Protein interaction networks : structure

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

DNA → RNA → protein

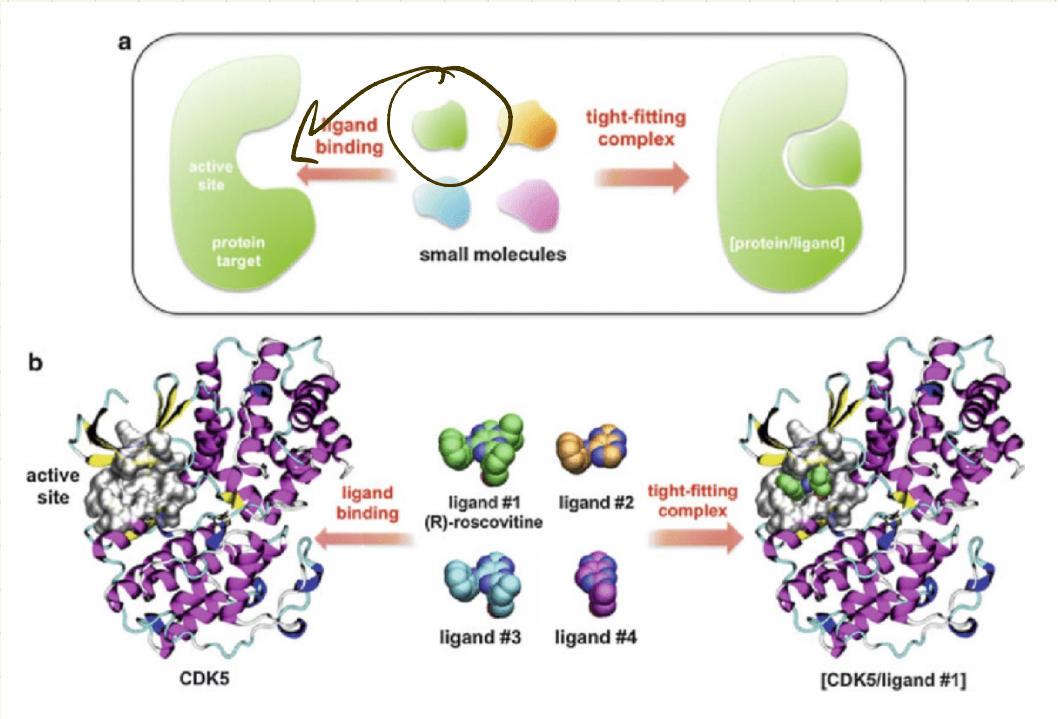
primary → secondary → tertiary structure

amino acid sequence — helices & sheets — 3D shape

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

Mouse : ~55k proteins

Amoeba : ~14k proteins



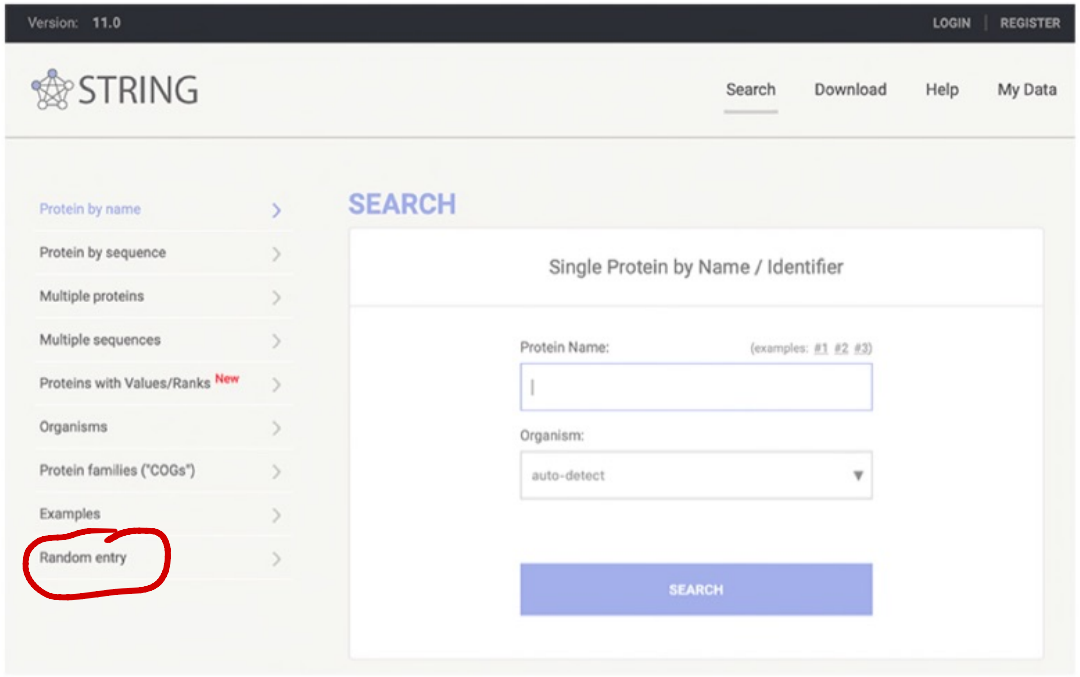
two types of binding

1. stable binding (building infrastructure)
2. transient binding (information flow)

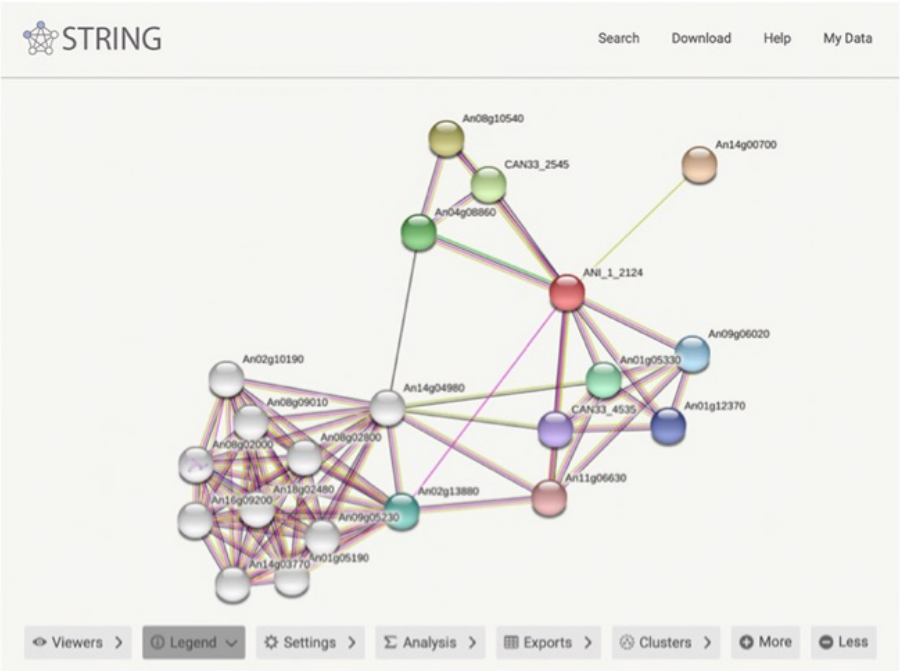
protein interaction databases: STRING

v11
covers 10,000,000 proteins
across 2000+ organisms

HTTPS://string-db.org/

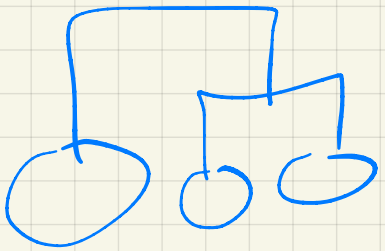


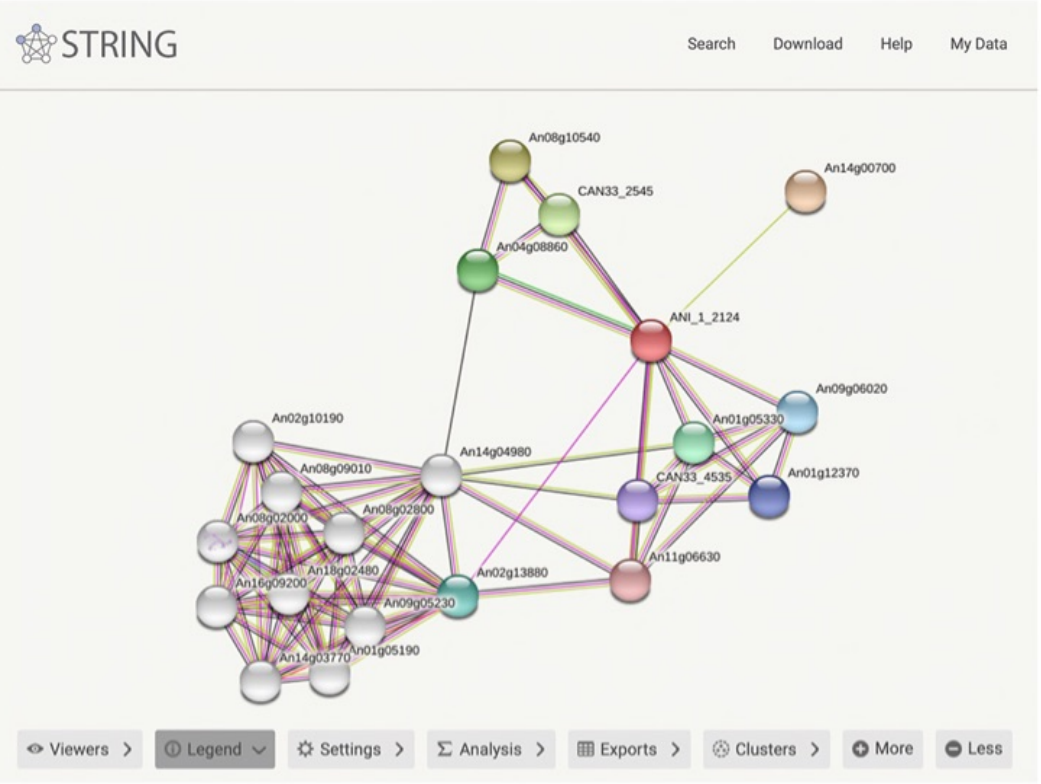
Random entry: ANI_1_2124



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

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





rich metadata! but very inconsistent coverage

node annotations

edge annotations

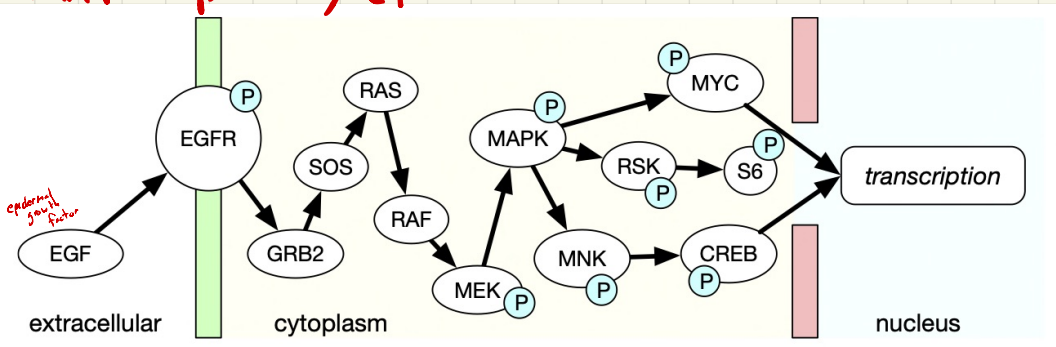
some experimentally validated data

some model-based predicted data

Zooming in: Signaling pathways

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

MAPK pathway (partial)



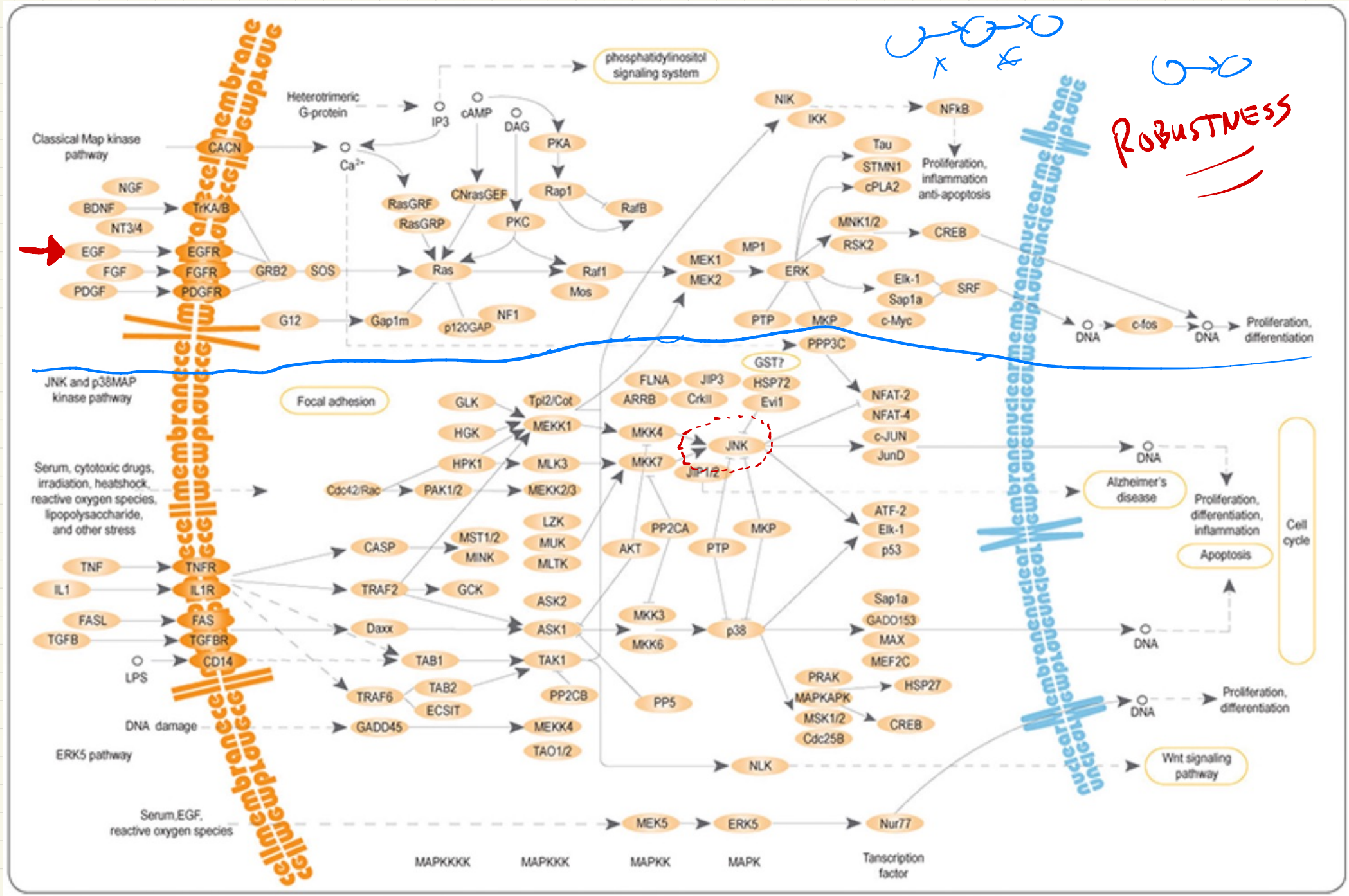
A → B activation

A ⊣ B inhibition

changes to MAPK: cancer
inflamatory diseases
obesity
diabetes

Zooming out a little: MAPK + WNT signalling

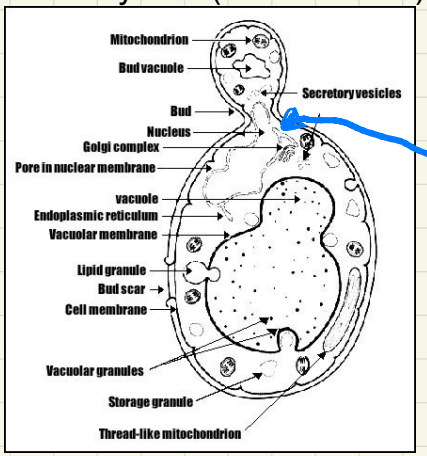
activation AND inhibition
wow crosstalk → why so much?



ROBUSTNESS

Where does the data come from?

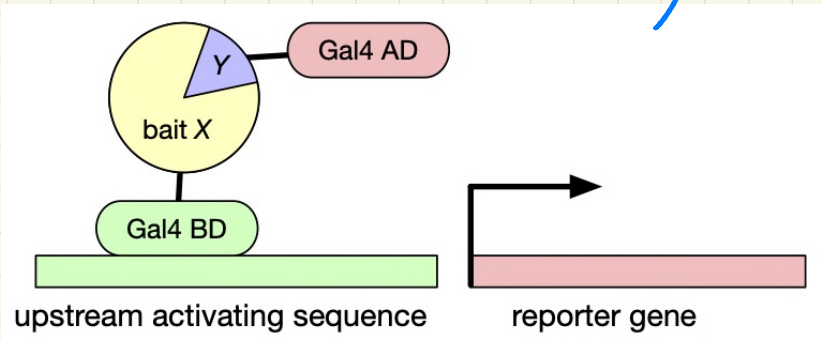
Bakers yeast (S. cerevisiae)



Experiments

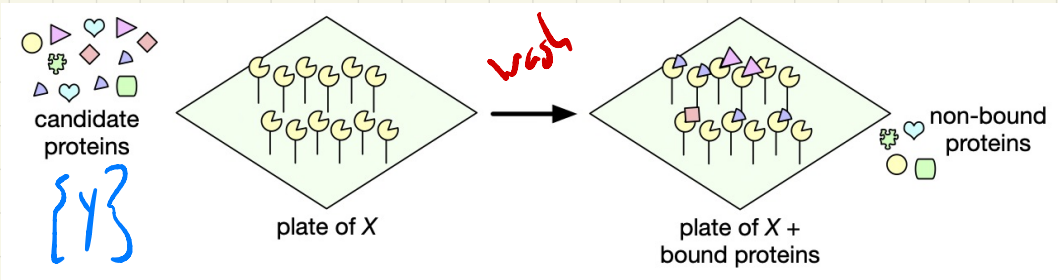
① yeast two-hybrid (Y2H)

- edge sampling (X)—(Y)
- insert (X)-(Y) inside an existing nuclear binding event
- limitations
  - nuclear protein only
  - high false positive rate
  - high false neg rate



② affinity purification + mass spectrometry (AP-MS)

- node sampling (X)
- wash {Y} over X
  - high FP
- limitations
  - high FN

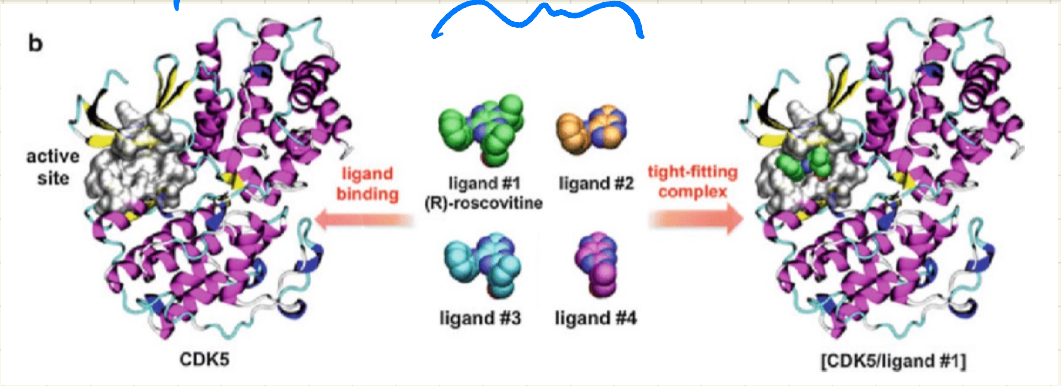


Models

3. text mining
   - use NLP tech to "read" scientific literature
   - manly low throughput experiments
   - limitations - false pos, false neg } complicated

4. predictions from protein domain analysis
   - secondary / tertiary structure
     - shape + location of binding domains
   - limitations - need structure
     - FP
     - FN

X {Y}



error rates

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

$\Theta(n)$ TP
$\Theta(n^2)$ TN

- $n \approx 10k$ (canonical proteins)
  - → 100,000,000 possible interactions most of which never occur
    - ∴ why?

high false positive rates →

high false negative rates →

TP: 0.99

TN: 0.99

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

FN: 100

$0.99 \times (100{,}000{,}000 - 10{,}000) = 98{,}990{,}100$

FP: 1,000,000

the reality of PPIN data → messy : lots of FP

1) binary interactions only ⇒ 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.

4) systems biology
what does G tell us about life?
complexity?
evolution?
robustness?

tools

1) motifs (small subgraphs)

- decompose complexity into smaller units
- FFLN = 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

simplicity (tractability) ↓

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

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

boolean networks
- approx linear models w/ simple boolean logic

---

rate equations:

$$2I^- + H_2O_2 + 2H^+ \rightarrow I_2 + 2H_2O$$

iodine · hydrogen peroxide · protons · molecular iodine · water

$$\text{rate} = k[I^-][H_2O_2] = \frac{\Delta[\text{things}]}{\Delta t}$$

* logistic equation solution

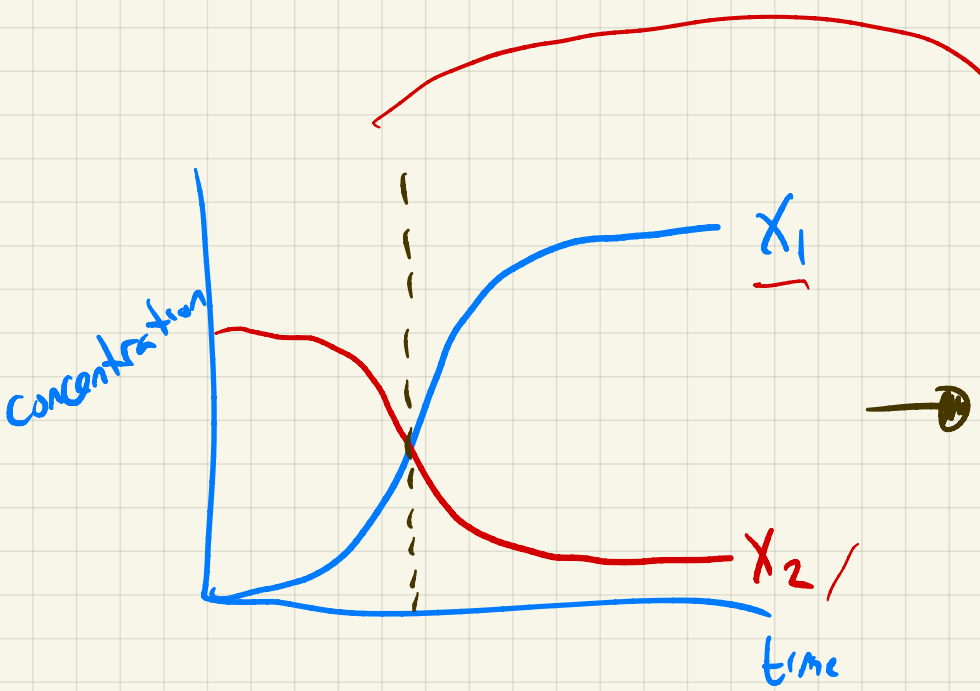
← exponential increase (recall SIR)

Advantages

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

Disadvantages

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



rate eq'tns

$X_1 \in [0,1]$

$X_2 \in [0,1]$

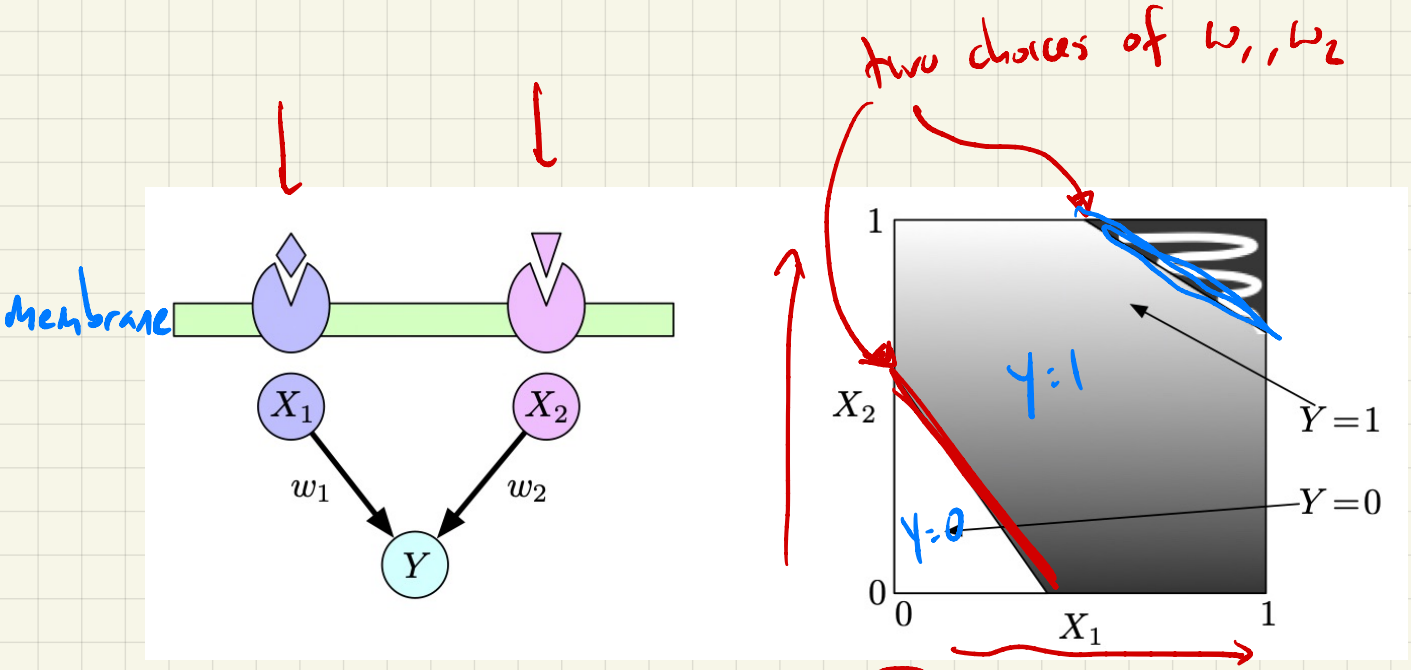
linear model

$X_1 \in \{0,1\}$

$X_2 \in \{0,1\}$

boolean logic

Simple linear circuit



$$y = \begin{cases} 1 & \text{if } X_1 w_1 + X_2 w_2 > 1 \\ 0 & \text{otherwise} \end{cases}$$

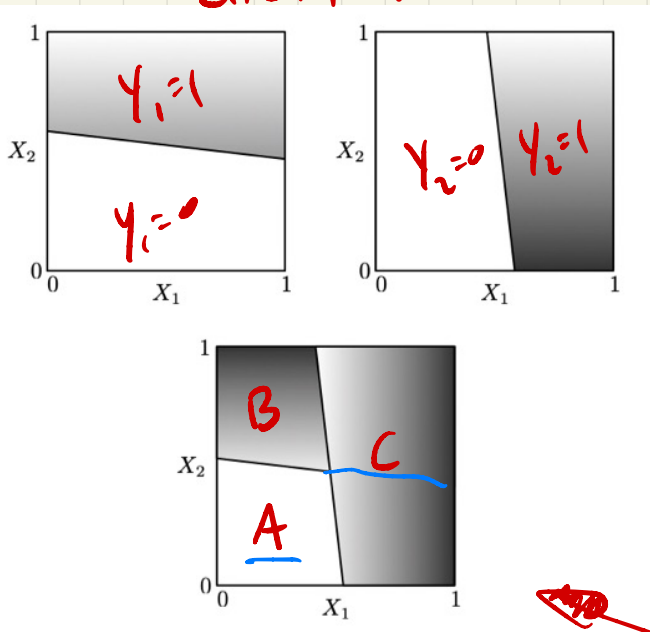
- What cellular "function" does this motif compute?

  Is it an AND gate?

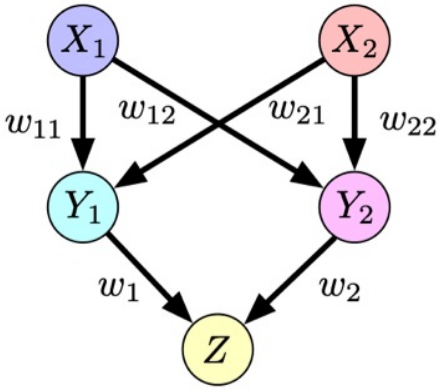
Combining linear circuits

→ put two in parallel

circuit 1

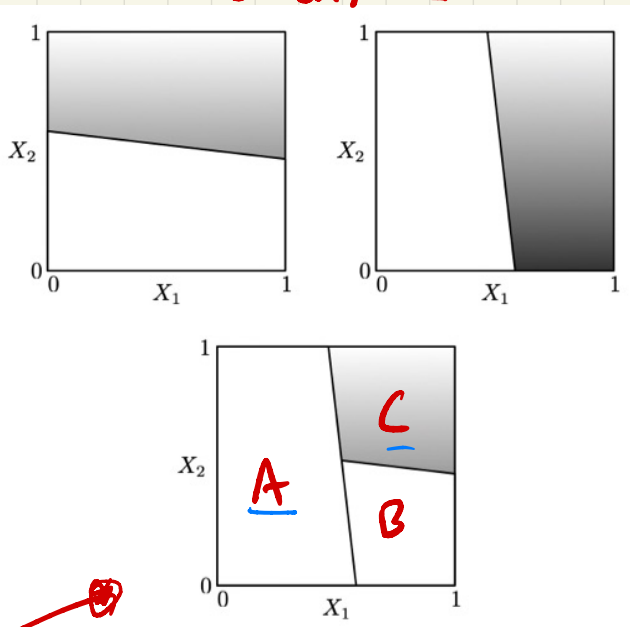


$w_1 > 1, w_2 > 1$



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

$$x_1 w_1 + x_2 w_2 + \cdots$$

Disadvantages

1) still need weights (= binding strength)

2) only know these for small subset of PPIN

3) only decent model for small PPIN motifs, not big networks

Boolean networks

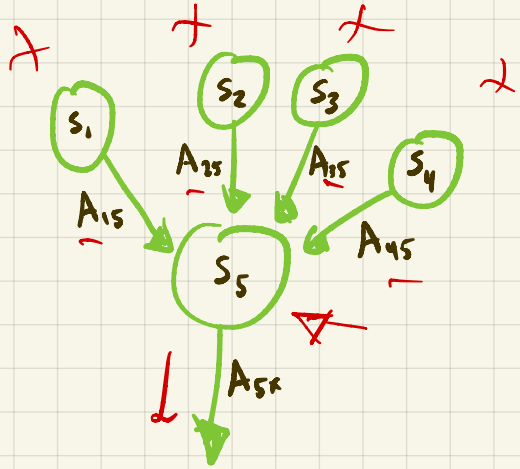
* EVEN SIMPLER *

$A_{ij} \in \{-1, 0, +1\}$ (inhibition, activation) or $A_{ij} \in \{0, 1\}$

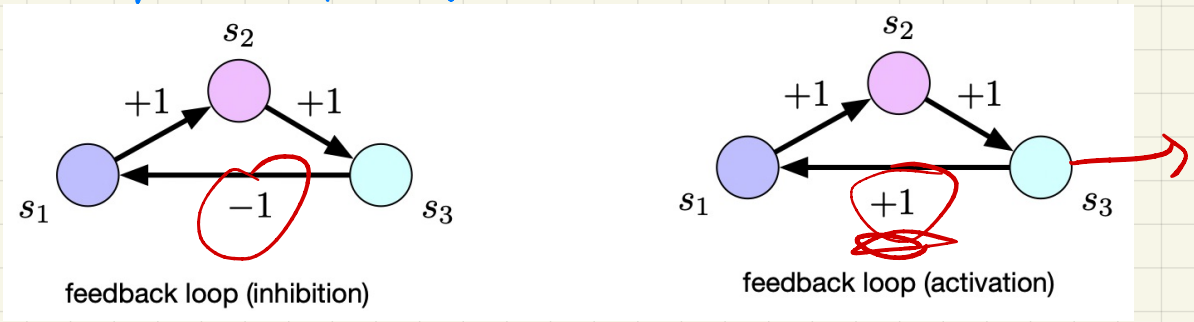
$s_i \in \{0, 1\}$

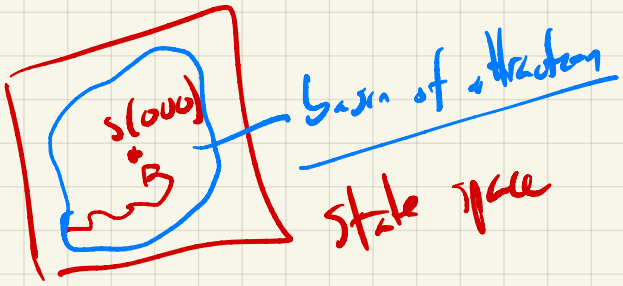
$$s_i(t+1) = \begin{cases} 1 & \text{if } \sum_{j=1}^{n} A_{ij}\, s_j(t) + c > 0 \\ 0 & \text{otherwise} \end{cases}$$

($\sum_{j=1}^{n} A_{ij}\, s_j(t)$: activation; threshold)



a simple example ($c = 0$)





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

- $s(0) = 100$ [inhibition]

$100 \to 010 \to 001 \to 000$

- $s(0) = 100$ [activation]

$100 \to 010 \to 001 \to 100$

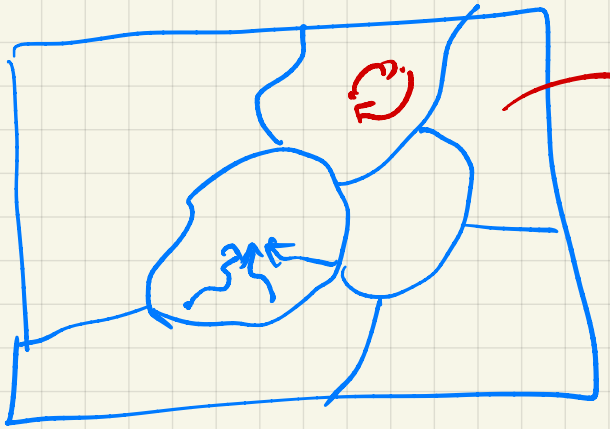
limit cycle

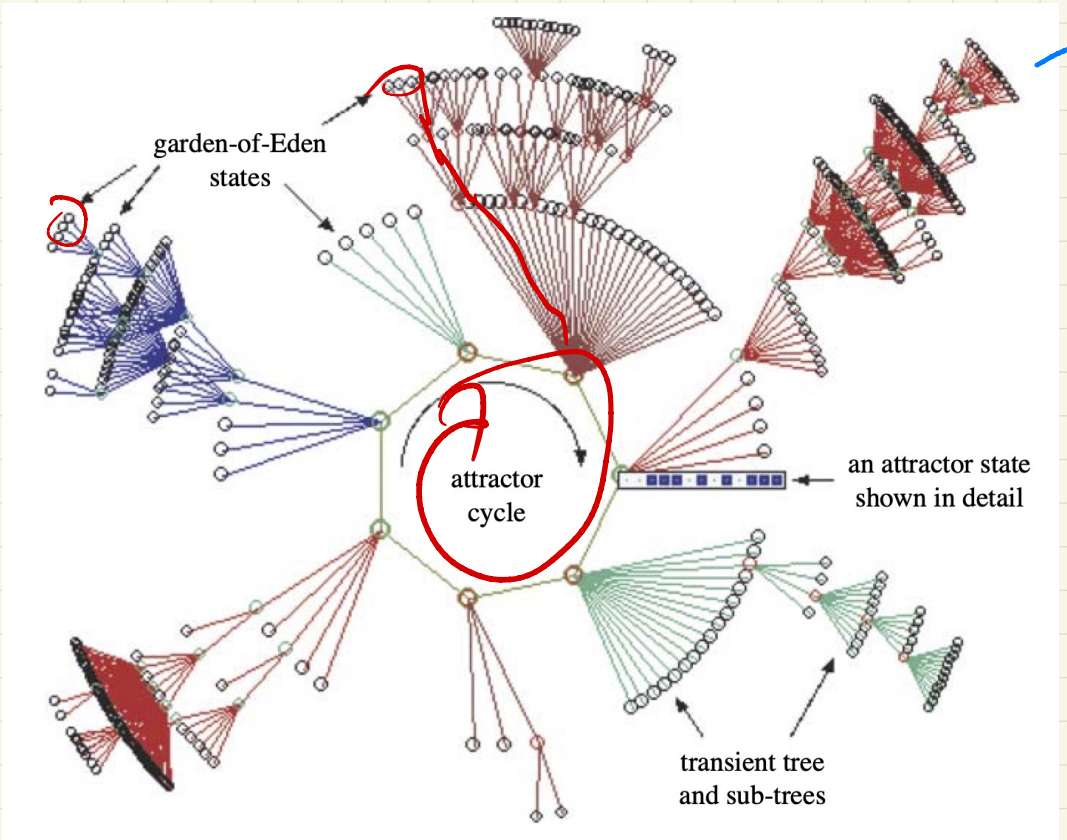
basins of attraction of random boolean networks

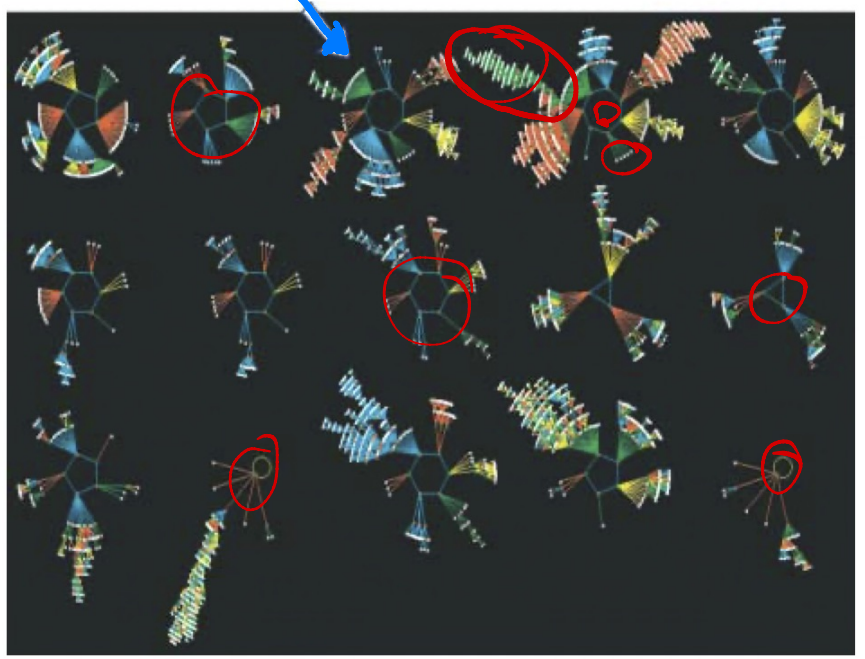
$$s_i(t+1) = \begin{cases} 1 & \text{if } \sum_{j=1}^{n} A_{ij}\, s_j(t) + c > 0 \\ 0 & \text{otherwise} \end{cases}$$

random graph: $G(n,p)$

enumerate all possible $s(\cdot)$
evolve forward







Wuensche (1994) Artificial Life III, vol XVII, p465-501

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)

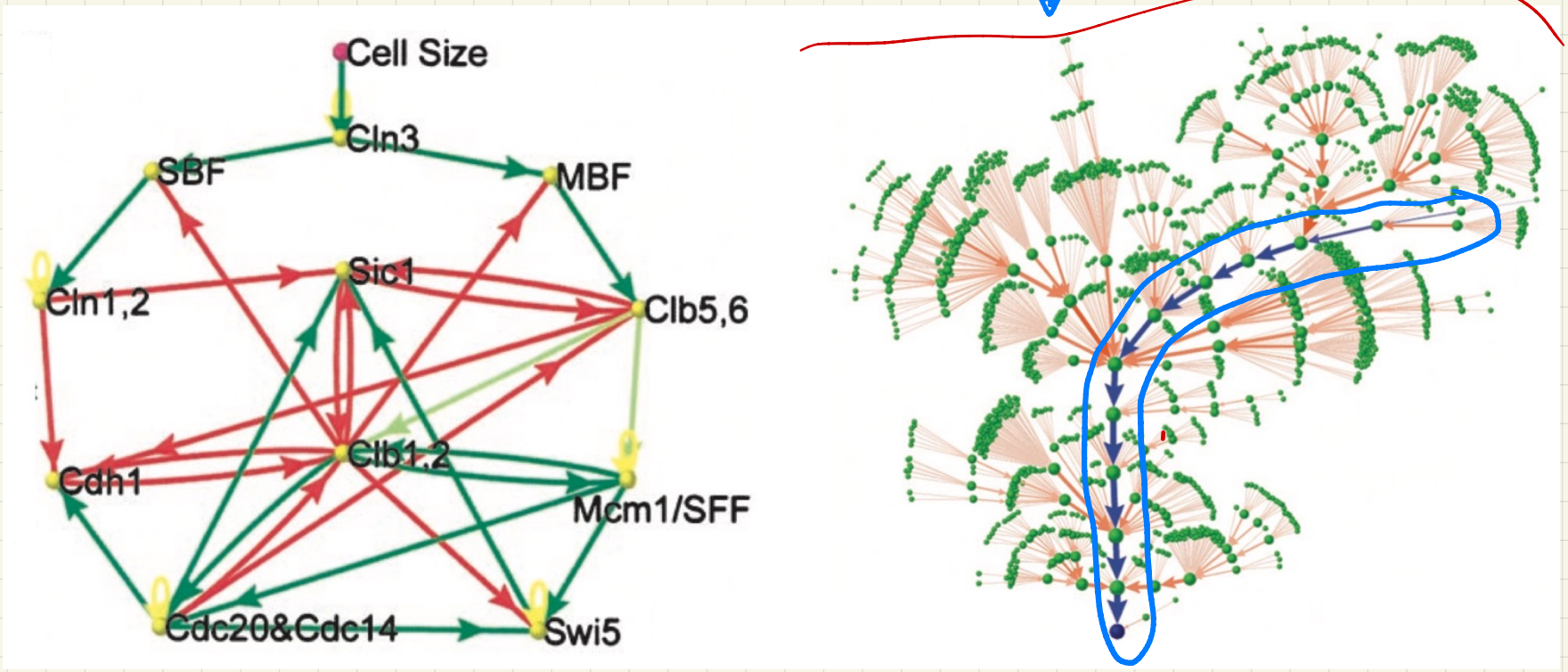
yeast cell cycle:

G1 → S → G2 → M

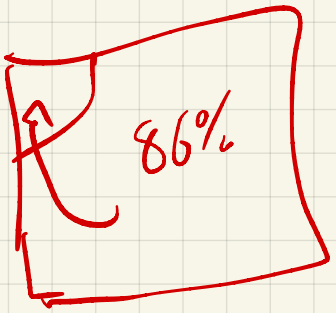
$2^{800} \simeq 10^{800 \log_{10} 2} \simeq 10^{230}$

* 800 proteins involved
  - → simple model on 11 + "check point" state
  - → 11 nodes ⇒ $2^{11} = 2048$ possible state vectors

all s(t) → s(t+1) transitions



- 1784 (86%) of states in one component



- the "main" path of 13 states recapitulates the cell cycle:

START → G1 → G1 → G1 → S → M → M → M → M → M → G1 → G1

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

⇒ 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?

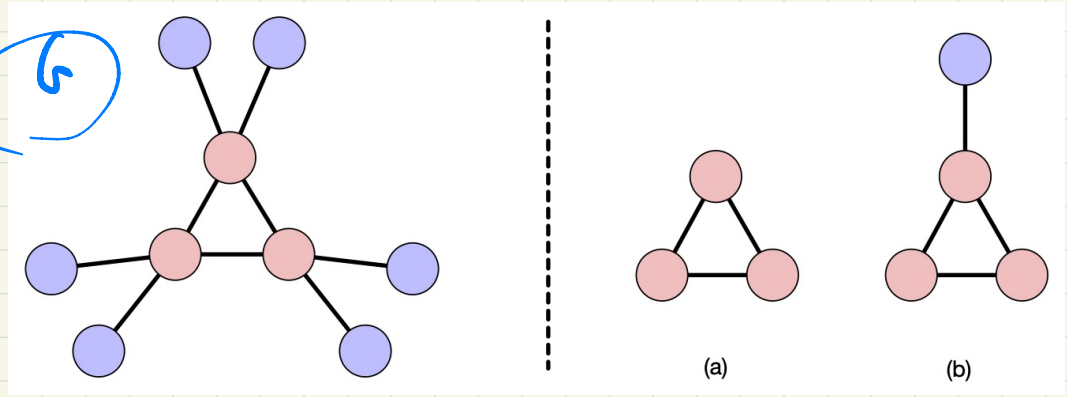
what counts as different

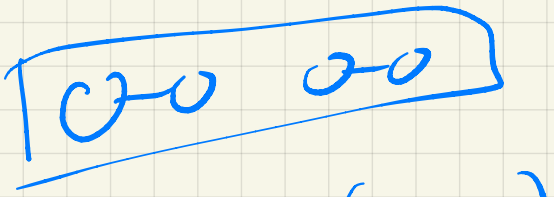
∀ ① non-identical subgraphs: G_1 and G_2 differ by at least one edge or at least one node

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

example

motifs





G_{2m} (connected)

lots of algs for counting

(1) non-identical subgraphs: (a) 1 AND (b) 6

(2) edge-disjoint subgraphs: (a) 1 OR 0 (b) 0 OR 1

let G_k be a motif of size k nodes

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

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

→ solution: z-scores!

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

Erdös-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

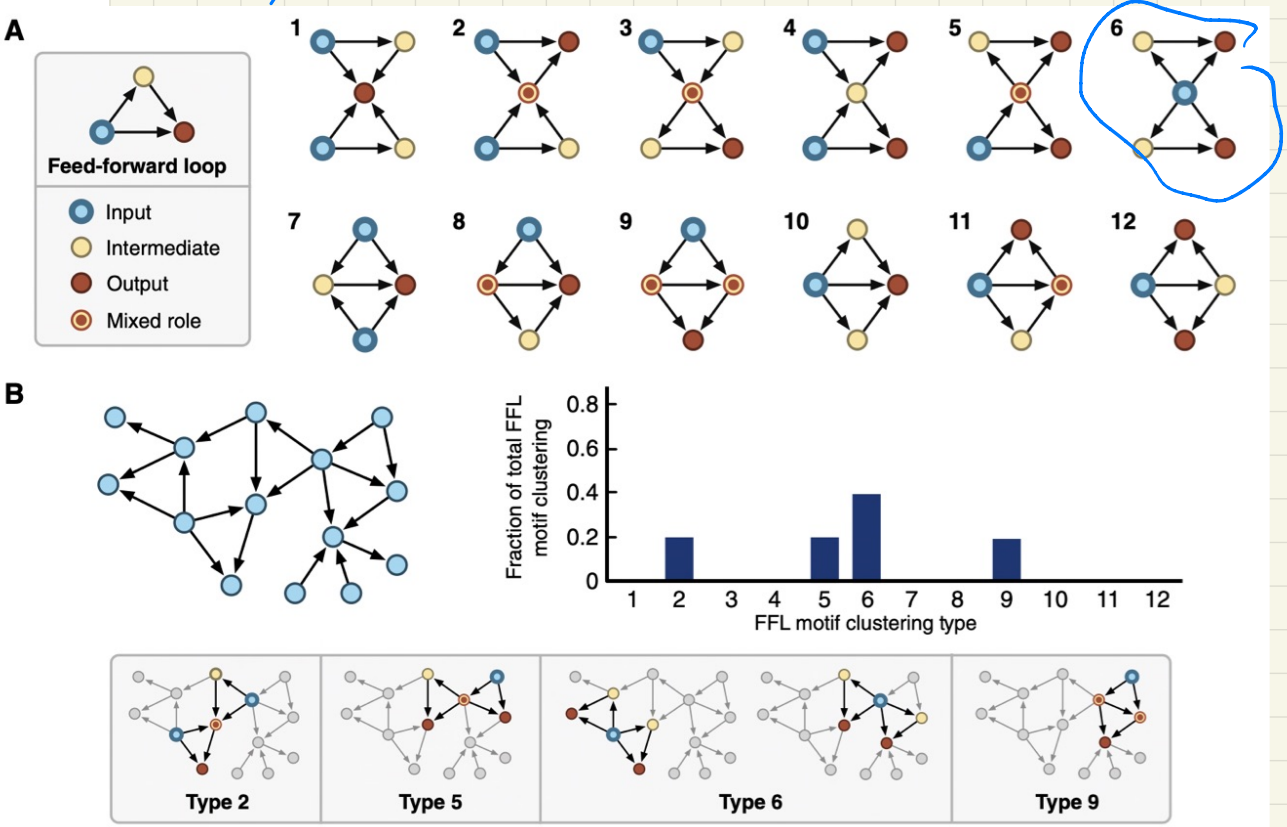
SCIENCE ADVANCES | RESEARCH ARTICLE

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†]

motifs of interest



null model