Hearing the Shape of Life

Rosemary Braun
rbraun@northwestern.edu

Assistant Professor

Biostatistics / Preventive Medicine
Engineering Sciences and Applied Mathematics

Northwestern Institute on Complex Systems

NSF-Simons Center for Quantitative Biology

Northwestern University

Or: the emergence of long-range research behavior from local interactions ;)

SFI Complex Systems Summer School, 2009

- Adapation of the Partition Decoupling Method (Leibon &al 2008) for molecular subtyping of biological samples
  - Application of spectral graph theory
    to articulate clusters of samples at multiple scales
    through iterative partitioning of a similarity graph
- Braun, Leibon, Pauls, & Rockmore. Partition Decoupling for Multi-gene Analysis of Gene Expression Profiling Data, *BMC Bioinformatics* **497**:12, 2011.

Other uses of spectral graph theory to investigate living systems . . . ?

Or: the emergence of long-range research behavior from local interactions ;)

Other uses of spectral graph theory to investigate living systems?

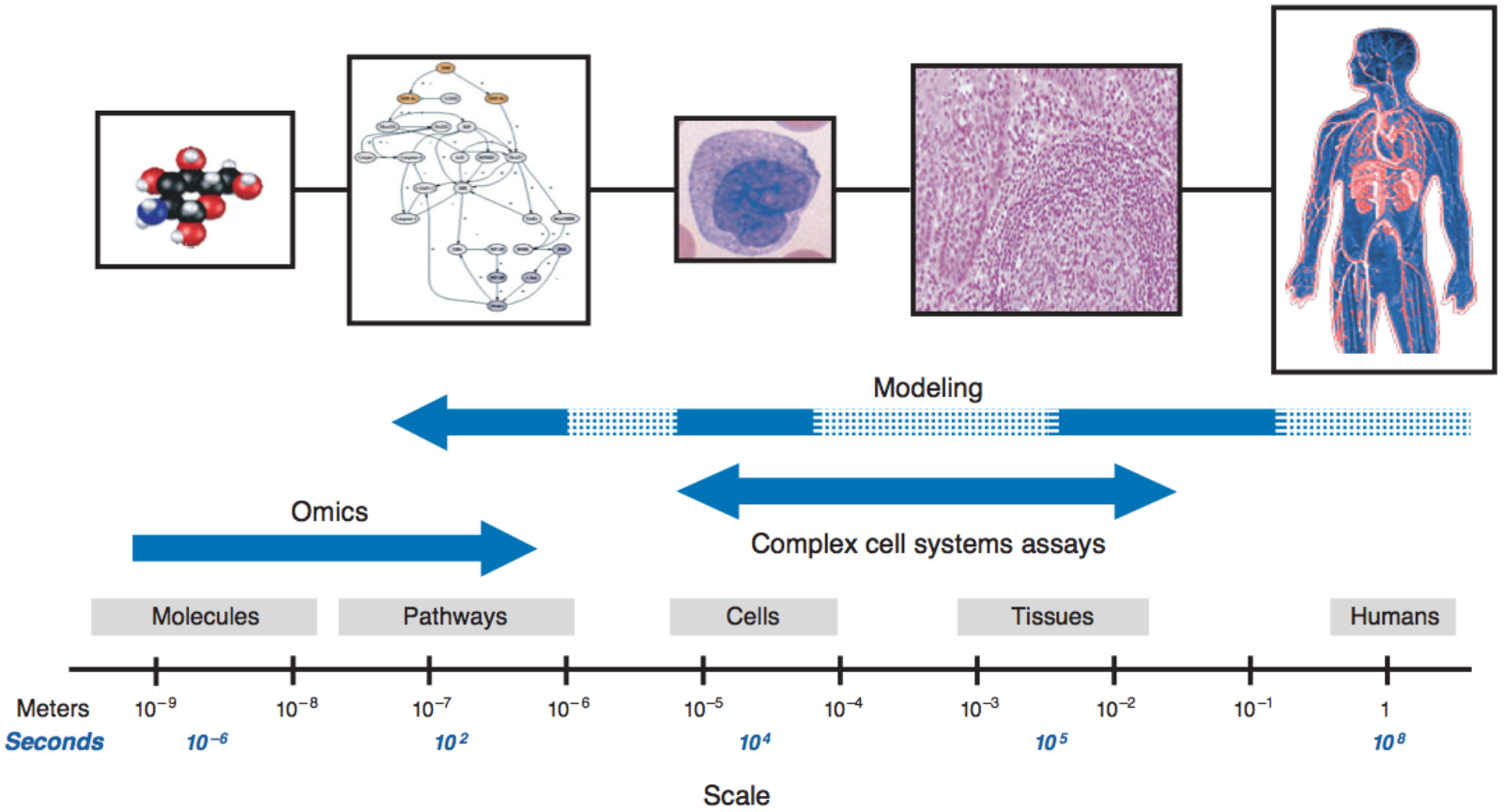
- Networks of samples (cells, tissues, organisms)
- Networks of genes (pathways; molecular interaction networks)

Northwestern (Biostats, Applied Math, NICO) [2011–]

- Novel Pathway Analysis Methods for Identifying Genomic Causes of Cancer (NIH/NCI K22 2012-2015)
- Modeling biocomplexity: From molecular interactions to population genetics (JSMF Complex Systems Scholar Award 2014-2020)

Motivation

Identify and control disease mechanisms by linking *detailed microscopic "∗omic" data* . . .



. . . to the *macroscopic phenotype*.

Modern molecular biology

We measure:

- high-throughput, *high dimensional* "omic" data: $10^4 - 10^6$ molecular features per sample

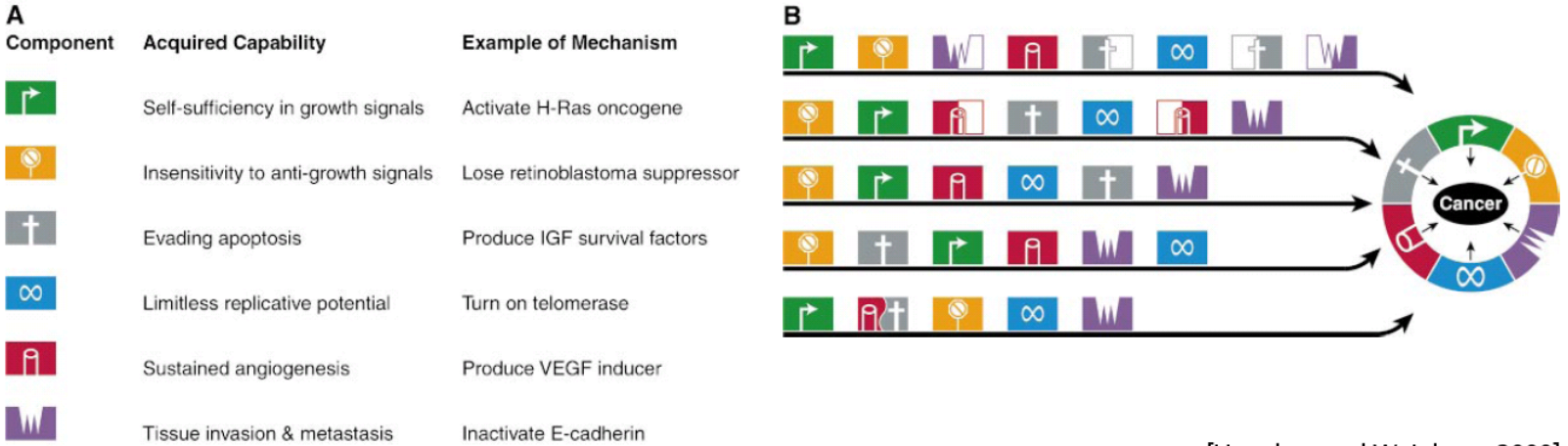
We want:

- to understand the emergence of *low dimensional* biological phenotypes

We know:

- the structure of molecular interaction networks (pathways)
- (but our knowledge is limited)

Challenge: Complex Diseases

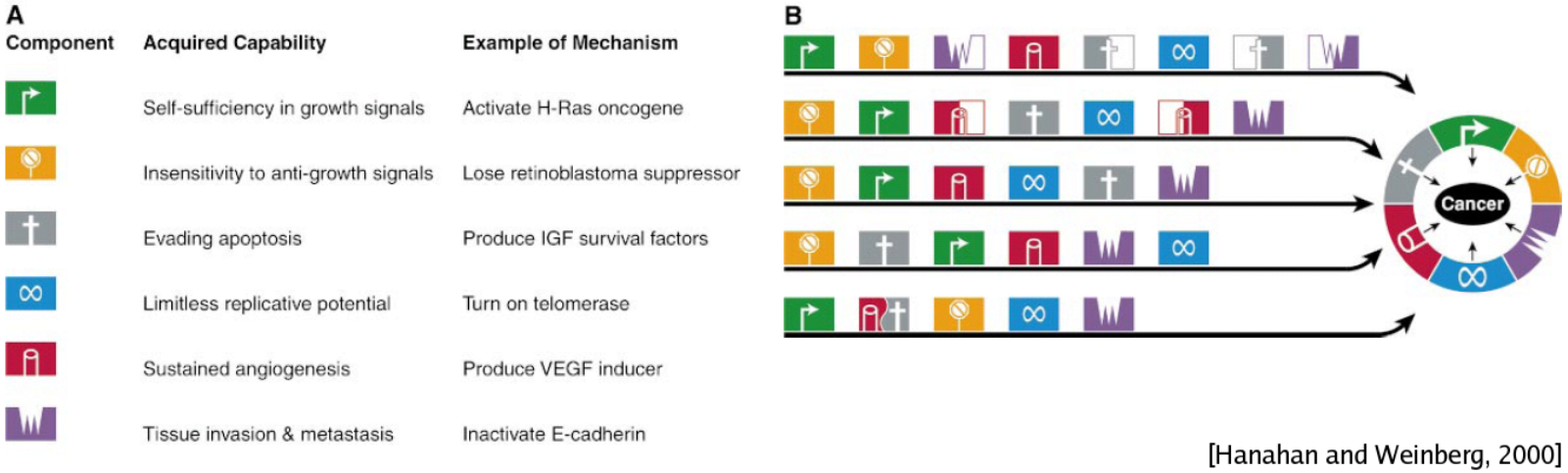


[Hanahan and Weinberg, 2000]

- Complex diseases (e.g. cancers) are driven by damage to *systems*, not individual genes.
- Complex diseases are *heterogeneous*: there's more than one way to damage a system.
- Systems are *robust*: cells may sustain damage to some genes in the system, but still function.

$\Rightarrow$ Important genes may not reach significance when tested individually.

Challenge: Complex Diseases



[Hanahan and Weinberg, 2000]

- Complex diseases (e.g. cancers) are driven by damage to *systems*, not individual genes.
- Complex diseases are *heterogeneous*: there's more than one way to damage a system.
- Systems are *robust*: cells may sustain damage to some genes in the system, but still function.

$\Rightarrow$ Important genes may not reach significance when tested individually.

Instead: take a *systems-level* view (pathways).

NICO

Why networks?

- Everything is connected!
  - Living systems — from the cell to entire populations — comprise interaction networks
  - Network structure $\Rightarrow$ system behavior

Why networks?

- Everything is connected!
  - Living systems — from the cell to entire populations — comprise interaction networks
  - Network structure $\Rightarrow$ system behavior
- As a way to make sense of high dimensional data
  - Modern molecular biology can measure 10^4–10^6 different genes in every sample
  - Finding key genes is a hunt for a needle in this haystack
  - Genes don't act alone
  - It's likely that there's more than one way to affect a system

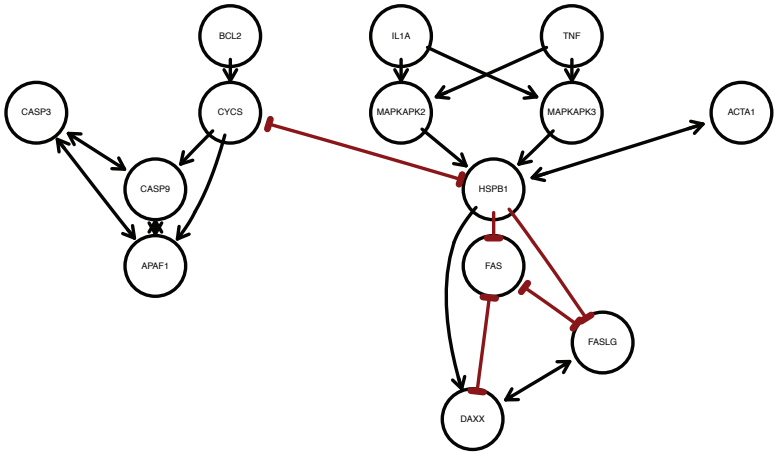
Why networks?

- Everything is connected!
  - Living systems — from the cell to entire populations — comprise interaction networks
  - Network structure $\Rightarrow$ system behavior
- As a way to make sense of high dimensional data
  - Modern molecular biology can measure 10^4–10^6 different genes in every sample
  - Finding key genes is a hunt for a needle in this haystack
  - Genes don't act alone
  - It's likely that there's more than one way to affect a system
- Spectral graph theory is beautiful and useful :)
  - How will a change in the network structure affect the overall properties of the network?
  - Can the network adapt/compensate for changes in one area with changes in another?
  - Can we infer something about the dynamics of the network, even if all we have is its topology?

Spectral Graph Theory

Graph Laplacian

Given a weighted, *un*directed pathway graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with gene vertices $\mathcal{V}$ and edges $\mathcal{E}$ respresenting a direct interaction of two genes or gene–products,



- (Normalized) Laplacian,

$$\overline{\boldsymbol{L}} = \boldsymbol{I} - \boldsymbol{D}^{-1/2}\,\boldsymbol{A}\,\boldsymbol{D}^{-1/2}\,,$$

- $\boldsymbol{A}$ is the weighted adjacency matrix,

$$A_{ij} = \begin{cases} \rho_{ij} & \text{if } (v_i, v_j) \in \mathcal{E} \\ 0 & \text{otherwise} \end{cases},$$

- $\boldsymbol{D}$ is a diagonal degree matrix with

$$D_{ii} = d_i = \sum_{j \neq i} |A_{ij}|\,.$$

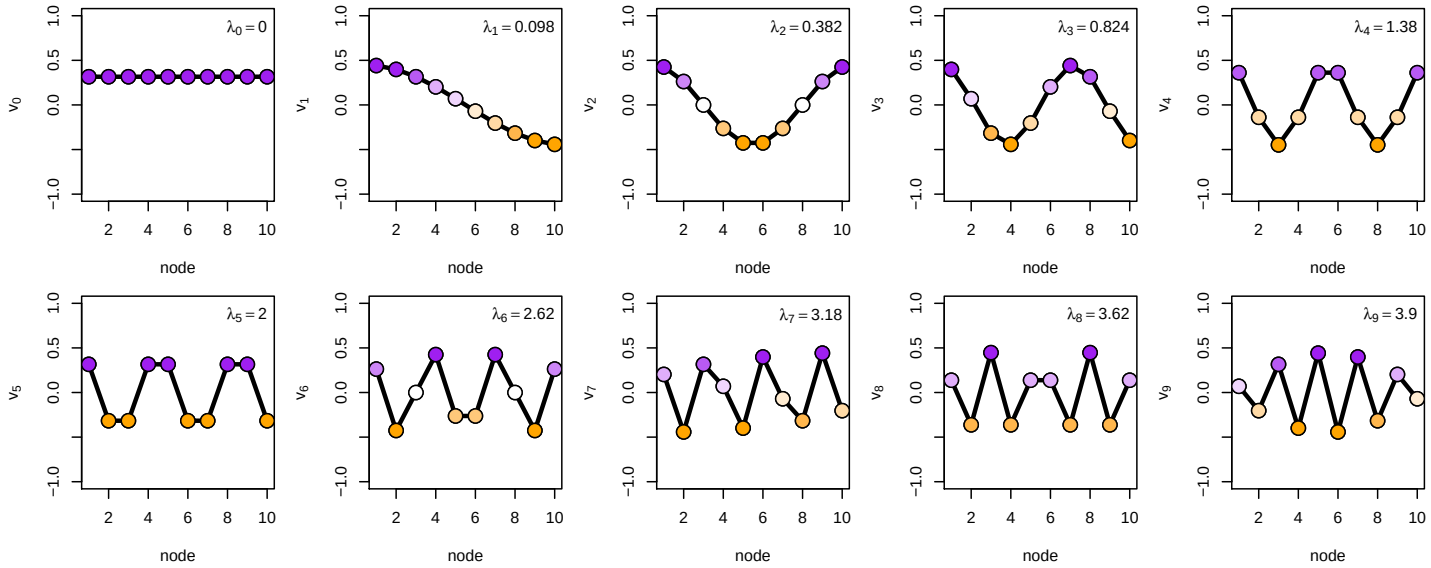
Graph Laplacian

- Encapsulates the network topology:
  - Which nodes interact (adjacency $\boldsymbol{A}$)
  - How many interactions a given node participates in (degree $\boldsymbol{D}$)
- Encapsulates the strength of interactions:
  - e.g. weights A_{ij} assigned from gene-gene correlations
- "Laplacian"?
  - Easy to show that $\boldsymbol{Lx}$ is the discrete form of the Laplace operator on a function $\boldsymbol{x_i} = f(\boldsymbol{v_i})$ of the vertices v_i
  - Minimizing $\boldsymbol{x}^\intercal \boldsymbol{Lx}$ represents minimizing the energy for many physical systems (e.g. displacements in a network of springs; voltages in a network of resistors)

Spectral decomposition of $\boldsymbol{L}$

Eigenvalues λ_k and eigenvectors v_k of $\boldsymbol{L}$:

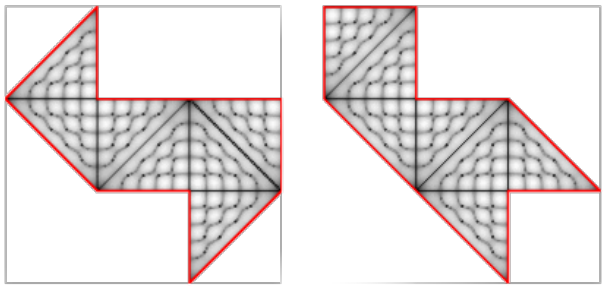
- Analogous to frequencies and normal modes



- Small-eigenvalued vectors describe the coarsest geometry
- $\lambda_1 =$ algebraic connectivity; indicates how easily the graph is partitioned (relaxation of min-cut), or, conversely, how readily the network will synchronize
- (Similar to PCA; in the case of a complete graph, can show equivalence to kernel-PCA)

"Hearing the Shape" of a network

- The geometry of the network can tell us something about dynamics of processes on the network (e.g. displacements, flow of current).
- Changing the edge weights can result in changes to the spectrum λ.
- Atay &al 2006: a network's *spectral* properties, rather than other network statistics, determines the dynamics.
- Isospectral graphs exist! Much like isospectral drums:

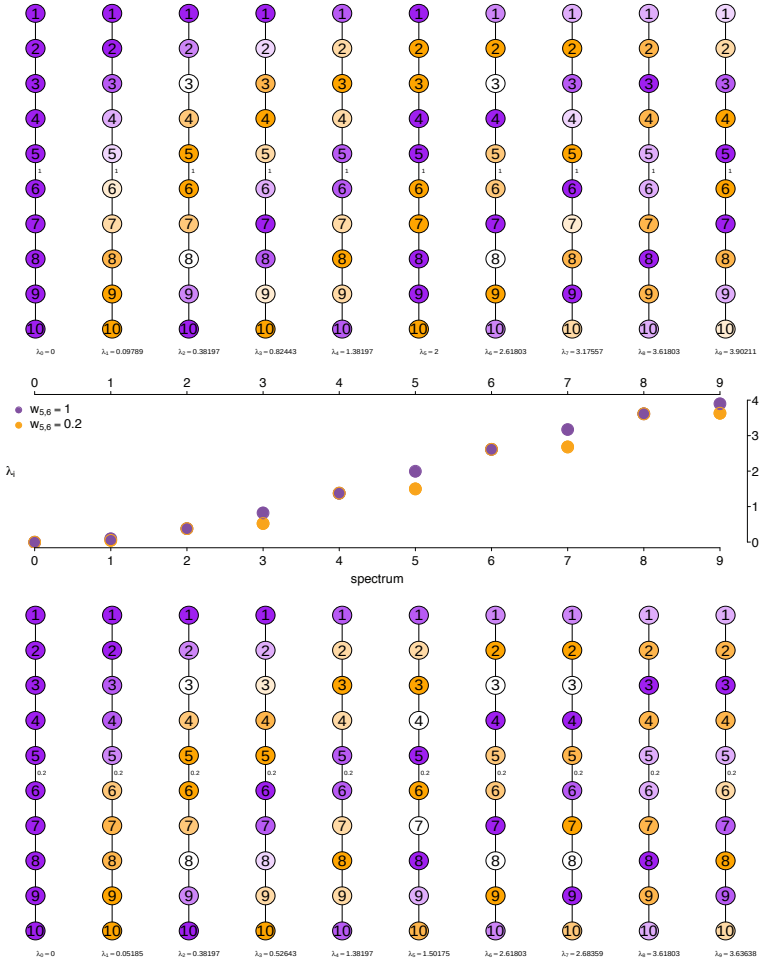


(Gordon, Webb, Wolpert 1992)

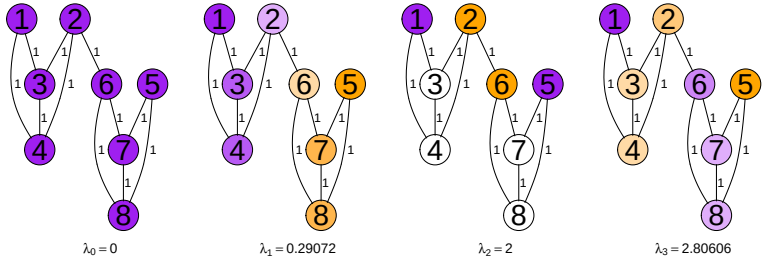
We can exploit these properties to summarize/analyze gene regulatory networks

Position matters

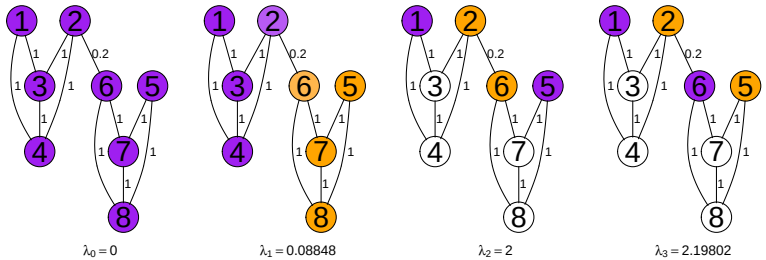
How strongly will a change to an edge weight affect the spectrum?



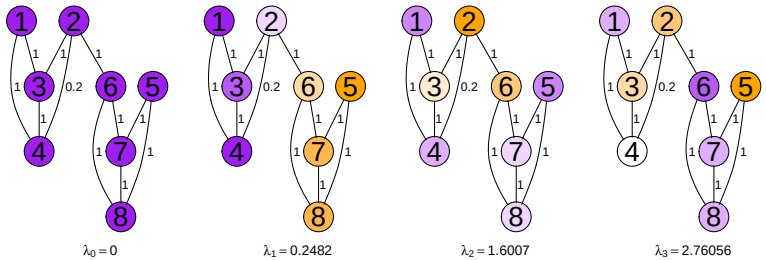
$a_{2,6} = 1,\ a_{2,4} = 1 \Rightarrow \lambda_1 = 0.29$:



$a_{2,6} = 0.2,\ a_{2,4} = 1 \Rightarrow \lambda_1 = 0.08$:



$a_{2,6} = 1,\ a_{2,4} = 0.2 \Rightarrow \lambda_1 = 0.25$:



Spectral Pathway Analysis

We can use these properties to:

- Detect pathways (networks) that appear to be differentially connected in cases vs. controls;
- Identify elements that contribute to network-wide gene regulatory differences;
- Make inferences about the time evolution of the network (under certain assumptions of gene regulation);
- Identify new regulators of network dynamics.

Several appealing features:

- No reliance on single-gene association statistics – consider "bulk" pathway behavior;
- Natural way to prioritize critical interactions;
- Noise reduction/robustness via filtering high-eigenvalued eigenvectors.

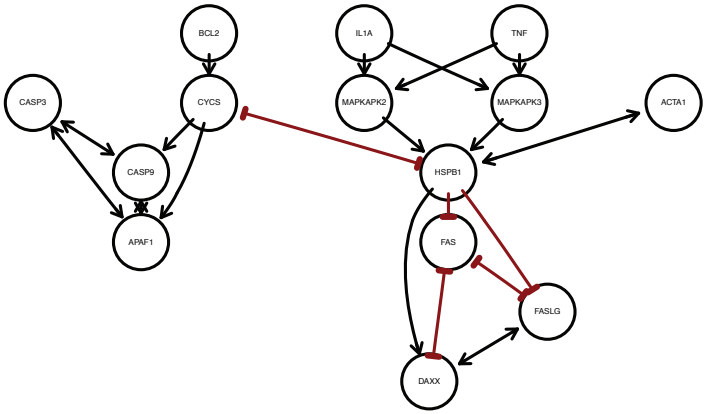
"Hearing the Shape" of Cancer

Spectral methods to infer aberrant network regulation

Pathway–level view

For a given pathway, we want:

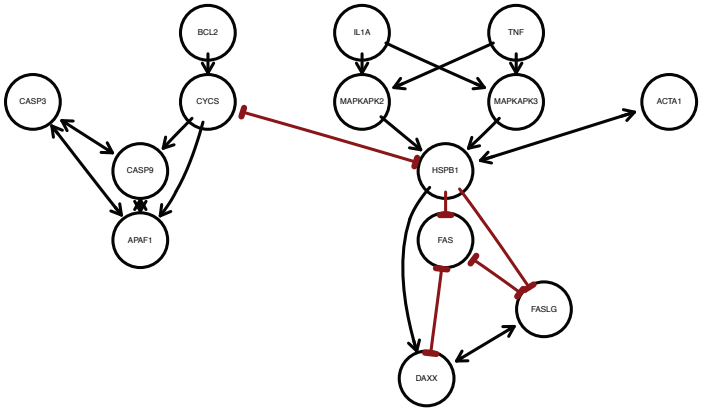
- a *pathway–wide summary statistic* that encapsulates the bulk variation of its constituent genes in the context of its network topology;
- to infer something about the pathway's *dynamical* properties from our static snapshot data.



Nodes in a network; the head of a drum.

- The graph Laplacian uniquely describes the geometry of a network (adjacency & degree of nodes, edge weights);
- Spectral decomposition of the graph Laplacian yields eigenvalue-eigenvector pairs that summarize the connectivity of the network and reveal its dynamical properties.

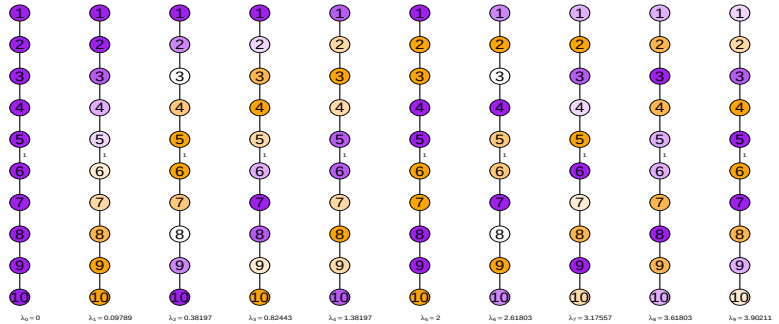
Pathway–level view

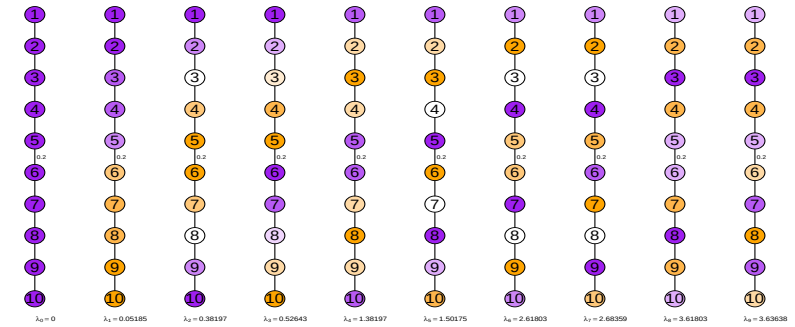


Idea: overlay experimental data onto a known interaction network and use the graph's spectral properties to say something about the behavior of the system as a whole.

- Integrates both gene expression and gene *co*-expression (correlation, MI, etc) data;
- Incorporates the pathway network topology (not all edges/nodes are equally critical);
- Encapsulates the bulk variation in the data for genes on that pathway;
- Robust to noise in gene expression measurements;
- Permits inferences about gene expression dynamics.

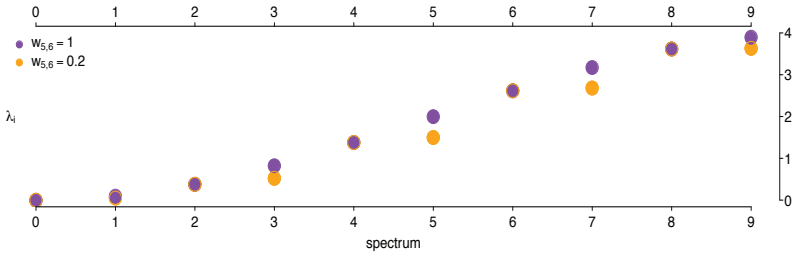
Pathway-wide coexpression





Comparing spectra:

Identify coexpression changes that are likely to influence bulk pathway characteristics.



Starting with putative pathway topology:

1. Weight the edges based on class-conditional gene-gene coexpression data;
2. Calculate eigenvalues and take differences between phenotypes;
3. Permute phenotype labels to assess statistical significance and flag pathways with significant spectral differences.

Application

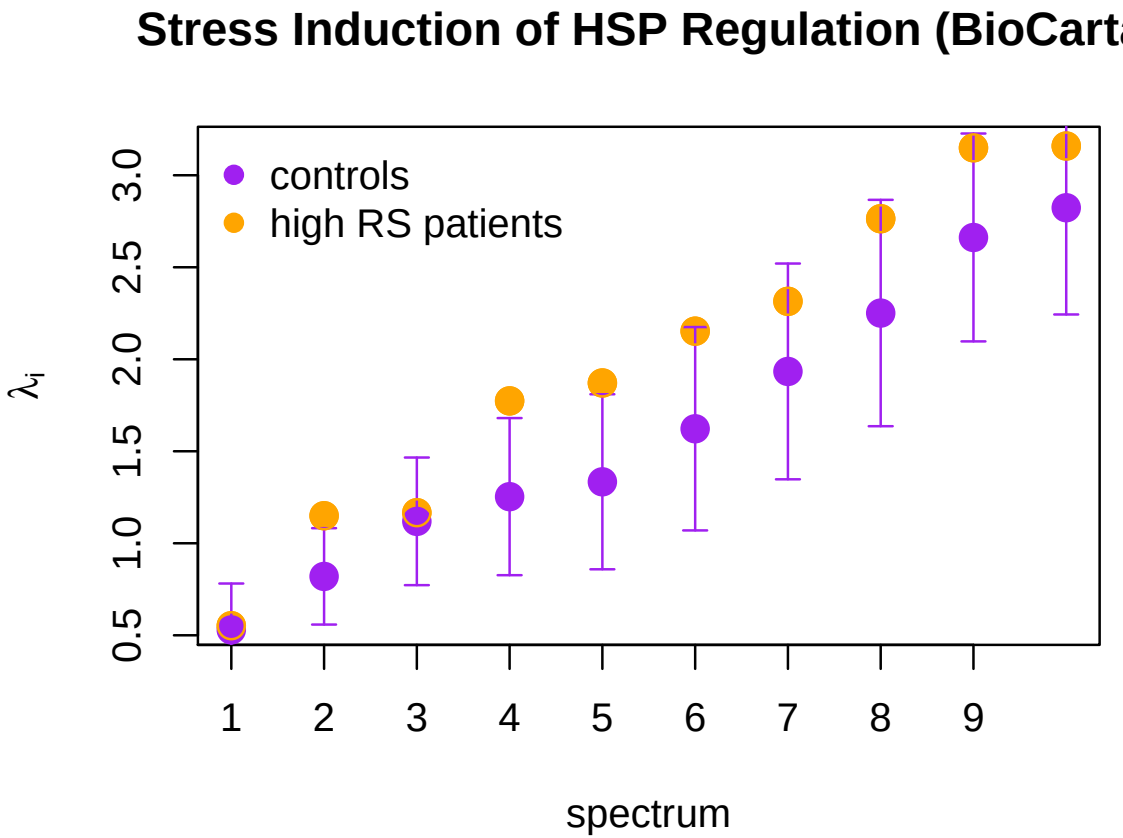
Radiation sensitivity study (public data, Reiger 2004, GEO accession GSE1725):

- Four phenotypes:
  - high radiation sensitivity cases (n=14)
  - low radiation sensitivity controls (n=13)
  - healthy controls (n=15)
  - skin cancer patients (n=15);
- Three radiation exposures: UV, ionizing radiation, mock;
- RNA from 171 samples hybridized to Affy HGU95Av2 chips (12625 probes);
- Intensities normalized using RMA [Bolstad 2003];
- Pathways retrieved from the NCI-PID database (663 pathways, 1195 connected components).

Systematically search all connected components for significant spectral differences in cases vs. controls.

Results: HSP pathway

An illustrative example (13th most significant):



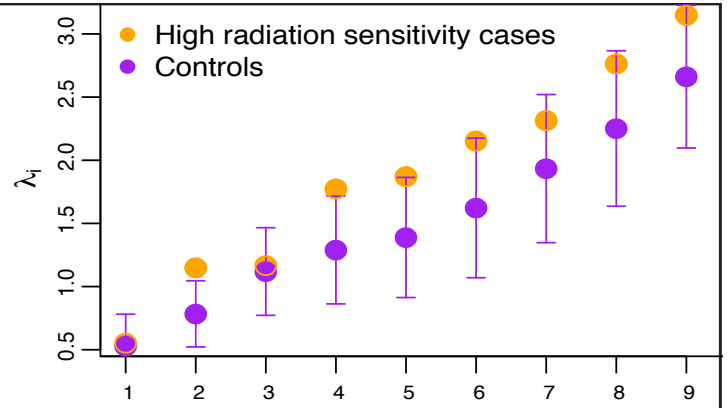
High λ_2 in the high radiation-sensitivity patients corresponds to increased coupling across the pathway...

HSP pathway

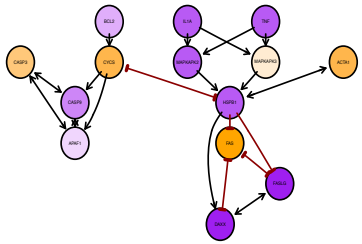
Example pathway: Stress Induction of HSP [BioCarta]

Inset right: spectrum of the pathway in cases vs. controls for first 9 eigenvalues. Errorbars indicate difference between case and control spectra under random label permutations, centered about true control values.

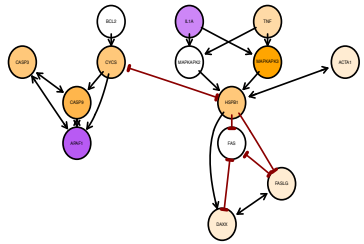
Below: network colored by eigenvectors values for the first four mode in cases vs. controls. Intensity of color indicates magnitude; purple and orange are of opposite sign.



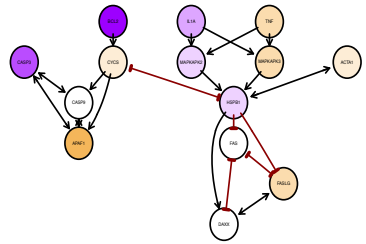
High radiation sensitivity cases:



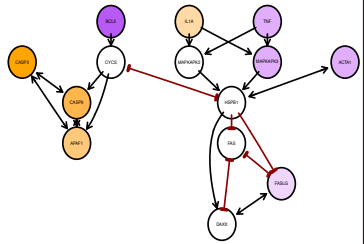
λ_1=0.55211



λ_2=1.14978

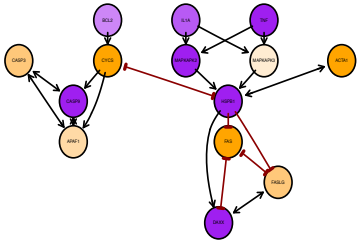


λ_3=1.16619

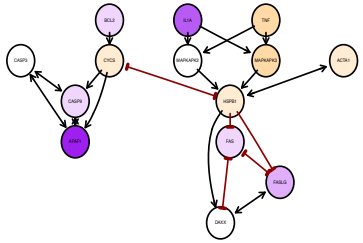


λ_4=1.77413

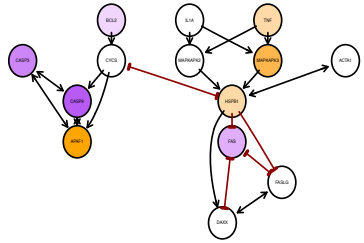
Controls:



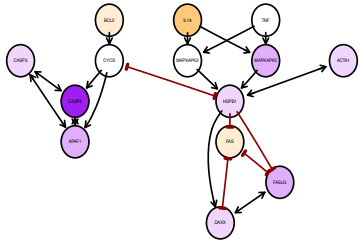
λ_1=0.52951



λ_2=0.82013

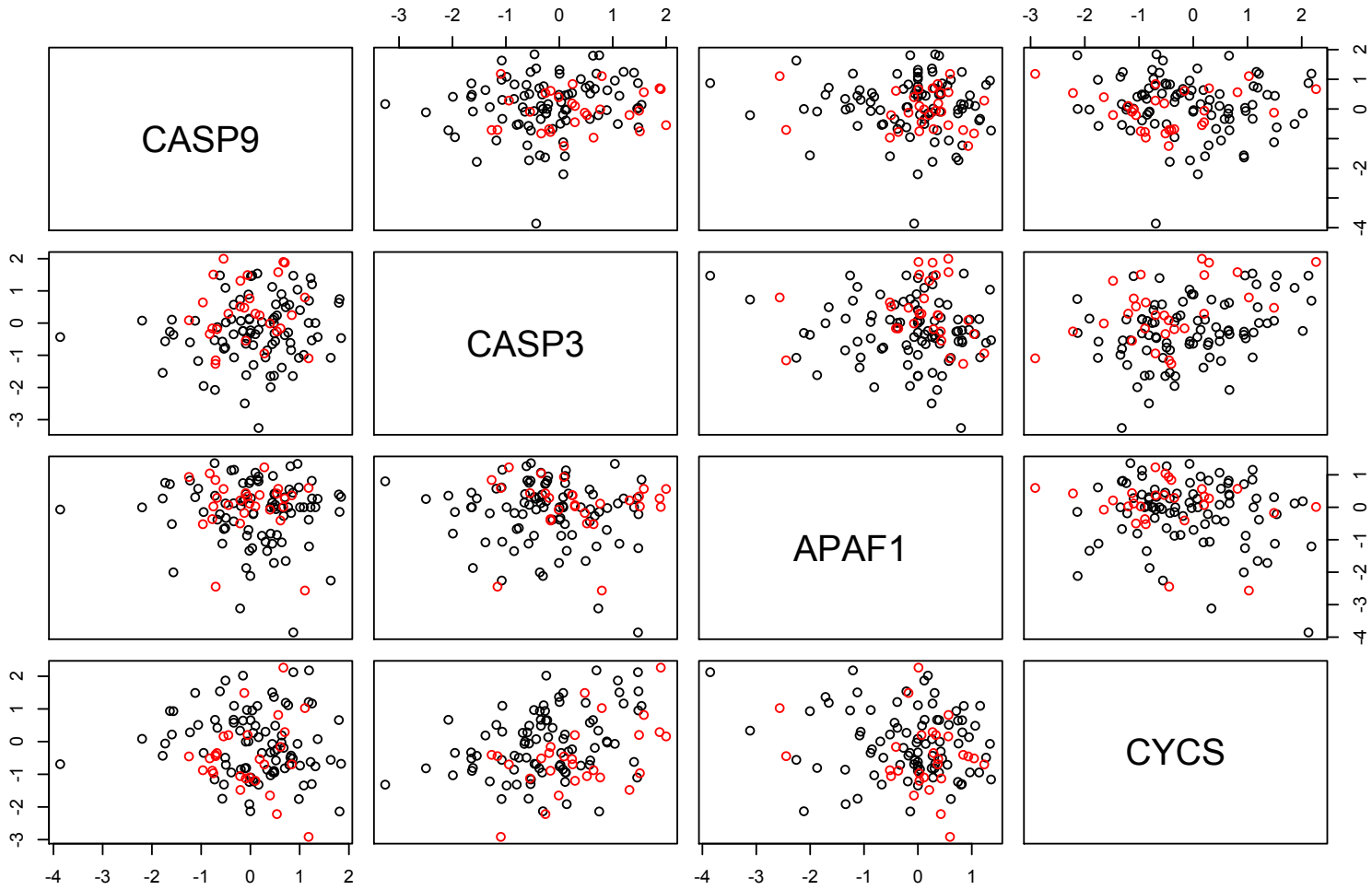


λ_3=1.1192



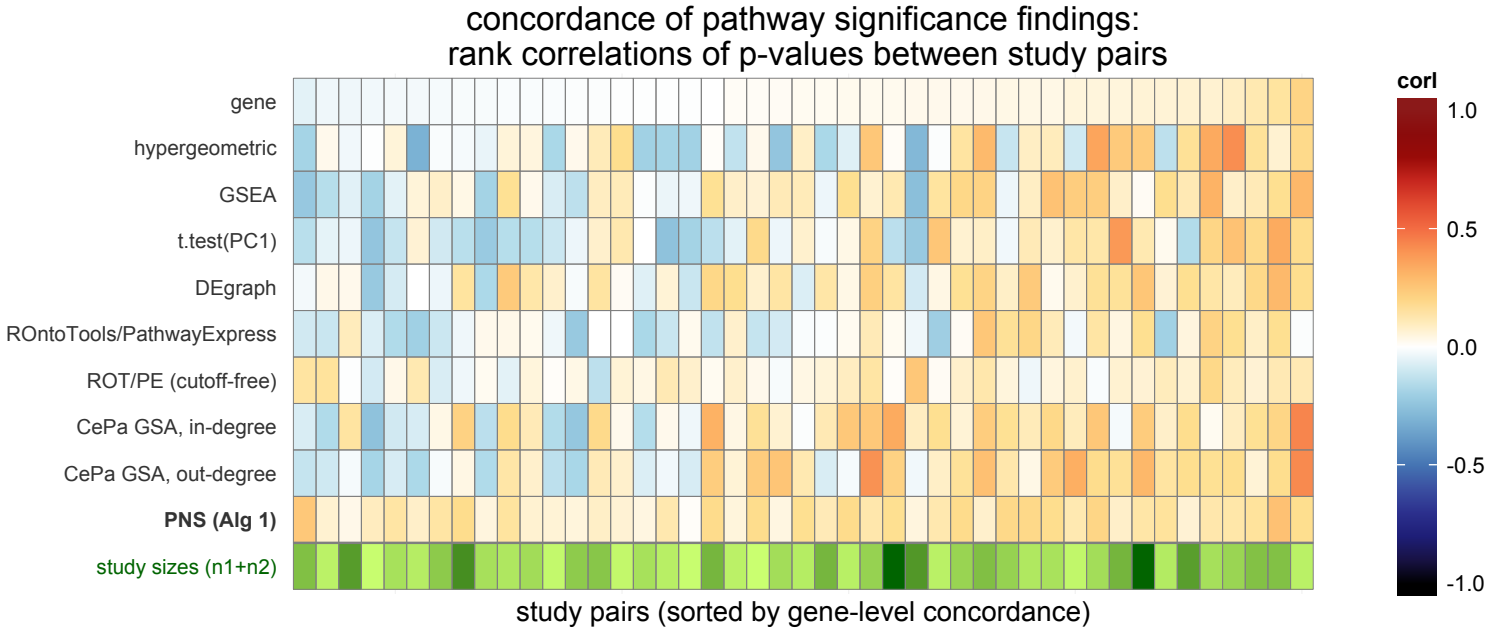
λ_4=1.25352

Subtle differences



Reproducible results

Greater reproducibility of findings across distinct studies of the same phenotypes compared to other statistical analysis methods:



From Network Structure to Network Function

Inferring differences in pathway dynamics from analysis of "snapshot" data.

From structure to function

Projection onto the network eigenvectors:

- Analogous to using PCA for dimension reduction, but "topology–aware;"
- Assess which modes are being hit in the phenotype of interest, *without* requiring that all samples do so in the same way.

  E.g., the same mode may be excited by down regulating one subnetwork or upregulating another, admitting molecular heterogeneity of complex diseases.
- In principle, these may be predictive of the pathway's dynamical response to the perturbation of a gene.

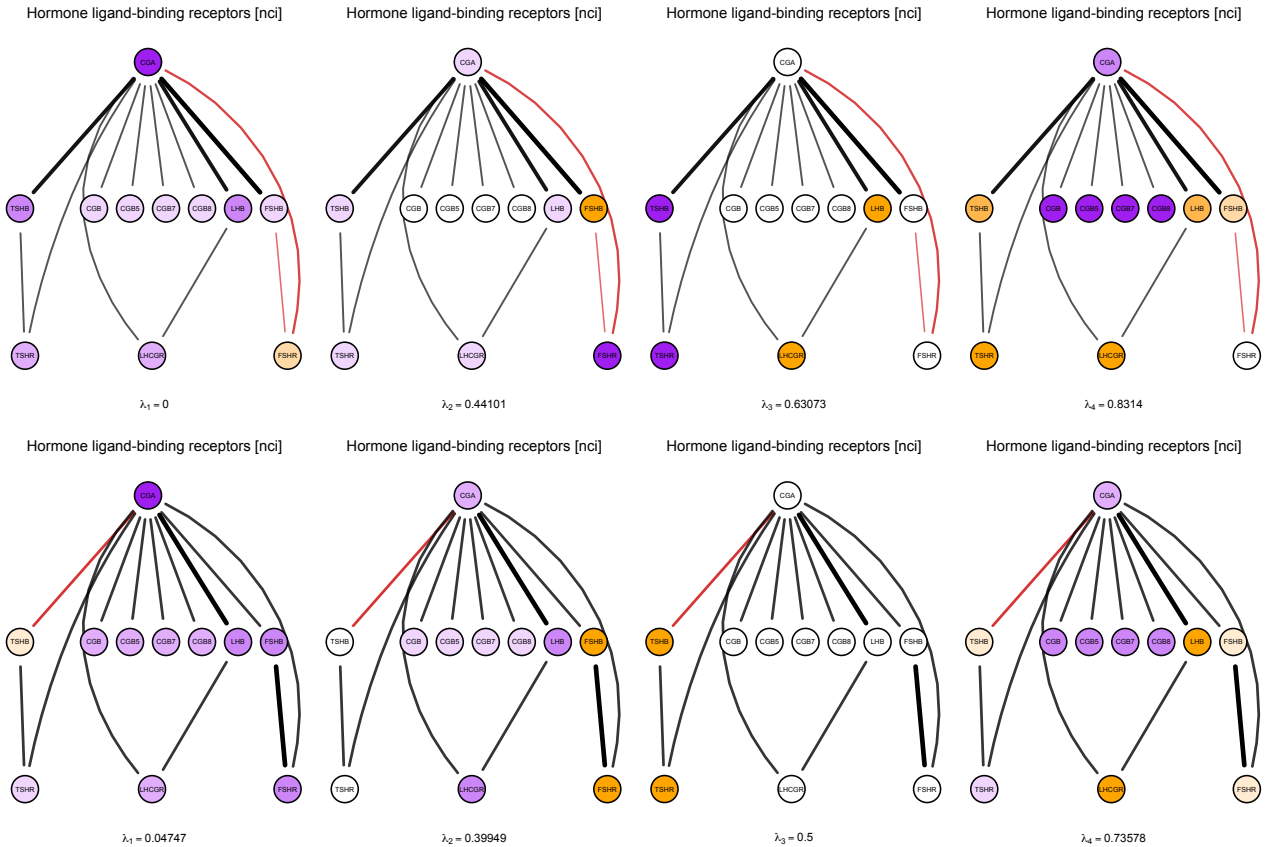
Spectral differences $\Rightarrow$ dynamical differences

Projection onto the network eigenvectors:

- ER+ breast cancer study using MCF-7 cells:
  - WS8 estrogen-dependent growth (typical ER+: deprive estrogen);
  - 2A non-responsive to estrogen deprivation;
  - 5C apoptoses in reponse to estrogen (after long-term estrogen deprivation).
- Edge weights assigned from a static "snapshot" study of cells under normal growth conditions;
- Pathway with significantly different spectra are flagged;
- Data from a separate time-course study following estrogen exposure is projected onto the eigenspace of those networks weighted by the WS8 data.

Differential connectivity

Hormone ligand binding receptors: first 4 "modes"



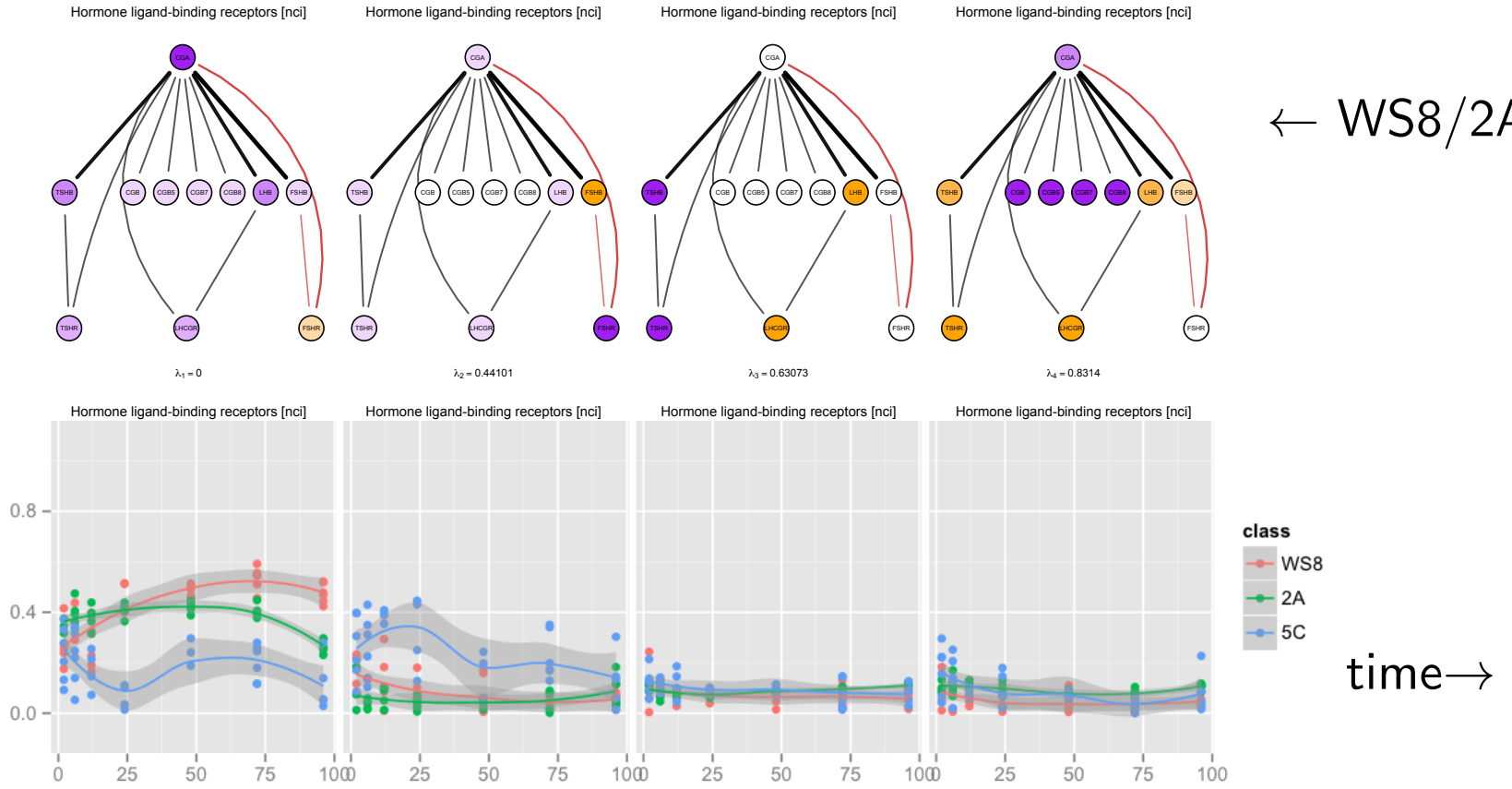
← WS8/2A

← 5C

(Recall: WS8 requires estrogen; 2A does not; 5C dies.)

Differential dynamics

Projection of gene expression onto pathway eigenvectors:



← WS8/2A

time→

Significantly different projections over time. Notably, the WS8 cells all tend toward the first (lowest eigenvalued) mode over time, while 2A cells do not sustain the response and 5C move away from it.

Summary

- Spectral decomposition of the weighted pathway Laplacian summarizes the correlation structure across the pathway in the context of known biological interactions.
- Significance testing does not rely on single–gene (or single–interaction) metrics, permitting the discovery of pathway–wide differences in network connectivity.
- Differences in weights of "bridge" / "bottleneck" edges necessarily have a stronger influence on the spectra than differences in a circumnavigable edge.
- Projection of gene expression values onto the eigenvectors may yield insights into the dynamics of the network.

Open questions

We assumed (to ensure $\boldsymbol{L} \succeq 0$):

- undirected graphs;
- no self links.

However, real biological networks:

- are directed (i may control j but not vice-versa);
- have both activating ($+$) and inhibiting ($-$) interactions;
- have autoregulating nodes (self loops).

What is the best way to handle this?

- If $\boldsymbol{L} \not\succeq 0$, how should we interpret the complex spectrum or non-orthogonal eigenvectors?
- Is there a way to formulate the analysis to ensure $\boldsymbol{L} \succeq 0$?

Open questions

Supposing we have a pathway with a spectrum that is modified in disease. . .

- What is the minimal number of edge-weight changes required to approximately recover the spectral properties of a graph?
- If I can only perturb specific edges (e.g., with a drug), what magnitude of changes is needed to approximately recoverthe spectral properties?
- Is recovering the spectral properties enough to recover the biological properties?

Thanks!

Braun lab:

- Sahil Shah
- Gary Wilk
- Marta Iwanaszko
- Phan Nguyen

NU collaborators:

- Bill Kath
- Ravi Allada
- Lifang Hou
- Wei Zhang

VCU:

- Seth Corey

Rice:

- Marek Kimmel

Santa Fe Institute (PDM):

- Dan Rockmore
- Scott Pauls
- Greg Leibon
- Fellow CSSS '09 students

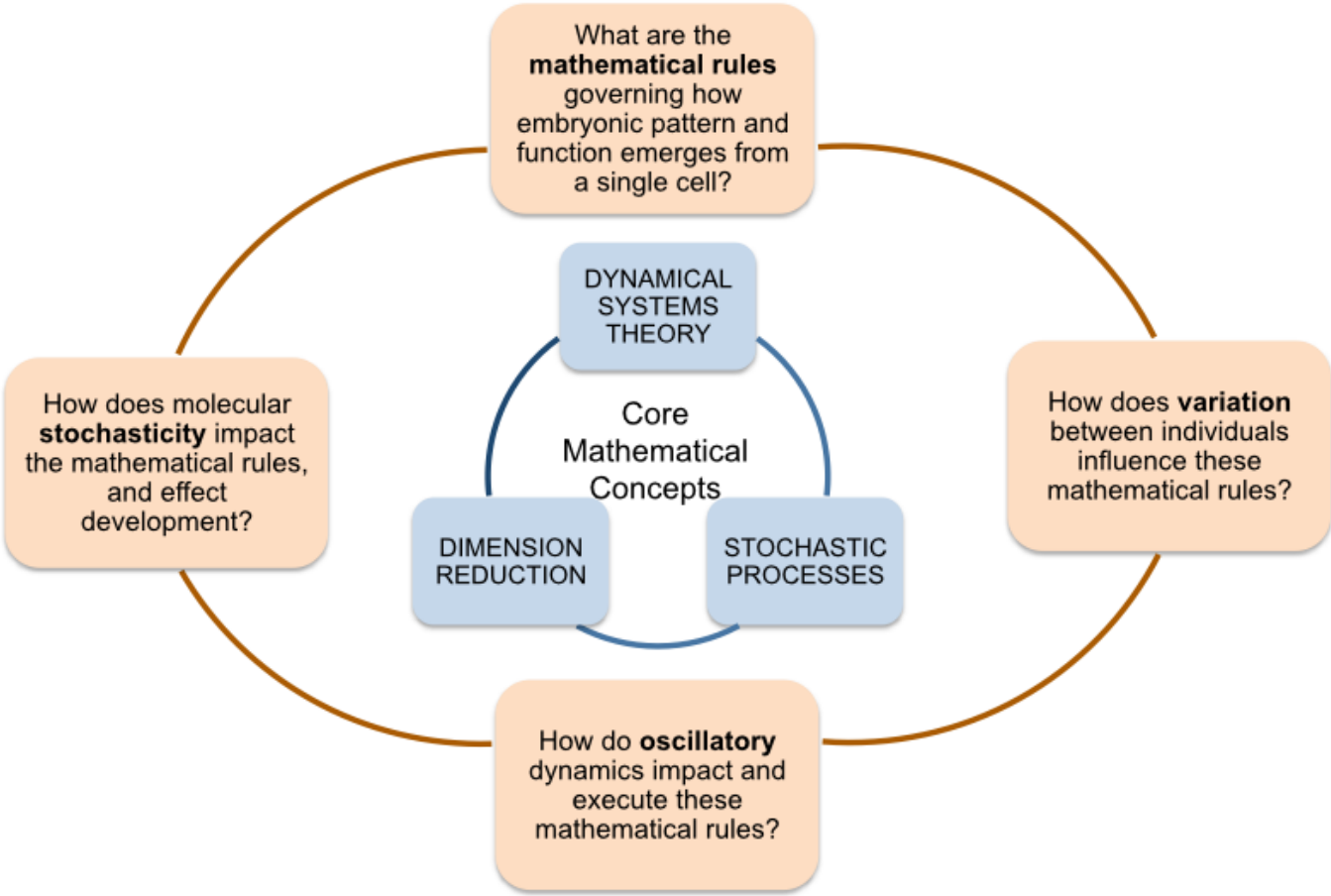
Funding:

- NIH/NCI (K22-CA148779)
- James S. McDonnell Foundation
- NU Data Science Initiative
- NSF-Simons Center for Quantitative Biology

NSF - Simons Center for Quantitative Biology

Northwestern University, Evanston Campus

The Application of Mathematics to Biological Dynamics



NSF - Simons Center for Quantitative Biology

Northwestern University, Evanston Campus

VISITING SCHOLARS PROGRAM

Eligibility
- Faculty, Postdocs, Students from US and abroad
- Rolling application
- Interest in applying mathematics to developmental biology

Duration
- 2 weeks - 6 months
- Capacity for six scholars at a time

Benefits
- Embedded in interdisciplinary groups
- Housing provided five min walk from campus
- Opportunities for training & collaboration

CONTACT:
NSF-Simons-QBio@northwestern.edu
http://sites.northwestern.edu/mathbiosys/