

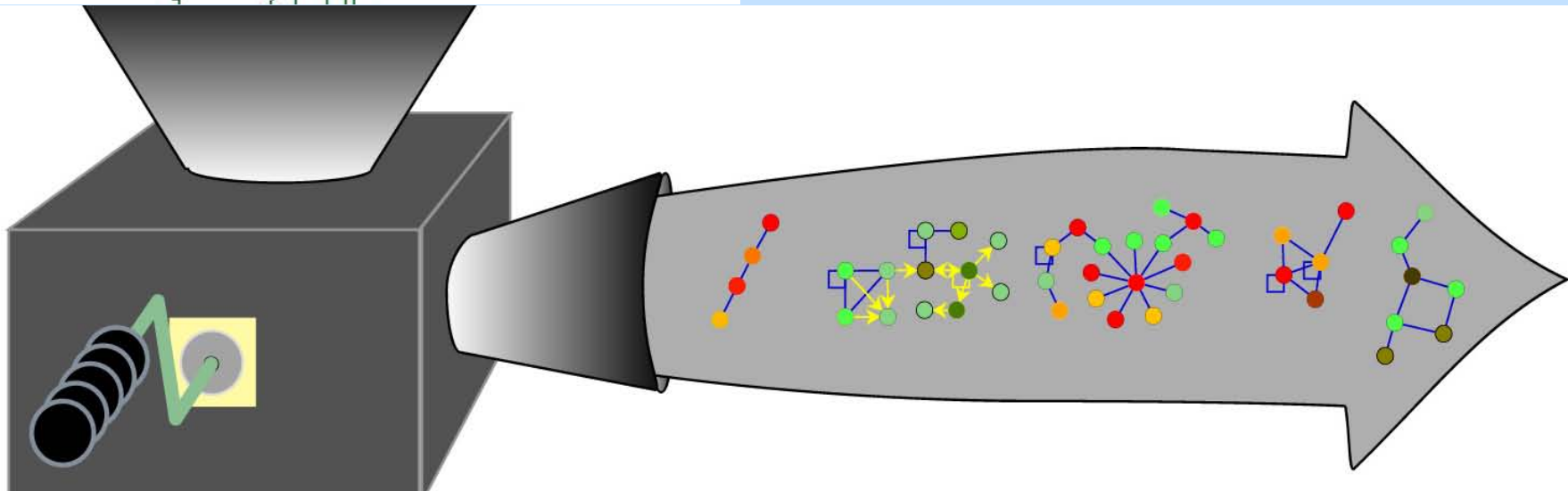
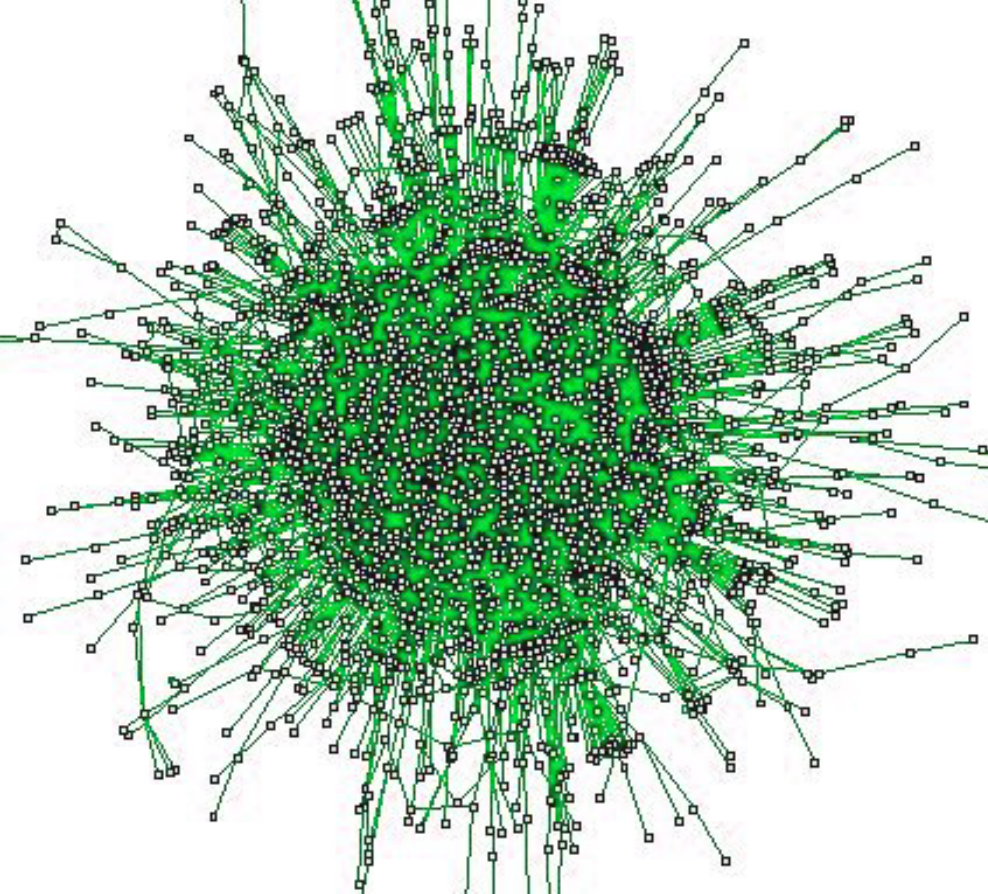
CG 2011-12

Module Identification

Gene/Protein Modules

- A *module* is a set of genes/proteins performing a distinct biological function.
- Characterized by a coherent behavior of its genes w.r.t. a certain biological property.
- Examples:
 - *transcriptional module*: a set of co-expressed genes sharing a common function.
 - *protein complex*: assembly of proteins that build up some cellular machinery.
 - *signaling pathway*: a chain of interacting proteins propagating a signal in the cell.

Distilling Modules from Networks



NetworkBLAST:
S. et al., JCB & PNAS 2005

Scoring

- Based on a likelihood ratio score.
- Protein complex model: edges occur indep. with high probability p .
- Random model: degree-preserving. Probability of edge $p(u, v)$ depends on degrees of proteins u, v .

$$C = (V', E')$$

$$L(C) = \prod_{(u,v) \in E'} \frac{p}{p(u,v)} \prod_{(u,v) \notin E'} \frac{1-p}{1-p(u,v)}$$

The Search Algorithm

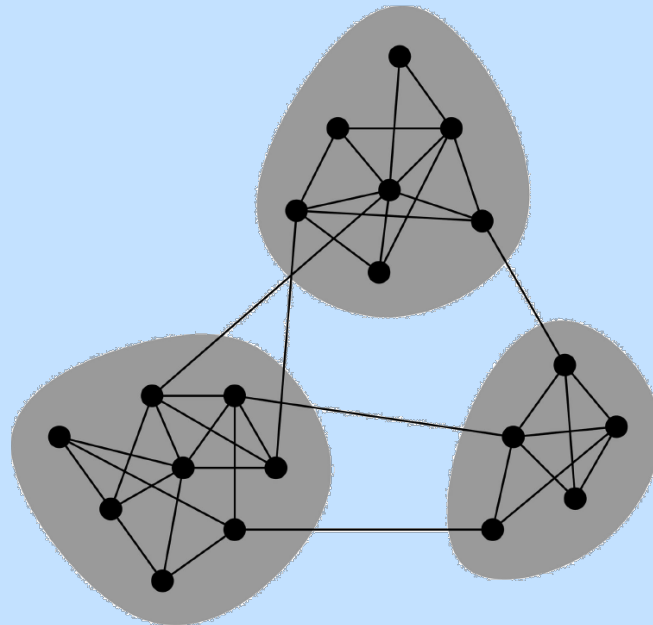
Algorithm:

- Construct small seeds around each vertex.
- Expand seeds using local search.
- Filter overlapping subgraphs.

While local-based, often detects the highest scoring subgraphs.

Modularity and Community Structure in Networks

M.E.J Newman, PNAS 2006



Modularity of a division (Q)

$Q = \#(\text{edges within groups}) - E(\#(\text{edges within groups in a RANDOM graph with same node degrees}))$

Trivial division: all vertices in one group
 $\implies Q(\text{trivial division}) = 0$

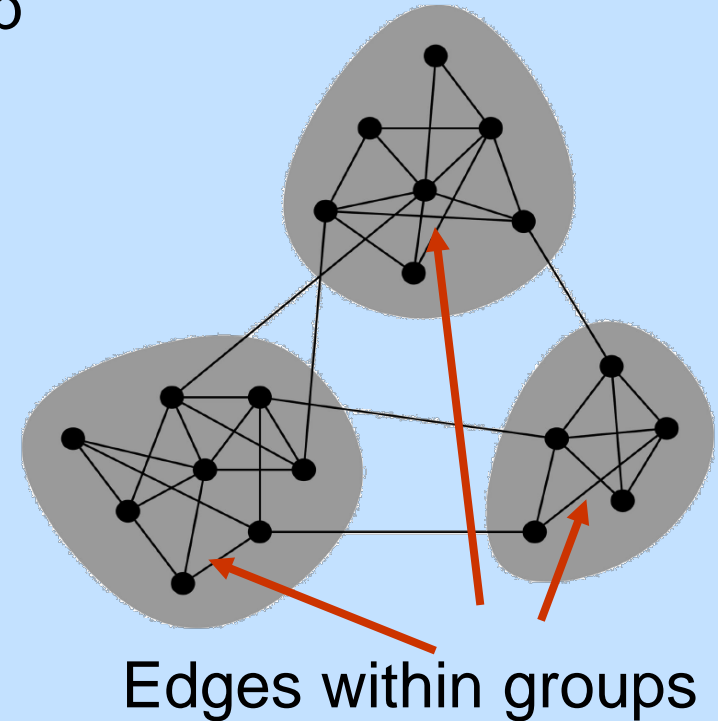
k_i = degree of node i

$M = \sum k_i = 2|E|$

$A_{ij} = 1$ if $(i,j) \in E$, 0 otherwise

E_{ij} = expected number of edges between i and j in a random graph with same node degrees.

Lemma: $E_{ij} \approx k_i * k_j / M$



$$Q = \sum (A_{ij} - k_i * k_j / M \mid i, j \text{ in the same group})$$

Division into two groups

$$Q = \sum (A_{ij} - k_i k_j / M \mid i, j \text{ in the same group})$$

- Suppose we have n vertices $\{1, \dots, n\}$
- \mathbf{s} - $\{\pm 1\}$ vector of size n .

Represent a 2-division:

- $s_i == s_j$ iff i and j are in the same group
- $\frac{1}{2} (s_i s_j + 1) = 1$ if $s_i == s_j$, 0 otherwise

$$\bullet \implies Q = \frac{1}{2} \sum_{i,j} \left(A_{ij} - \frac{k_i k_j}{M} \right) (s_i s_j + 1)$$

Division into two groups (2)

$$Q = \frac{1}{2} \sum_{i,j} \left(A_{ij} - \frac{k_i k_j}{M} \right) (s_i s_j + 1)$$

Since $\sum_{i,j} A_{ij} = \sum_i k_i = M$

$$Q = \frac{1}{2} \sum_{i,j} \left(A_{ij} - \frac{k_i k_j}{M} \right) s_i s_j$$

B = the modularity matrix
- symmetric

$$Q = \frac{1}{2} \mathbf{s}^T \mathbf{B} \mathbf{s}$$

where

$$B_{ij} = A_{ij} - \frac{k_i k_j}{M}$$

Division into two groups (3)

B is symmetric \Rightarrow **B** is diagonalizable (real eigenvalues)

B's eigenvalues

$$\beta_1 \geq \beta_2 \geq \dots \geq \beta_n$$

B's orthonormal eigenvectors

$$\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_n$$

$$\mathbf{B}\mathbf{u}_i = \beta_i \mathbf{u}_i$$

$$Q = \frac{1}{2} \mathbf{s}^T \mathbf{B} \mathbf{s} \quad \longrightarrow \quad Q = \frac{1}{2} \sum_i \beta_i a_i^2$$

$\left[\mathbf{s} = \sum_i a_i \mathbf{u}_i \right]$

where $a_i = \mathbf{u}_i^T \cdot \mathbf{s}$

- Which vector \mathbf{s} maximizes Q ?
 - clearly $\mathbf{s} \sim \mathbf{u}_1$ maximizes Q , but \mathbf{u}_1 may not be $\{\pm 1\}$ vector
 - Greedy heuristic: maximize the projection a_1 : choose $s_i = +1$ if $u_{i1} > 0$, $s_i = -1$ otherwise

Identifying protein pathways

Subgraph Isomorphism

Problem: Given a graph G and a pattern graph H , decide if G contains a *subgraph* isomorphic to H .

- NPC.
- Note that the subgraph is not required to be induced.
- Trivial algorithm runs in $O(n^k)$.
- We will be interested in graph classes that admit a *fixed parameter* algorithm (Downey & Fellows '92) – i.e., time is exponential in k but polynomial in n .

Finding Simple Paths

Problem: Given a graph $G=(V,E)$ and a parameter k , find a simple path of length k in G .

- NPC by reduction from Hamiltonian path.
- The hardness stems from requiring the path to be *simple*.

Color Coding [AYZ'95]

Problem: Given a graph $G=(V,E)$ and a parameter k , find a simple path with k vertices (length $k-1$) in G .

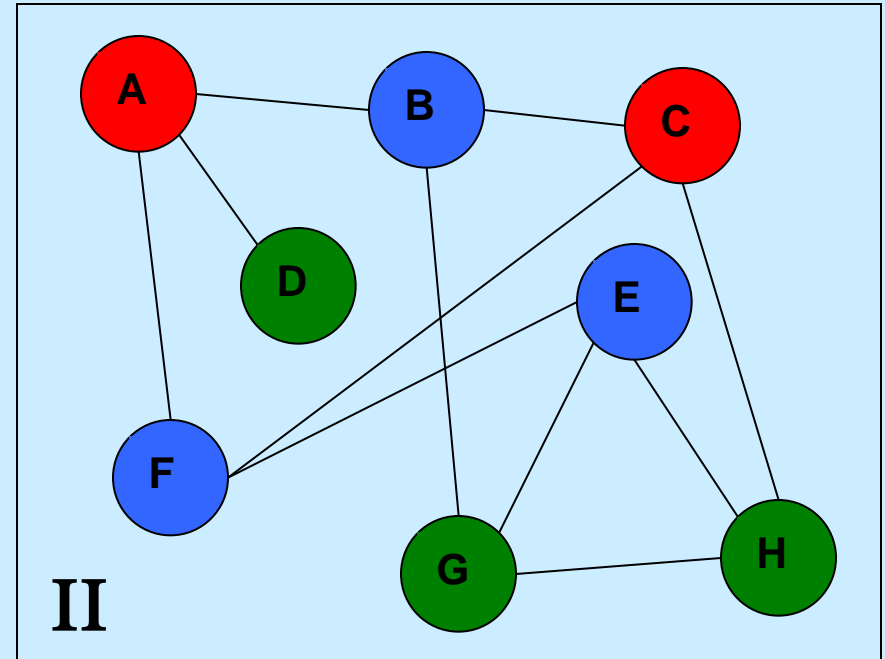
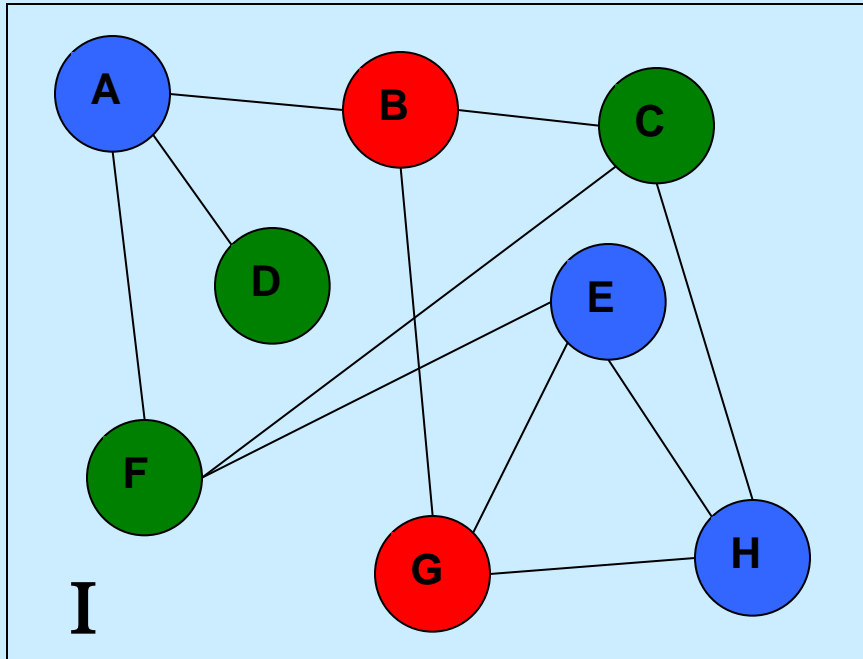
Algorithm: Randomly color vertices with k colors, and find a *colorful* path (distinct colors).

$$c : V \rightarrow [1, k]; S \in 2^{[1, k]}$$

$$P(v, S) = \max_{u:(u,v) \in E, c(u) \in S - \{c(v)\}} P(u, S - \{c(v)\})$$

Main idea: only 2^k color subsets vs. n^k node subsets.

Coloring Example



- Two different colorings on toy graph, $k=3$
- In coloring **I**, $P(A,RGB)$ is built $C \rightarrow BC \rightarrow ABC$
- In coloring **II**, $P(A,RGB)$ is built $G \rightarrow BG \rightarrow ABG$
- ABC is not colorful in coloring **II**

Randomization Analysis

- A colorful path is simple, but a simple path may not be colorful *under a given coloring*
- Solution: run multiple independent trials.
- After one trial:

$$\Pr(\text{Success}) = k! / k^k \geq 1 / e^k$$

Color Coding [AYZ'95]

Complexity:

- Space complexity is $O(2^k n)$.
- Colorful path found by DP in $O(km2^k)$.
- $O(e^k)$ iterations are sufficient.
- Overall time is $2^{O(k)}m$.
- Note that the exponential part involves the parameter only, that is, the problem is *fixed parameter tractable*.

Comparison of Running Times

| Path length | Color coding | Exhaustive |
|-------------|--------------|------------|
| 8 | 435 | 866 |
| 9 | 2,149 | 15,120 |
| 10 | 11,650 | -- |

- ~4500 vertices, ~14500 edges.

Biologically-Motivated Constraints

- Color-Coding gives an algorithmic basis, now introduce biologically motivated extensions.
- Can introduce edge weights (confidence).
- Can constrain the start or end of a path by type.
 - Steffen et al. '02: pathways from membrane to TF.
- Can force the inclusion of a specific protein on the path by ...

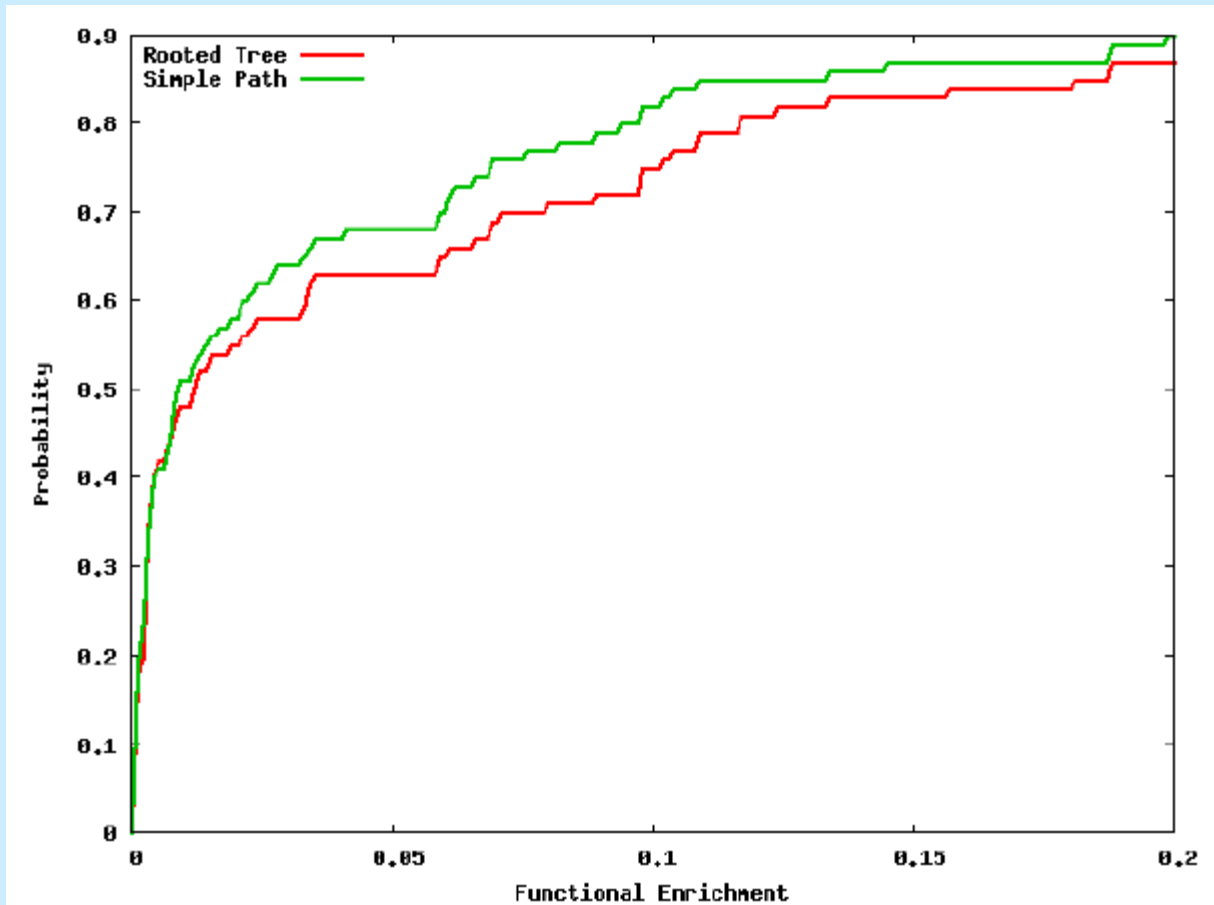
The Path Recursion

- Assume edge weights are $-\log(\text{prob.})$
- Goal: minimum weight k -path that starts at a vertex from I .
- Base: $W(v, \{v\}) = 0$ if v is in I (o/w ∞).

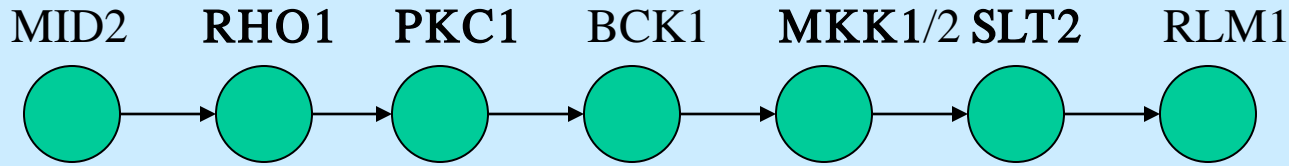
$$W(v, S) = \min_{c(u) \in \{S - c(v)\}} W(u, S - \{c(v)\}) + w(u, v), |S| > 1$$

Application to yeast (1)

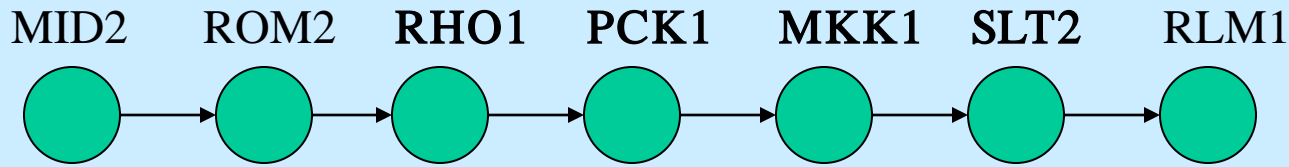
- 100 best paths from receptor to transcription factor of size 8 @ 99.9% success.



A) Cell wall integrity pathway in yeast

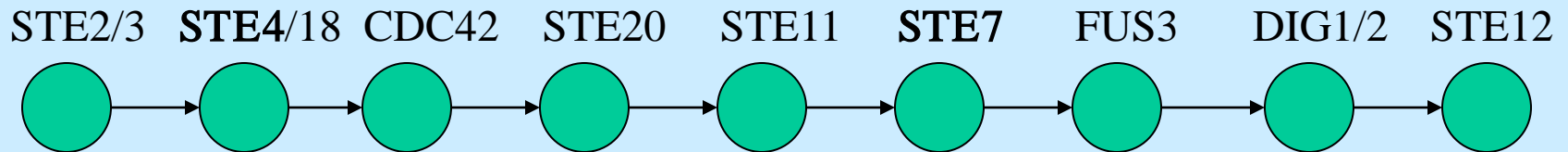


B) Best path of length 7 found from MID2 to RLM1

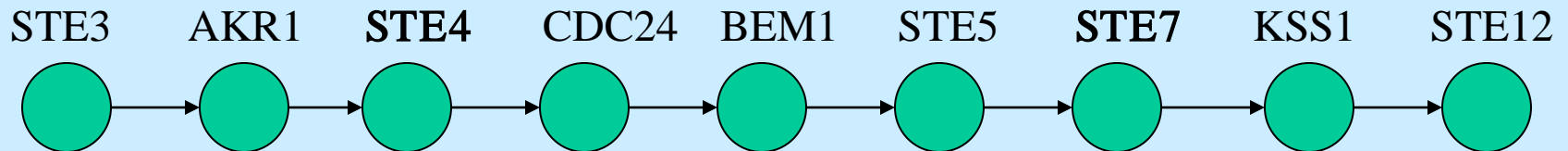


Appl. To
yeast (2)

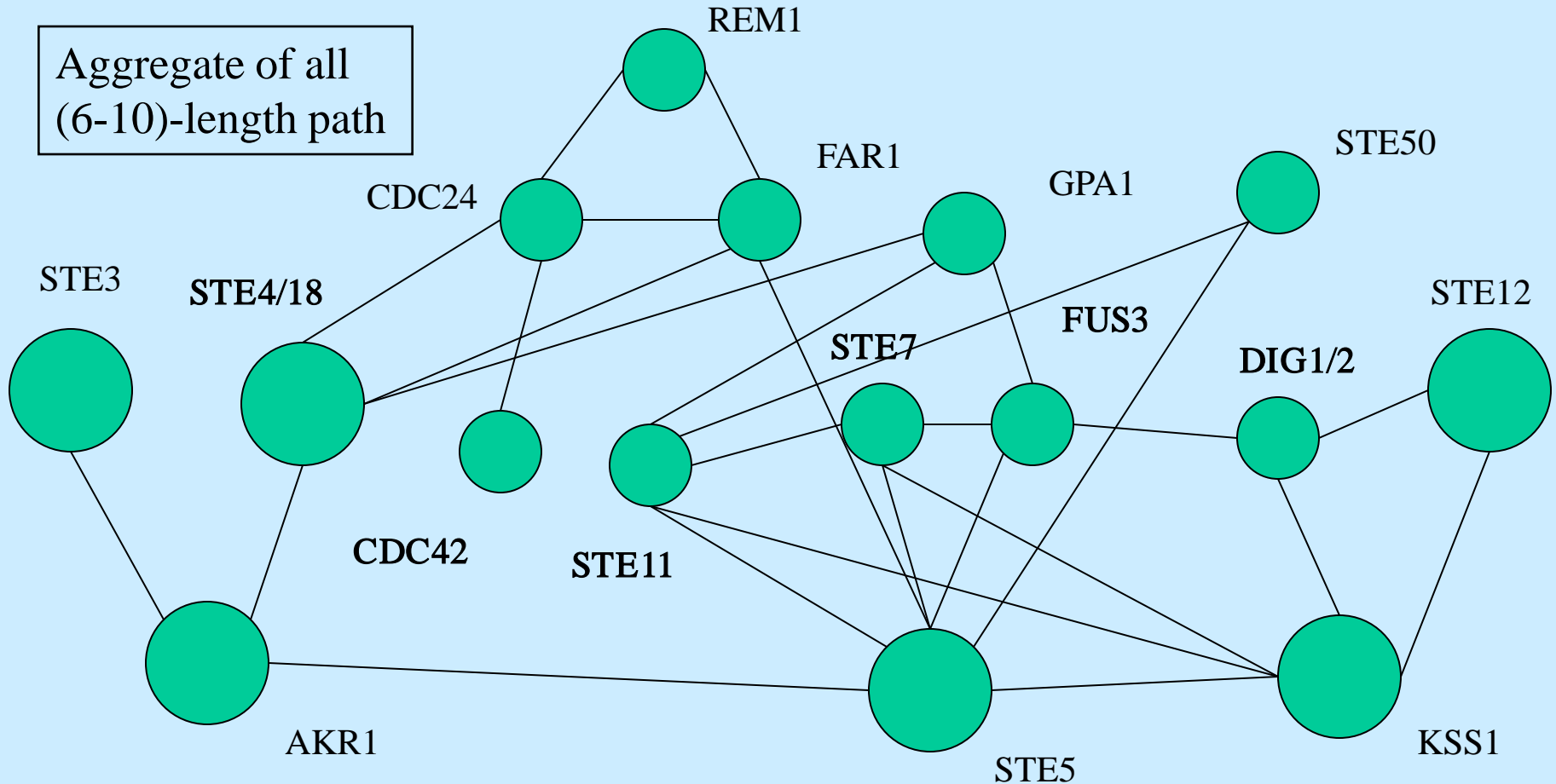
C) Pheromone response pathway in yeast



D) Best path of length 9 found from STE2/3 to STE12



A Closer Look at Pheromone Response



The real pathway (main chain):

