(maltases; genes expressed in hemocytes)

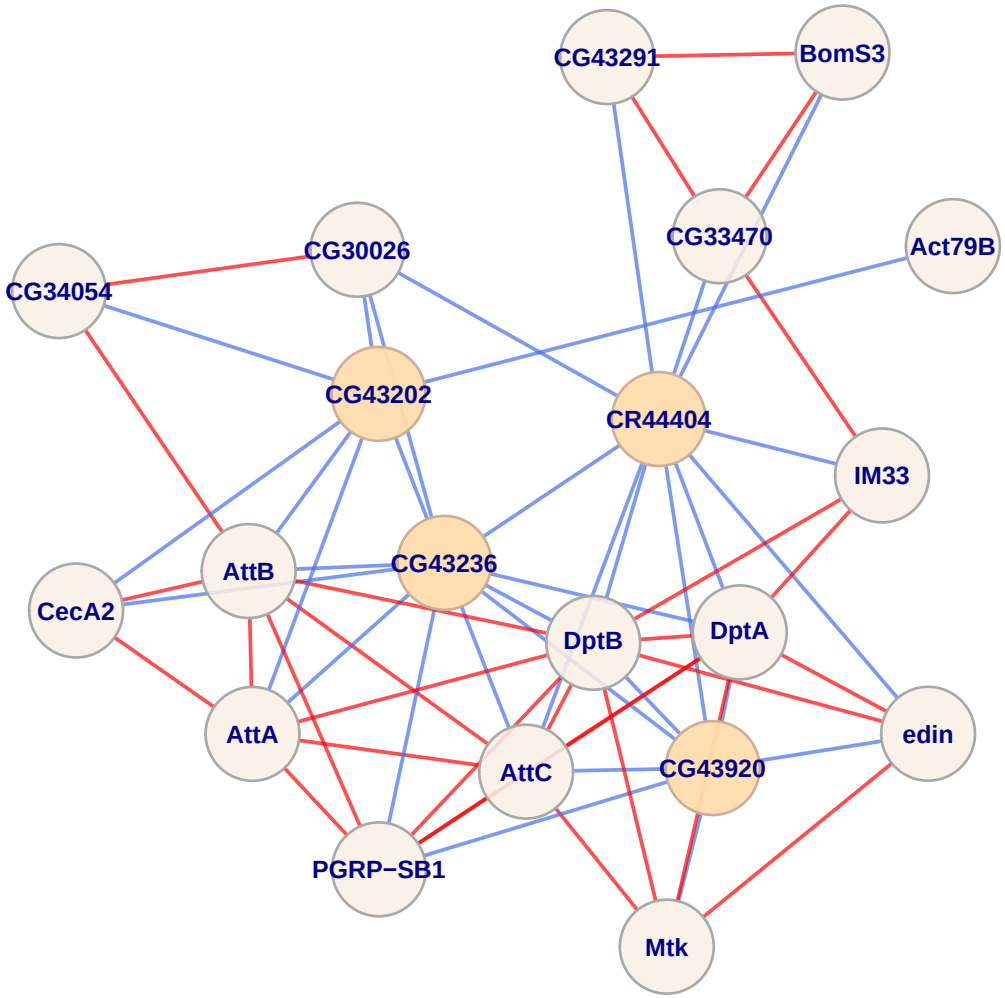


Genes involved in metabolic processes

(ribosome biogenesis; lipid catabolism; fatty acid biosynthesis; genes with uncharacterized relationships to *FASN1*)



Network reconstruction



Edge drawn between two genes if their Bayesian lead-lag $R^2 > 0.9$.

Red edges: previously known associations. Blue edges: previously unknown.

Thank you!

Sara Venkatraman

skv24@cornell.edu

https://sara-venkatraman.github.io

Appendix: Stein's unbiased risk estimate for linear models

Theorem [Fourdrinier, Strawderman, Wells 2018]

Let $\mathbf{Y} \sim N(\mathbf{X}\beta, \sigma^2 \mathbf{I}_n)$ where $\mathbf{X} \in \mathbb{R}^{n \times p}$. Let β_* be a weakly-differentiable function of the least-squares estimator $\hat{\beta}_{\text{OLS}}$ such that $\hat{\mathbf{Y}} = \mathbf{X}\beta_* = \mathbf{a} + \mathbf{S}\mathbf{Y}$ for some vector $\mathbf{a}$ and matrix $\mathbf{S}$. Then

$$\delta_0(\mathbf{Y}) = \|\mathbf{Y} - \mathbf{X}\beta_*\|^2 + (2\text{Tr}(\mathbf{S}) - n)\hat{\sigma}^2$$

is an unbiased estimator of $\|\hat{\mathbf{Y}} - \mathbf{X}\beta\|^2$, where $\hat{\sigma}^2 = \frac{\|\mathbf{Y} - \mathbf{X}\hat{\beta}_{\text{OLS}}\|^2}{n-p}$.

In this context, $\beta_* = \mathbb{E}(\beta|\mathbf{Y}) = \frac{1}{1+g}\beta_0 + \frac{g}{1+g}\hat{\beta}_{\text{OLS}}$:

- Then $\hat{\mathbf{Y}} = \mathbf{X}\beta_* = \frac{1}{1+g}\mathbf{X}\beta_0 + \frac{g}{1+g}\mathbf{H}\mathbf{Y}$, where $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$
- Therefore $\mathbf{a} = \frac{1}{1+g}\mathbf{X}\beta_0$ and $\mathbf{S} = \frac{g}{1+g}\mathbf{H}$, whose trace is $\frac{gp}{1+g}$

Appendix: Variants of the lead-lag R^2

Recall our model of gene expression – the R^2 from this model is called the lead-lag R^2 (LLR^2):

$$m_A(t) = c_1 m_B(t) + c_2 \int_0^t m_B(s)\,\mathrm{d}s + c_3 \int_0^t m_A(s)\,\mathrm{d}s + c_4 t + c_5.$$

Consider two "sub-models":

- Sub-model 1: R^2 from this model, called $\mathrm{LL}R^2_{\text{other}}$, captures variation in gene A explained by an*other* gene B.

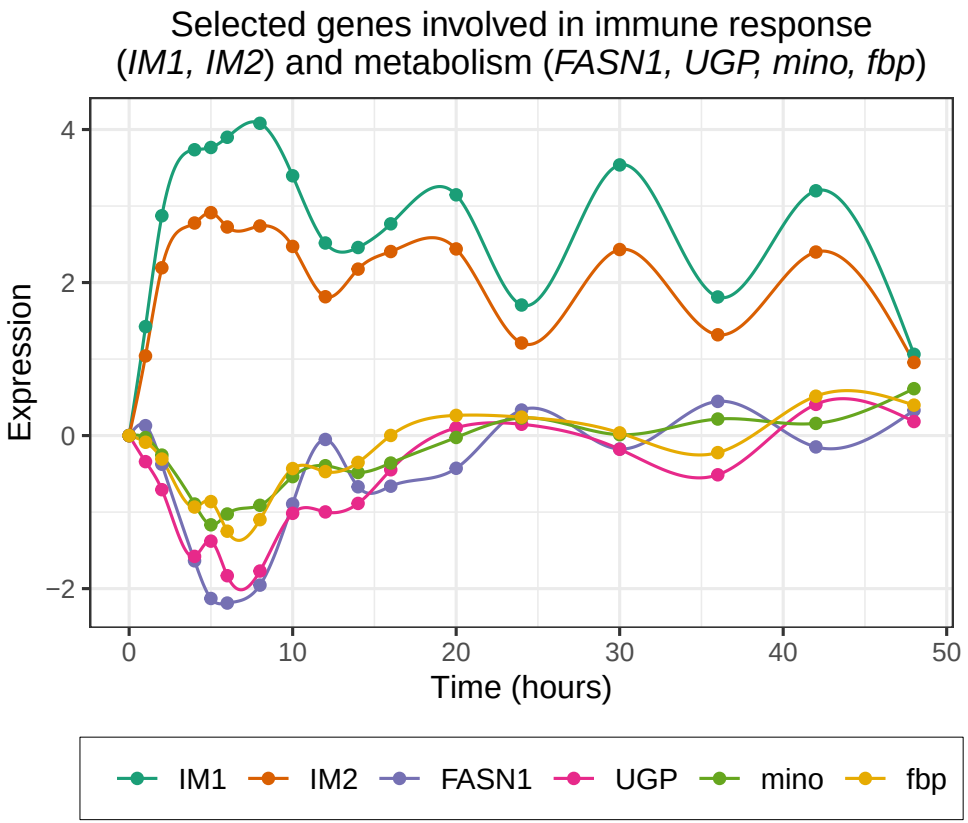
$$m_A(t) = c_1 m_B(t) + c_2 \int_0^t m_B(s)\,\mathrm{d}s + c_5$$

- Sub-model 2: R^2 from this model, called LLR^2_{own}, captures variation in gene A explained by its *own* past and linear time trends.

$$m_A(t) = c_3 \int_0^t m_A(s)\,\mathrm{d}s + c_4 t + c_5$$

In the scatterplots on slide 11, the x-axis shows $\mathrm{LL}R^2_{\text{other}}$ and the y-axis shows $\mathrm{LL}R^2 - \mathrm{LL}R^2_{\text{own}}$.

Appendix: Immune response and metabolism



Prior adjacency matrix **W**

	IM1	IM2	FASN1	UGP	mino	fbp
IM1	-	1	NA	0	0	NA
IM2	1	-	NA	0	0	NA
FASN1	NA	NA	-	NA	NA	NA
UGP	0	0	NA	-	1	NA
mino	0	0	NA	1	-	NA
fbp	NA	NA	NA	NA	NA	-

Bayesian lead-lag R^2 similarity matrix

	IM1	IM2	FASN1	UGP	mino	fbp
IM1	-	0.99	0.76	0.21	0.33	0.52
IM2	0.98	-	0.71	0.18	0.31	0.46
FASN1	0.82	0.80	-	0.77	0.97	0.78
UGP	0.30	0.30	0.83	-	0.88	0.99
mino	0.40	0.39	0.98	0.91	-	0.90
fbp	0.68	0.66	0.82	0.99	0.86	-

Red entries: previously known associations
Blue entries: previously unknown associations