

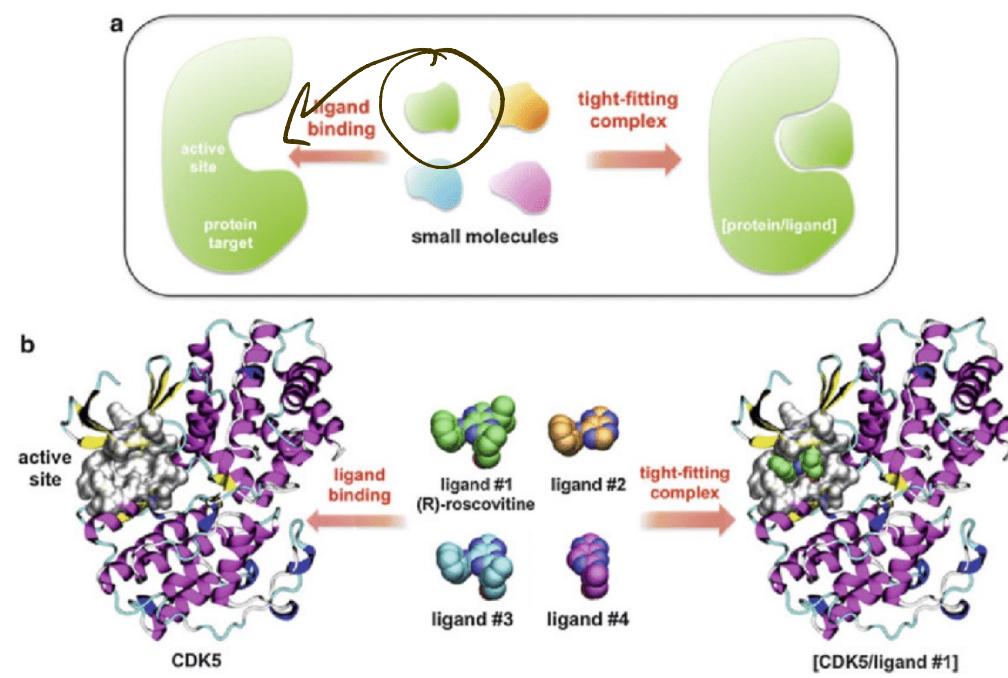
Recall On being the right size, revisited: The problem with engineering metaphors in molecular biology, by Nicholson (2020)

Protein interaction networks : structure

DNA → RNA → protein

primary → secondary → tertiary structure

amino acid sequence helices & sheets 3D shape



Humans: >10,000 "canonical" proteins
maybe 80k-400k with modifications

Mouse: ~55k proteins

Ameba: ~14kc proteins

two types of binding

① stable binding
(building infrastructure)

② transient binding
(information flow)

protein interaction databases: STRING

v7
covers 10,000,000 proteins
across 2000+ organisms

HTTPS://string-db.org/

Version: 11.0 LOGIN | REGISTER

STRING

Protein by name >
Protein by sequence >
Multiple proteins >
Multiple sequences >
Proteins with Values/Ranks New >
Organisms >
Protein families ("COGs") >
Examples >
Random entry >

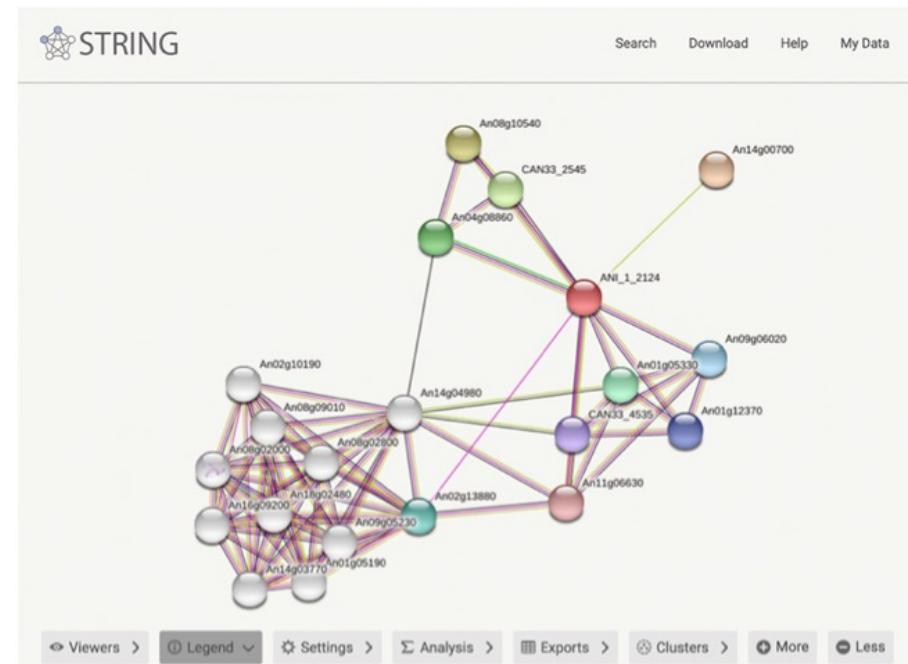
SEARCH

Single Protein by Name / Identifier

Protein Name: (examples: #1 #2 #3)
Organism: auto-detect

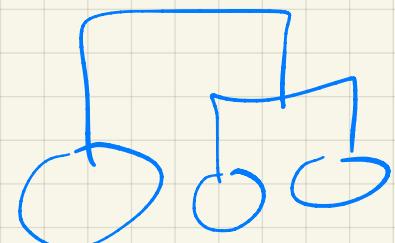
SEARCH

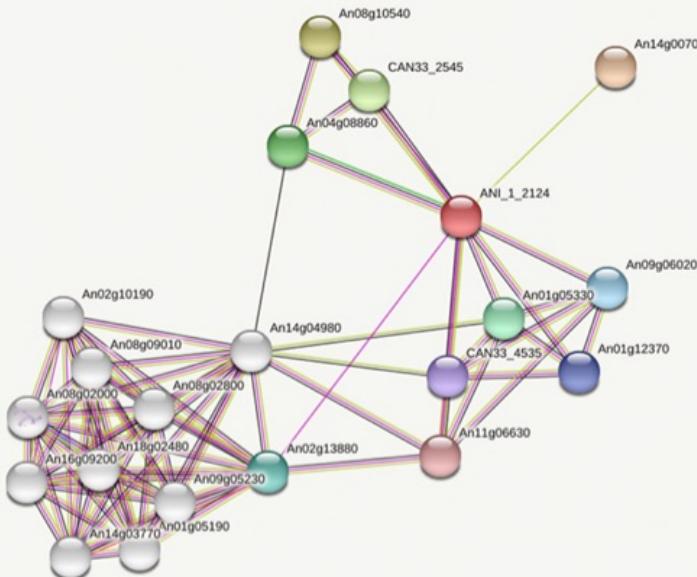
Random entry: ANI_1_2124



What do you notice about the local PPIN neighborhood of ANI_1_2124?

- lots of triangles
- maybe hierarchical?
- two "modules"





Viewers > Legend > Settings > Analysis > Exports > Clusters > More > Less

Nodes:

Network nodes represent proteins

splice isoforms or post-translational modifications are collapsed; i.e. each node represents all the proteins produced by a single, protein-coding gene locus.

Node Color

colored nodes:
query proteins and first shell of interactors
white nodes:
second shell of interactors

Node Content

empty nodes:
proteins of unknown 3D structure
filled nodes:
some 3D structure is known or predicted

Edges:

Edges represent protein-protein associations

associations are meant to be specific and meaningful, i.e. proteins jointly contribute to a shared function; this does not necessarily mean they are physically binding each other.

Known Interactions

from curated databases
experimentally determined

Predicted Interactions

gene neighborhood
gene fusions
gene co-occurrence

Others

textmining
co-expression
protein homology

Your Input:

ANI_1_2124 ATPase family associated with various cellular activities (AAA) family protein; Putative uncharacterized protein An14g00020; Belongs to the AAA ATPase family (728 aa)

Predicted Functional Partners:

- An14g00700 AFG1-like ATPase family protein; Similarity to hypothetical ATPase HFN2B - Haematobia irritans (552 aa)
- An08g10540 Zinc-finger double-stranded RNA-binding family protein; Similarity to hypothetical protein CAD60750.1 - Podospora...
- CAN33_2545 Function- ribosomal protein L24 interacts with the 5. 8 S rRNA of *S. cerevisiae* (190 aa)
- An04g08860 Putative uncharacterized protein An04g08860 (406 aa)
- An01g05330 NPL4 family protein; Nuclear protein localization protein 4 (654 aa)
- An02g13880 Pre-mRNA splicing factor component family protein; Putative uncharacterized protein An02g13880 (791 aa)
- An09g06020 SEP domain family protein; Putative uncharacterized protein An09g06020 (388 aa)
- An01g12370 Similarity to hypothetical protein SPCC285.11 - *Schizosaccharomyces pombe* (525 aa)
- CAN33_4535 Ubiquitin elongating factor core family protein; Pathway- involved in N-terminal ubiquitin fusion degradation proteo...
- An11g06630 HECT-domain (Ubiquitin-transferase) family protein; Putative uncharacterized protein An11g06630 (1797 aa)

Your Current Organism:

Aspergillus niger

NCBI taxonomy ID: 5061

Other names: *A. niger*, ATCC 16888, *Aspergillus niger*, *Aspergillus niger* Tiegh., CBS 554.65, NRRL 326

rich metadata! but very inconsistent coverage

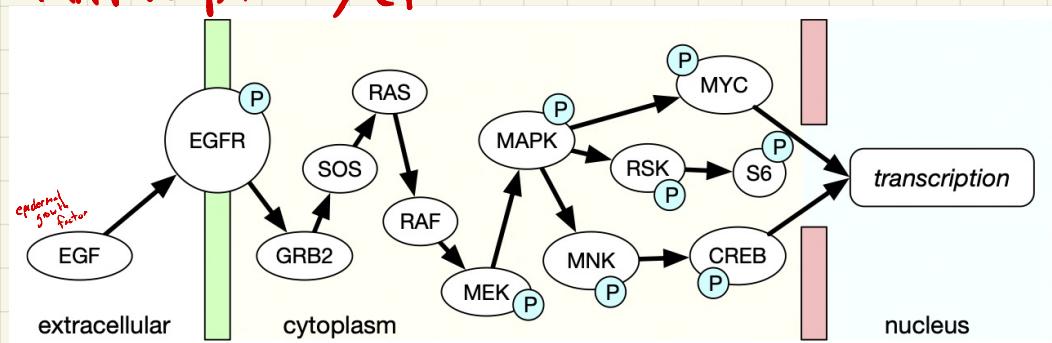
node annotations
edge annotations

some experimentally validated data
some model-based predicted data

Zooming in : Signaling Pathways

edges directed (flow of signal)
here, all activation (no inhibition)

MAPK pathway (partial)



activation

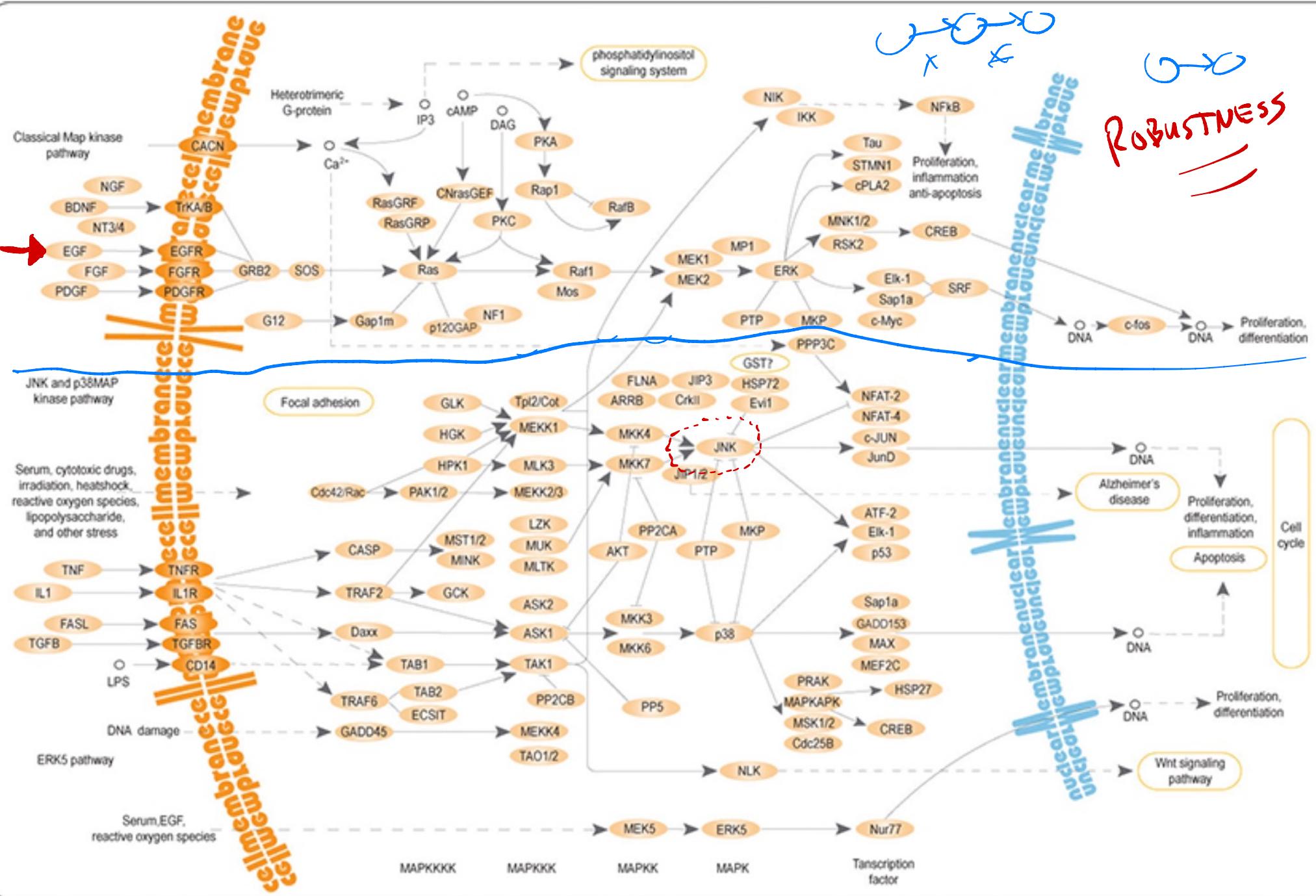


inhibition

changes to MAPK: cancer
inflammatory diseases
obesity
diabetes

Zooming out a little: MAPK + WNT signalling

activation AND inhibition
wow crosstalk → why so much?



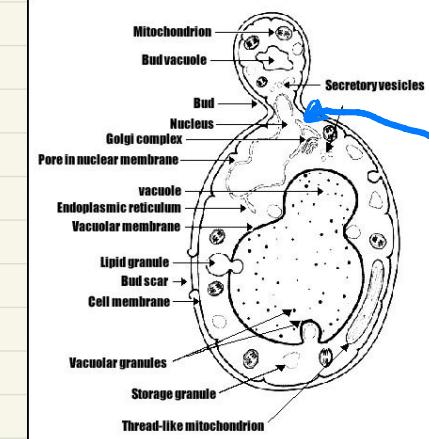
Where does the data come from?

Experiments

① yeast two-hybrid (Y2H)

- edge sampling $\textcircled{X} \rightarrow \textcircled{Y}$
- insert $\textcircled{X}-\textcircled{Y}$ inside an existing nuclear binding event
- limitations
 - nuclear protein only
 - high false positive rate
 - high false neg rate

Bakers yeast (*S. cerevisiae*)

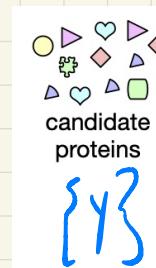
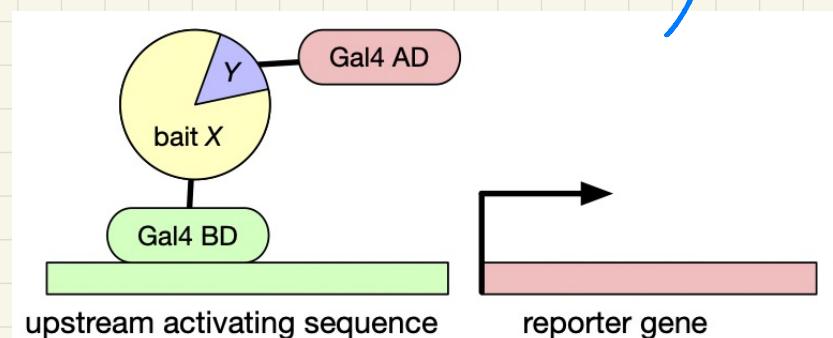


② affinity purification + mass spectrometry (AP-MS)

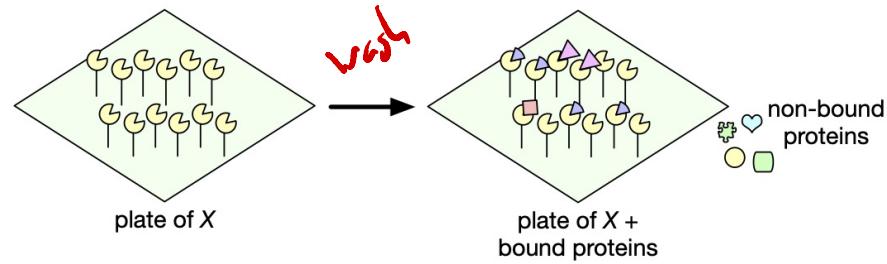
- node sampling



- wash $\{\textcircled{Y}\}$ over \textcircled{X}
 - high FP
- limitations
 - high FN



$\{\textcircled{Y}\}$



Models

③ text mining

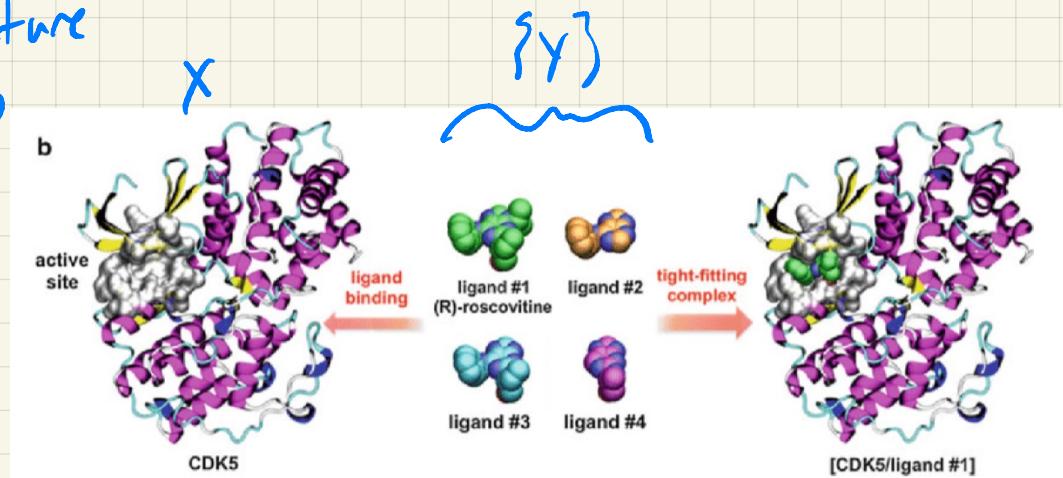
- use NLP tech to "read" scientific literature
- mainly low throughput experiments
- limitations
 - false pos
 - false neg

} complicated

④ predictions from protein domain analysis

- secondary / tertiary structure
 - ↳ shape + location of binding domains

- limitations - neck structure
 - FP
 - FN



error rates

$\Theta(n^2)$ possible interactions among $\{n\}$ proteins

$n \approx 10k$ (canonical proteins)

→ 100,000,000 possible interactions most of which never occur

So why?

high false positive rates



high false negative rates



$$\begin{cases} \Theta(n) \text{ TP} \\ \Theta(n^2) \text{ TN} \end{cases}$$

$$\begin{cases} \text{TP: 0.99} \\ \text{TN: 0.99} \end{cases}$$

$$0.99 \times 10,000 = 9,900$$

$$FN: 100$$

$$\begin{aligned} &0.99 \times (100,000,000 - 10,000) \\ &= 98,000,000 \\ &FP = 100,000 \end{aligned}$$

the reality of PPIN data \rightarrow messy + lots of FP

- 1) binary interactions only \Rightarrow no weights / binding affinities
no "sign" on edge: inhibition/activation
- 2) no data on where in the cell the $(X-Y)$ occurs
or when in cell cycle
- 3) no data on protein concentrations
- 4) missing functional labels

What to do with PPINs?

- 1) fill in the details : predict / observe missing attributes
" links
 - 2) find modules : "building blocks of complexity"
 - 3) understand diseases : pathway analysis
modules
interaction dynamics
etc. etc.
- ii) systems biology
- what does G tell us about life?
complexity?
evolution?
cooperation?

tools

1) motifs (small subgraphs)

- decompose complex into smaller units
- PIN = digital circuit metaphor

2) pathways or modules

- structure + dynamics → function

3) "hubs" (high-degree nodes)

- date hubs vs party hubs
- edge weight = importance

"Computing" with a PPIN

rate equations

- dynamics of protein concentrations (with noise?)
requires detailed biophys. kinematics info (hard)

linear Models

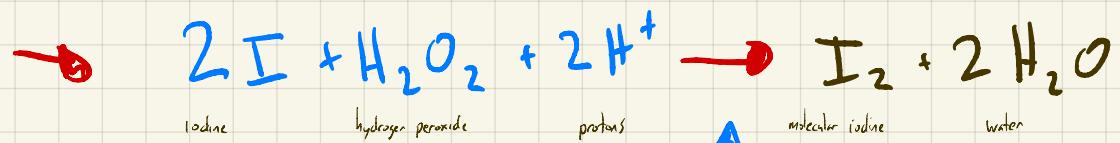
- approx rate eqtns w/ simple linear functions
need "weight" info

boolean networks

- approx linear models w/ simple boolean logic

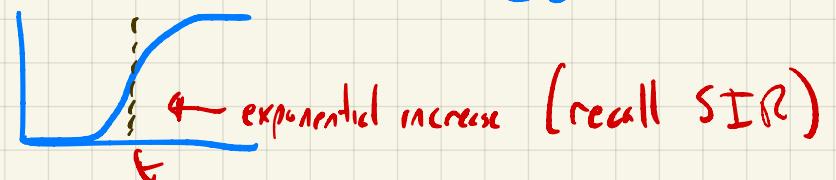
(simplicity)
(tractability)

Rate equations:



$$\text{rate} = k[\text{I}^-][\text{H}_2\text{O}_2] = \frac{\Delta [\text{things}]}{\Delta t}$$

* logistic equation solution

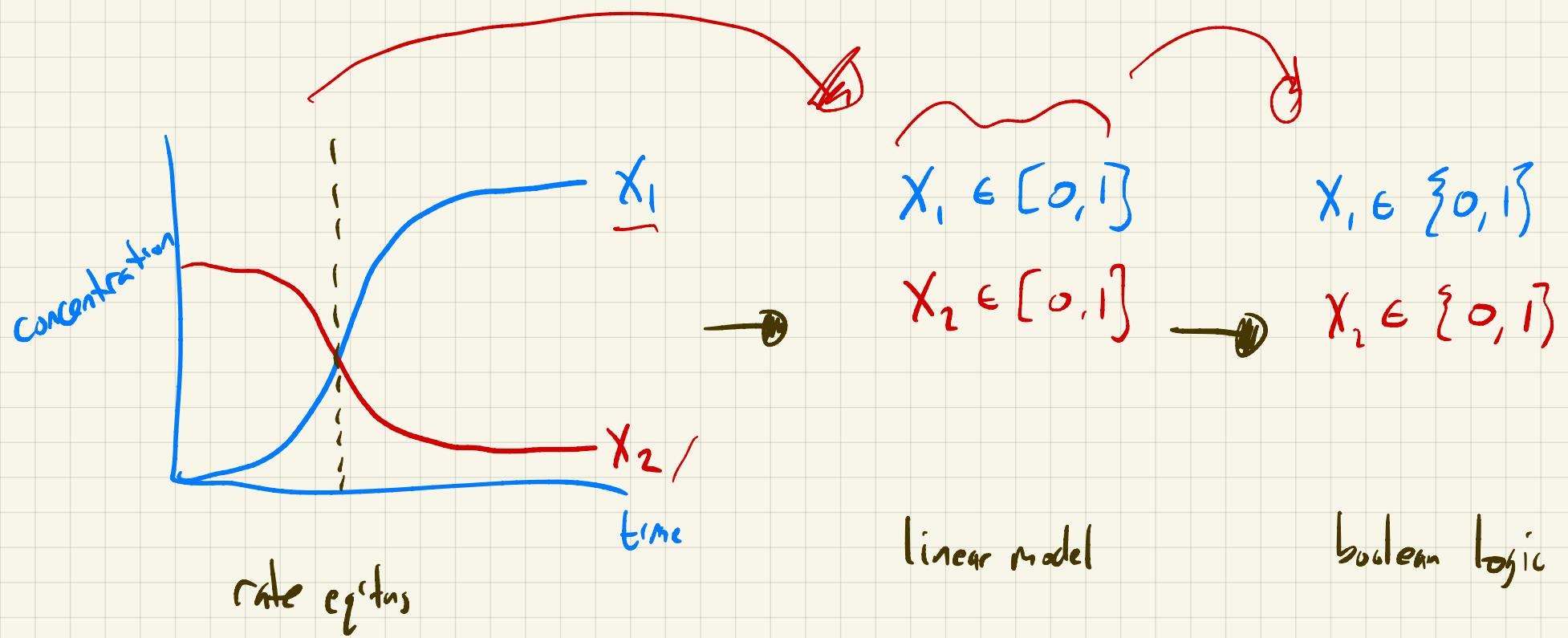


Advantages

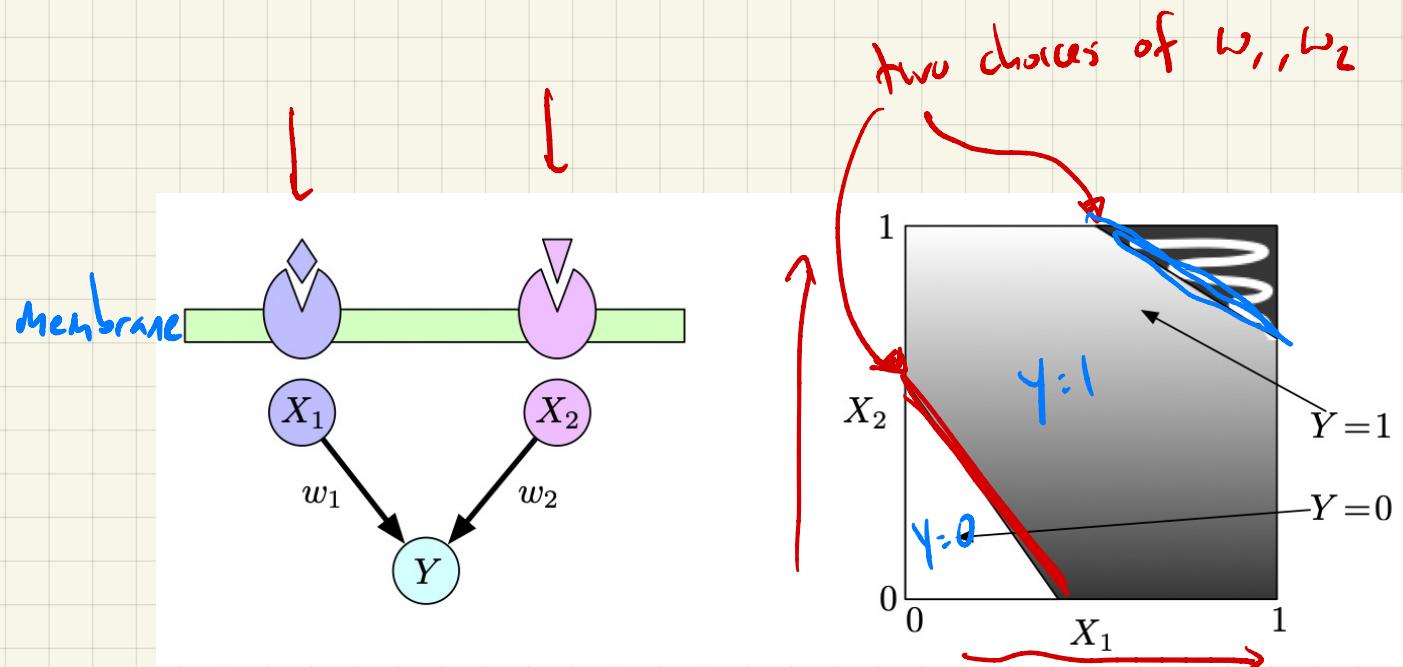
- 1) more accurate!
- 2) dynamics (can be complicated)

Disadvantages

- 1) need rate constants (experiments)
- 2) need to simulate dynamics (slow)



Simple linear circuit



$$y = \begin{cases} 1 & \text{if } X_1\omega_1 + X_2\omega_2 > 1 \\ 0 & \text{otherwise} \end{cases}$$

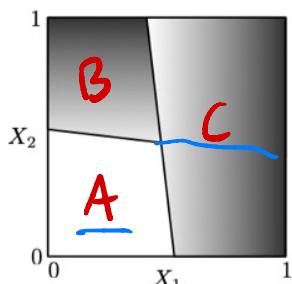
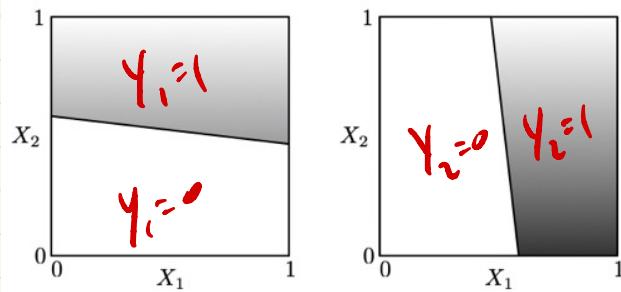
• What cellular "function" does this motif compute?

Is it an AND gate?

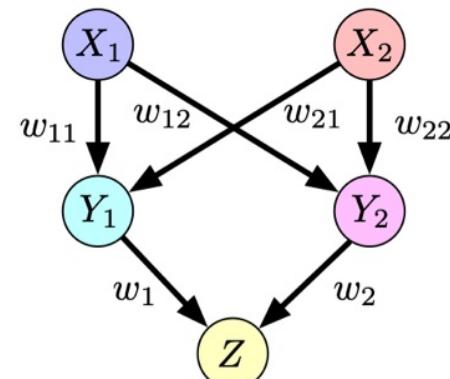
Combining linear circuits

→ put two in parallel

Circuit 1

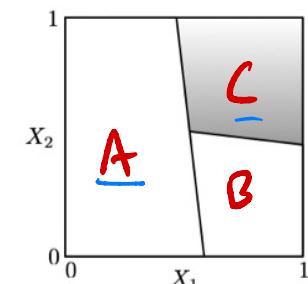
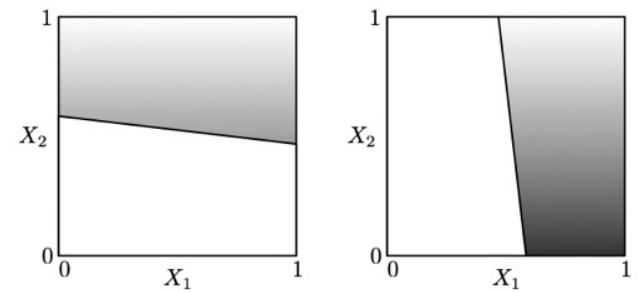


$$w_1 > 1, w_2 > 1$$



two different operations on $\{Y_1, Y_2\}$

Circuit 2



$$w_1 < 1, w_2 < 1$$

$$w_1 + w_2 > 1$$

- What's going on in the input space regions A, B, C ?

Advantages

- 1) no rate equations (= no rate constants)
- 2) everything is linear
- 3) rich space of computable functions

$$x_1 w_1 + x_2 w_2 + \dots$$

Disadvantages

- 1) still need weights (= binding strength)
- 2) only know these for small subset of PPIN
- 3) only decent model for small PPIN modif., not big networks

Boolean networks

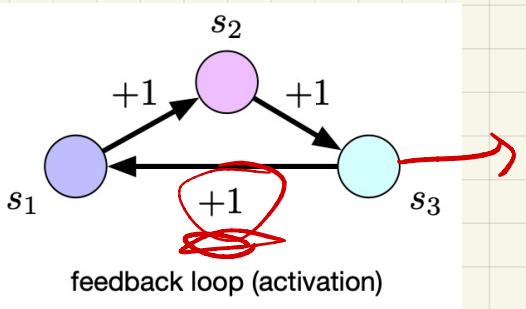
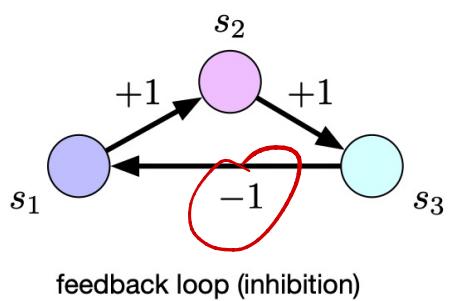
* EVEN SIMPLER *

$$s_i \in \{0, 1\}$$

$$s_i(t+1) = \begin{cases} 1 & \text{if } \sum_{j=1}^n A_{ij} s_j(t) + c > 0 \\ 0 & \text{otherwise} \end{cases}$$

activation threshold

a simple example ($c=0$)



• $s(0) = 000$. is an "attractor" or fixed point

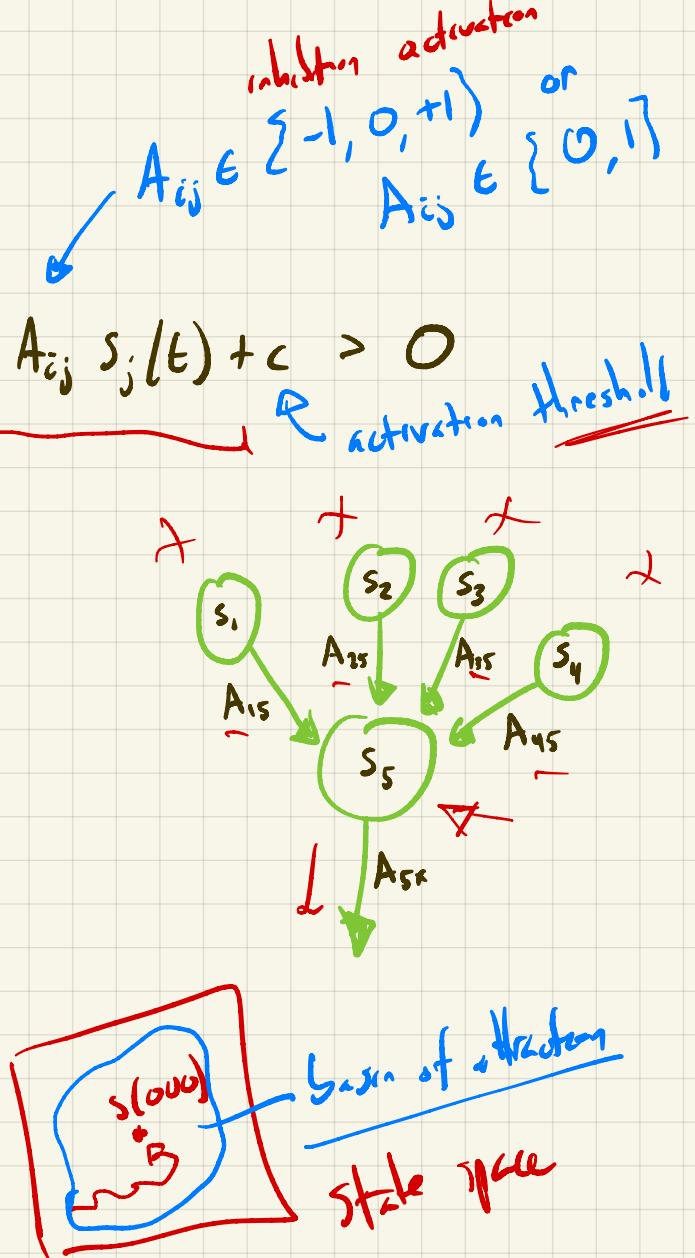
• $s(0) = 100$ [inhibition]

$$100 \rightarrow 010 \rightarrow 001 \rightarrow 000$$

• $s(0) = 100$ [activation]

$$100 \rightarrow 010 \rightarrow 001 \rightarrow 100$$

limit cycle

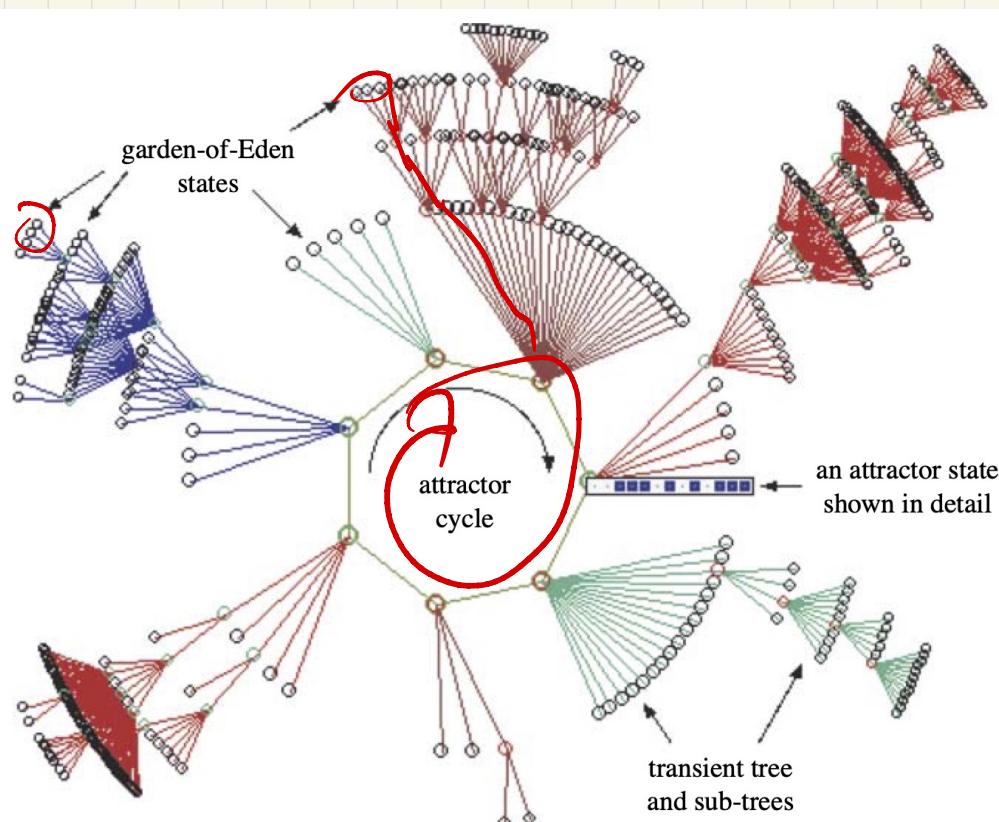
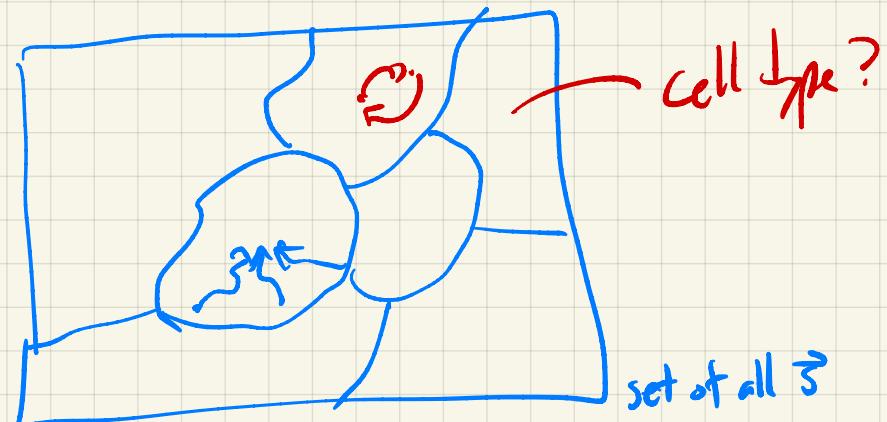


Basins of attraction of random boolean networks

$$s_i(t+1) = \begin{cases} 1 & \text{if } \sum_{j=1}^n A_{ij} s_j(t) + c > 0 \\ 0 & \text{otherwise} \end{cases}$$

random graph: $G(n,p)$

cumulate all possible $s(\omega)$
evolve forward



a complicated circuit

The yeast cell-cycle network is robustly designed

Fangting Li*†, Tao Long*†, Ying Lu*†, Qi Ouyang*‡, and Chao Tang*§

*Centre for Theoretical Biology and Department of Physics, Peking University, Beijing 100871, China; and §NEC Laboratories America, 4 Independence Way, Princeton, NJ 08540

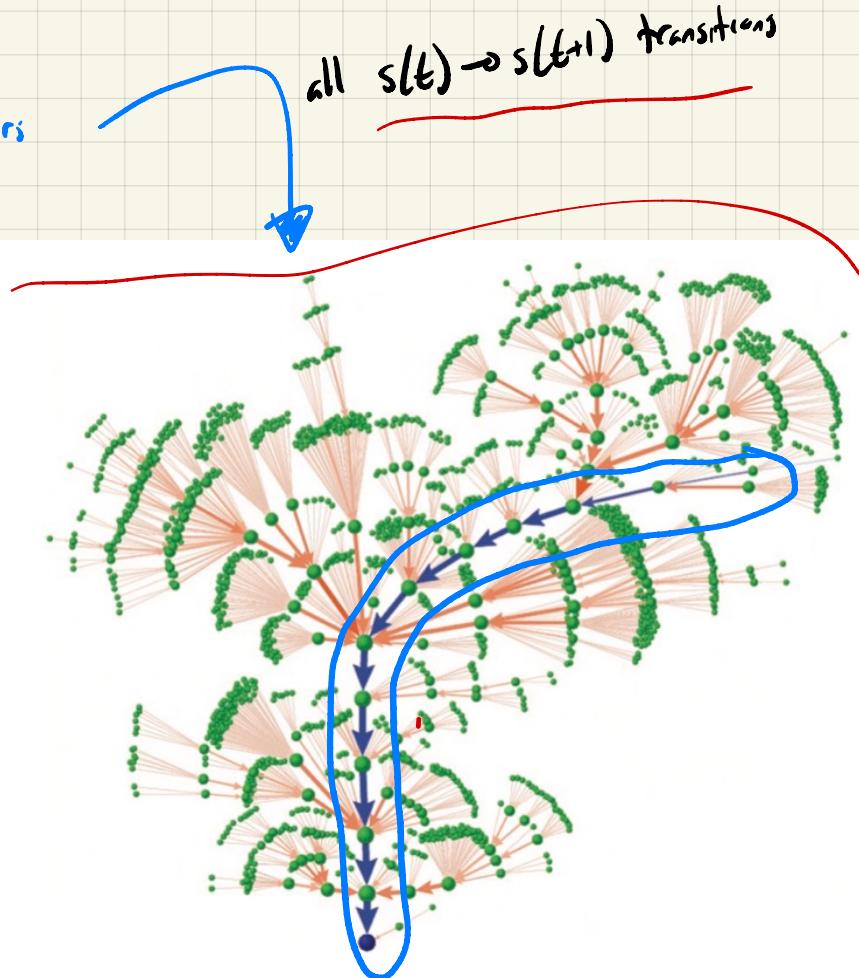
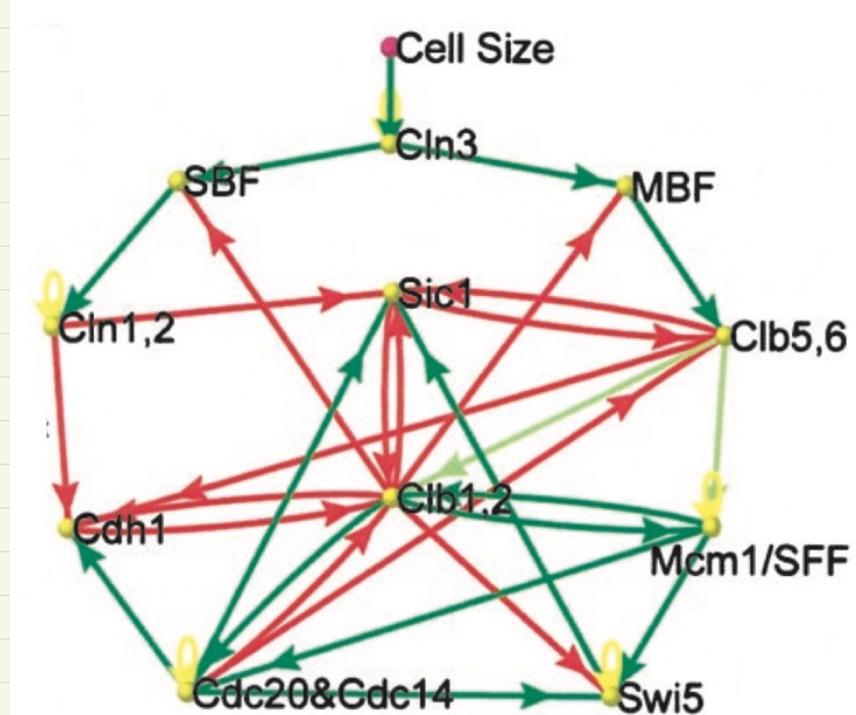
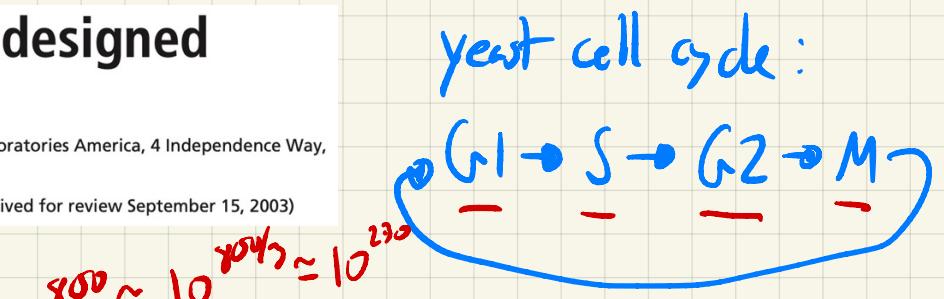
Edited by Peter G. Wolynes, University of California at San Diego, La Jolla, CA, and approved February 3, 2004 (received for review September 15, 2003)

PNAS 101(14), 4781 (2004)

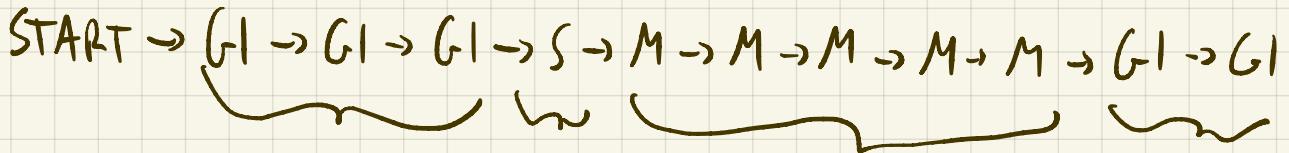
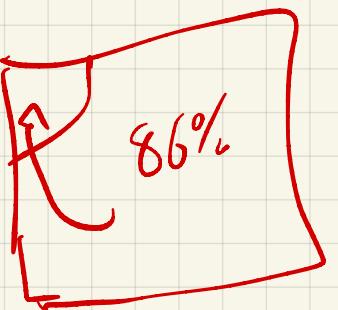
* 800 proteins involved

→ simple model on Π + "check point" state

→ Π nodes ≥ 2 $2^{\Pi} = 2^{800}$ possible state vectors



- 1784 (86%) of states in one component
- the "main" path of 13 states recapitulates the cell cycle:



- for 86% of all perturbations to \vec{s} , dynamics relax back to main cell cycle

\geq robustness!

- What's missing from this model? (what kind of questions can it not be used to answer?)

- how could we make this model better? (more realistic)

the structural compromise: counting motifs

to identify & count subgraphs that are computational building blocks (motifs)

A motif is a sub graph that appears more often than expected.

↓
Comparison ↓
null model (random graphs)

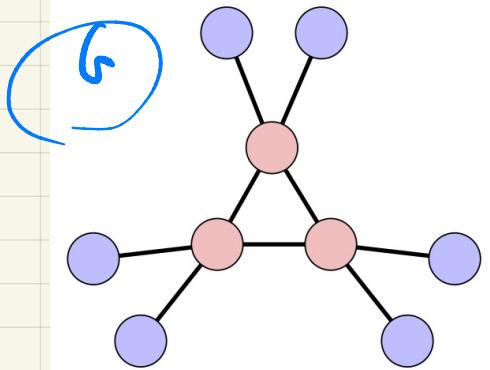
how to count subgraphs?

what counts as different

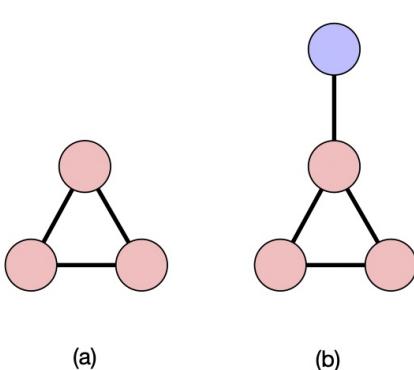
① non-identical subgraphs: G_1 and G_2 differ by at least one edge or
at least one node

② edge-disjoint subgraphs: if node $i \in G_1$ then $i \notin G_2$

example



motifs



$\boxed{C_2 \cup C_3}$
 $G_{2,3}$ (connected)

① non-identical subgraphs:

(a) 1
AND
(b) 6

lists of alg for
counting.

② edge-disjoint subgraphs:

(a) 1 0
OR
(b) 0 1

let G_k be a motif of size k nodes

assume f algorithm that can count: $f(G_k) = \# \text{ of } G_k \text{ in } G$

Does $f(G_k)$ occur more often than we expect?

→ solution: z-scores!

$$z\text{-score} = \frac{f(G_k) - E[f(G_k)]_{\text{null}}}{\sqrt{V[f(G_k)]_{\text{null}}}}$$

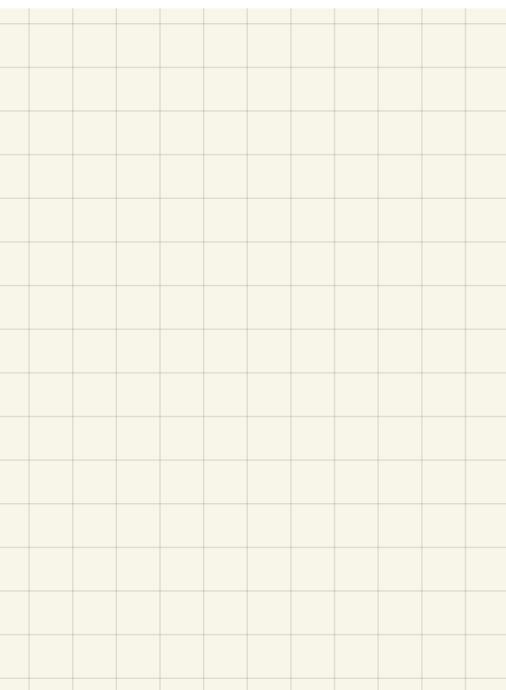
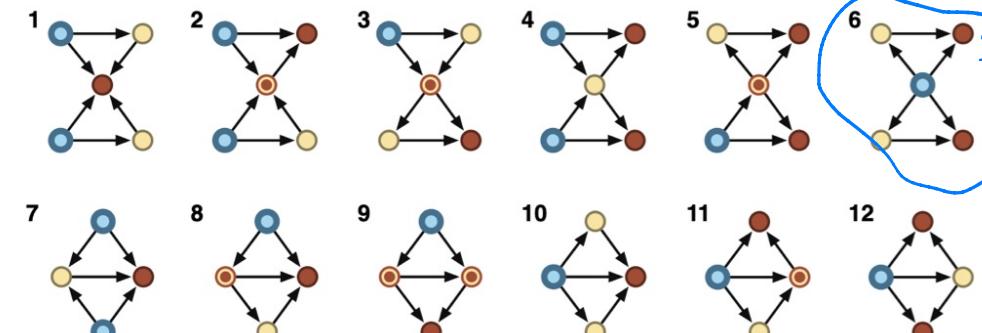
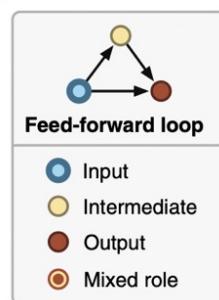
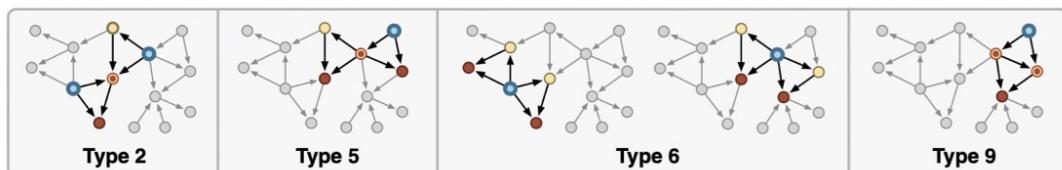
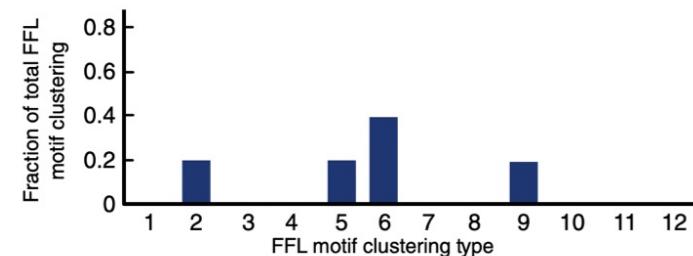
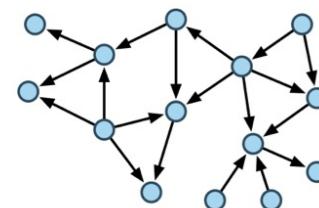
Erős-Renyi
Chung-Lu
SBM
DC-SBM
etc.

Random graph toolbox

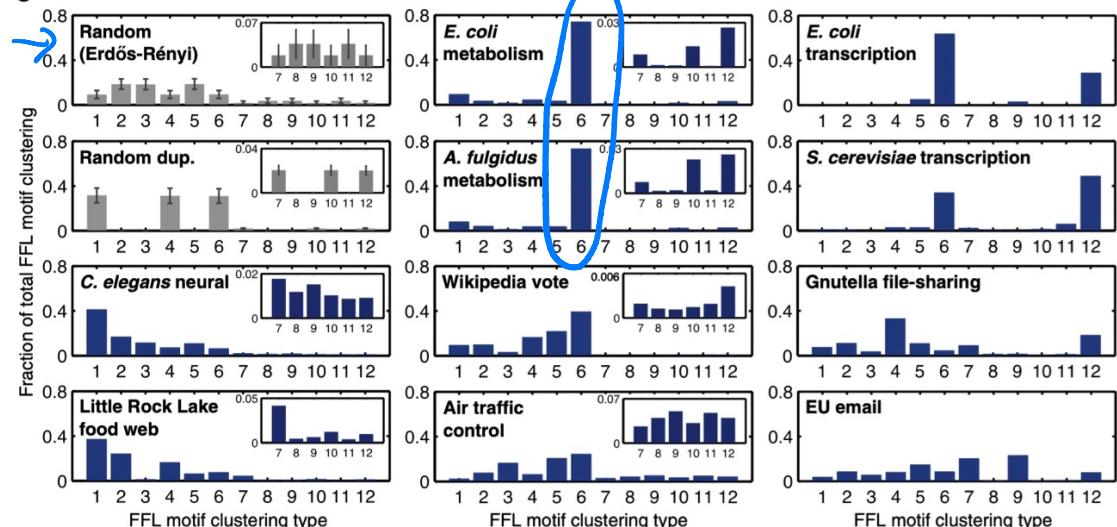
- $z=0$ (or "close" to it) → no more / less than expected
- $z > 0$ (by some margin) → G_k is over represented in G
- $z < 0$ (by some margin) → G_k is suppressed in G

NETWORK SCIENCE

Organization of feed-forward loop motifs reveals architectural principles in natural and engineered networks

Thomas E. Gorochowski,^{1,2*} Claire S. Grierson,^{1,2†} Mario di Bernardo^{1,3,4†}**A****B**

Null model

C

Motifs of interest