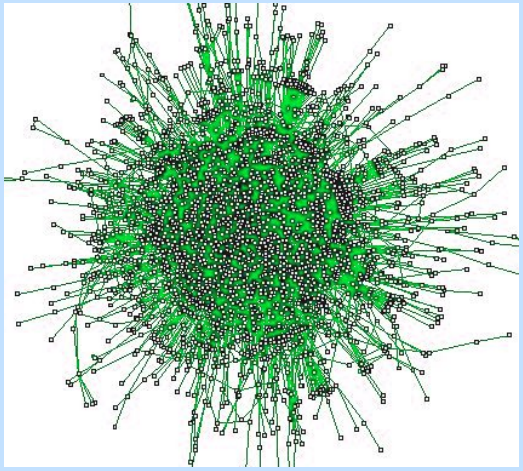


Network biology minicourse (part 3)

Algorithmic challenges in genomics

Identifying network modules

Roded Sharan

School of Computer Science, Tel Aviv University

Gene/Protein Modules

• A *module* is a set of genes/proteins performing a distinct biological function (Hartwell et al., Nature'99)

• Examples for PPI modules:

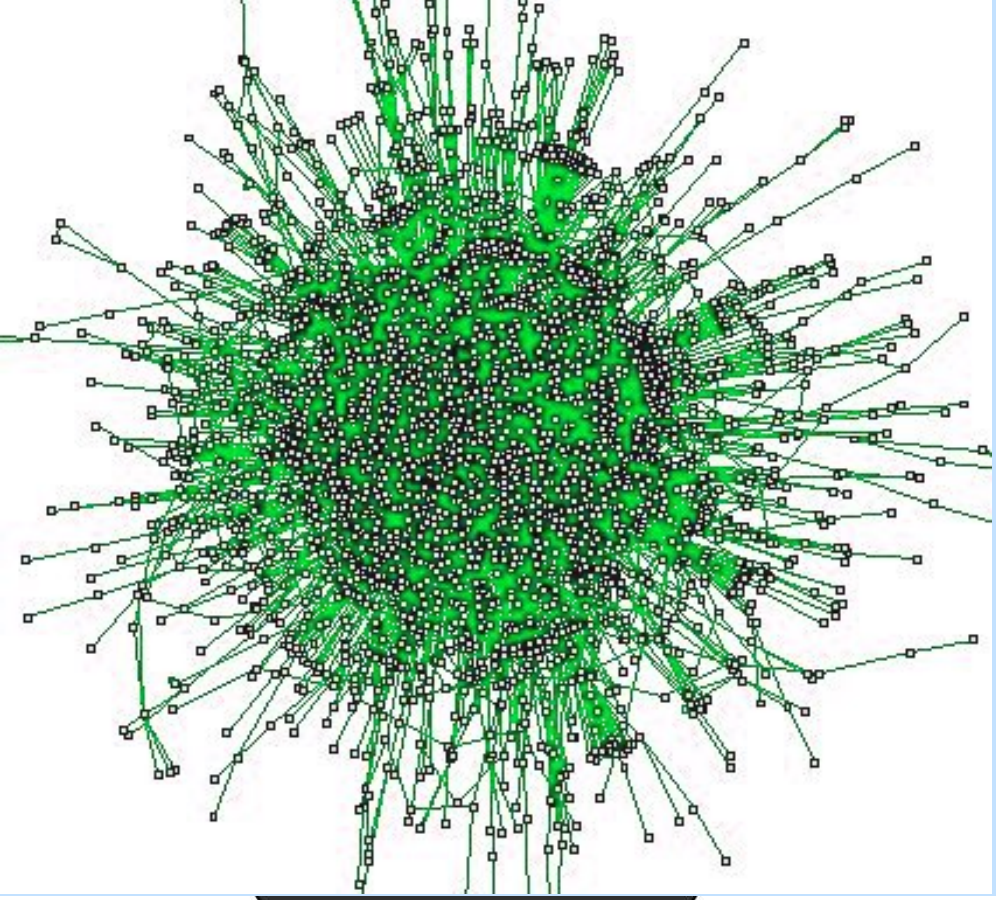
– *protein complex*: assembly of proteins that build up some cellular machinery.

– *signaling pathway*: a chain of interacting proteins propagating a signal in the cell.

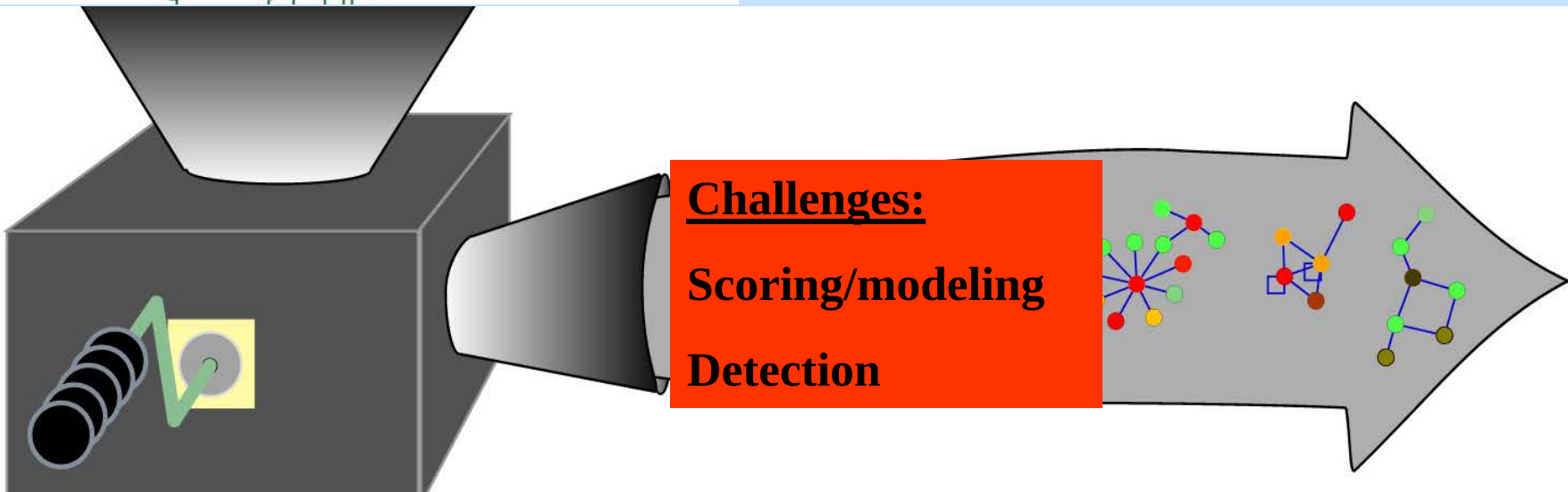
• A data-driven "definition": a module is characterized by a coherent behavior of its genes w.r.t. a certain biological property.

Module finding vs. clustering

- Modules can overlap
- Need not cover the entire network
- Some problems translate to biclustering…



Distilling Modules from Networks



Challenges:

Scoring/modeling

Detection

Outline

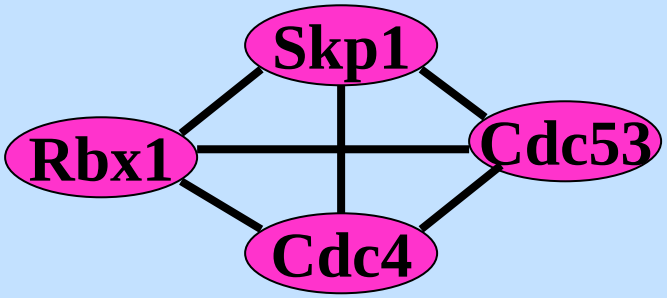
- **Protein complex**: local prediction strategies
- **Protein complex**: global (clustering) strategies
- **Protein complex**: biclustering
- **Pathway** inference
- Network **integration**

Outline

- **Protein complex**: local prediction strategies
- **Protein complex**: global (clustering) strategies
- **Protein complex**: biclustering
- **Pathway** inference
- Network **integration**

From complexes to heavy subgraphs

- Protein complexes are manifested as dense subgraphs.
- For example – the SCF complex:



Modeling problem: statistical scoring of density

Algorithmic problem: find high-scoring subgraphs

MCODE

- Vertex weighting based on density of its neighborhood
- Complex prediction:
  - Start from heaviest vertex of weight *w*.
  - Iteratively, add neighbors whose weight is above *pw*, where *p* is a parameter.
  - Repeat till all vertices are covered.
- Postprocessing

Bader & Hogue, Bioinformatics 2003

Details & limitations

- ***k*-core:** a graph of minimal degree *k*.
- **Density:** % edges out of all possible vertex pairs.
- The **weight** of a vertex is defined as the density of the highest *k*-core of its closed neighborhood, multiplied by the corresponding *k*.

Main limitations

- No underlying probabilistic model
- Complexes cannot overlap (up to postprocessing).

NetworkBLAST

- Use likelihood-ratio scoring.
- **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.

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

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

- Actual score takes into account edge reliabilities
- $\log L(C)$ is additive over edges and non-edges of C
- Complexes are found via greedy local search

S. et al. JCB & PNAS 2005

Outline

- **Protein complex**: local prediction strategies
- **Protein complex**: global (clustering) strategies
- **Protein complex**: biclustering
- **Pathway** inference
- Network **integration**

Markov Clustering (MCL)

Idea: Random walk tends to remain within clusters

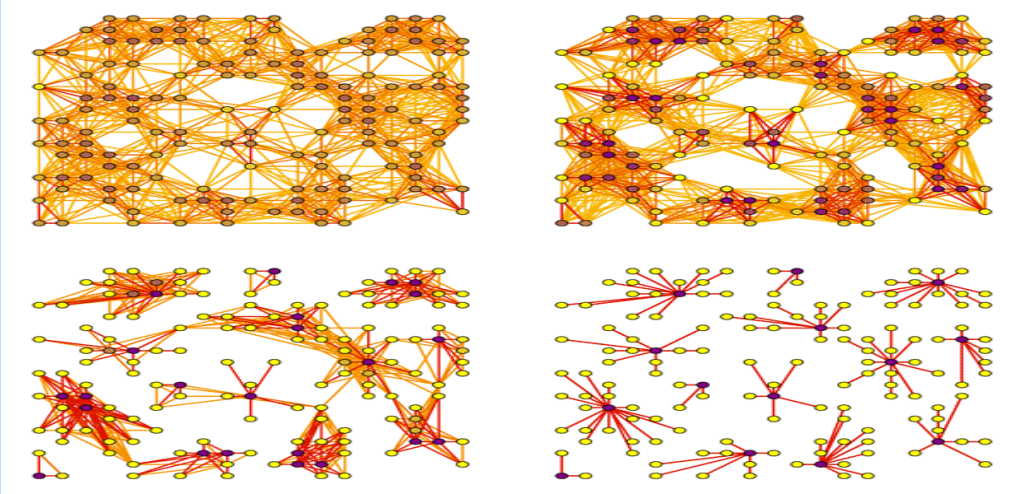
Algorithm:

- Input: stochastic matrix M of the graph; parameters e, r.
- Iterate until convergence:

  Expansion: $M \leftarrow M^e$ //higher-length walks

  Inflation: raise each entry to the power of r (and normalize)

  //boost probs of intra-cluster walks



Enright et al., NAR 2002

Modularity-based clustering

Q = #(edges within groups) - E(#(edges within groups in a RANDOM graph with same node degrees))

Trivial division: all vertices in one group

==> Q(trivial division) = 0

k_i = degree of node i

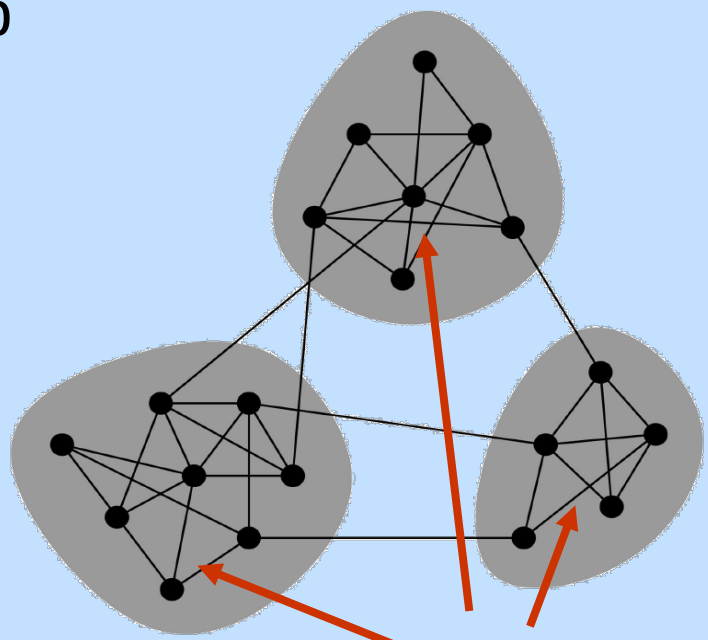
$M = \sum k_i = 2|E|$

Aij = 1 if $(i,j) \in E$, 0 otherwise

Eij = expected #edges between i and j in a random degree-preserving graph.

Lemma: $Eij \approx k_i * k_j / M$

$Q = \sum(Aij - ki*kj/M \mid i,j \text{ in the same group})$



Edges within groups

Newman, PNAS 2006

Division into two groups

Q = $\sum$(Aij - ki*kj/M | i,j in the same group)

- Suppose we have n vertices {1,…,n}
- **s** - {±1} vector of size n.
  Represent a 2-division:
  - si == sj iff i and j are in the same group
  - ½ (si*sj+1) = 1 if si==sj, 0 otherwise
- ==> $$Q = \frac{1}{2}\sum_{i,j}(A_{ij} - \frac{k_i k_j}{M})(s_i s_j + 1)$$

Division into two groups (2)

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

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

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

B = the modularity matrix
- symmetric

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

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

$$\beta_1 \geq \beta_2 \geq \cdots \geq \beta_n$$

B's orthonormal eigenvectors

$$\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots \boldsymbol{u}_n$$

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

$$Q = \frac{1}{2}\boldsymbol{s}^T \boldsymbol{B}\boldsymbol{s} \quad \left[\boldsymbol{s} = \sum_i a_i \boldsymbol{u}_i\right] \Rightarrow Q = \frac{1}{2}\sum_i \beta_i a_i^2$$

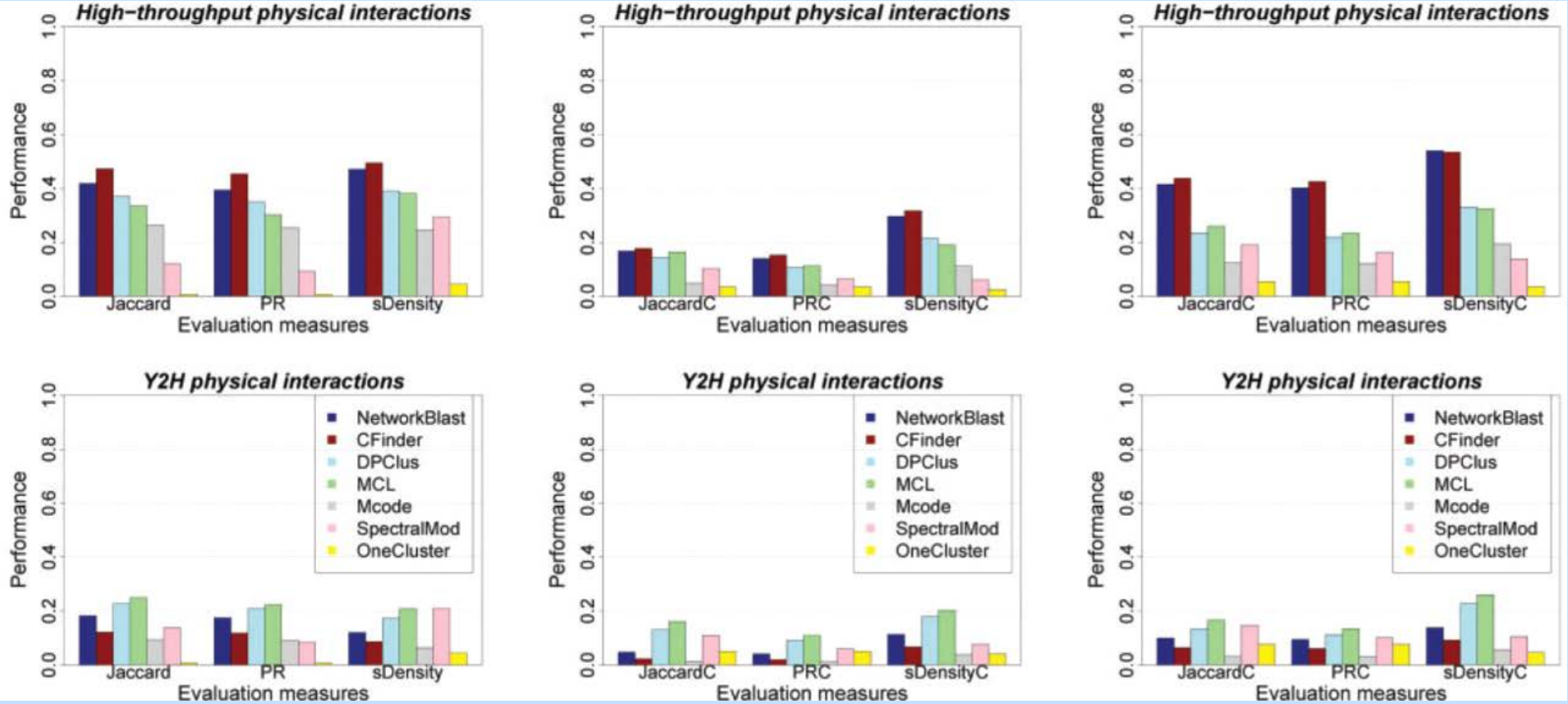
- Which vector **s** maximizes Q?
  - clearly **s** ~ **u1** maximizes Q, but **u1** may not be {±1} vector
  - Heuristic: maximize the projection of s on u1 (a_1): choose si= +1 if $u1_i$>0, si=-1 otherwise

Performance evaluation

- Based on gold-standards such as:
  - GO terms
  - GO complexes
  - MIPS complexes (yeast)

- Use measures of **precision** and **recall**
- Could be computed by overlaps (taking the mean, or combine into Jaccard indices) or statistically (hypergeometric encrichment).

Performance comparison



MIPS GO BP GO CC

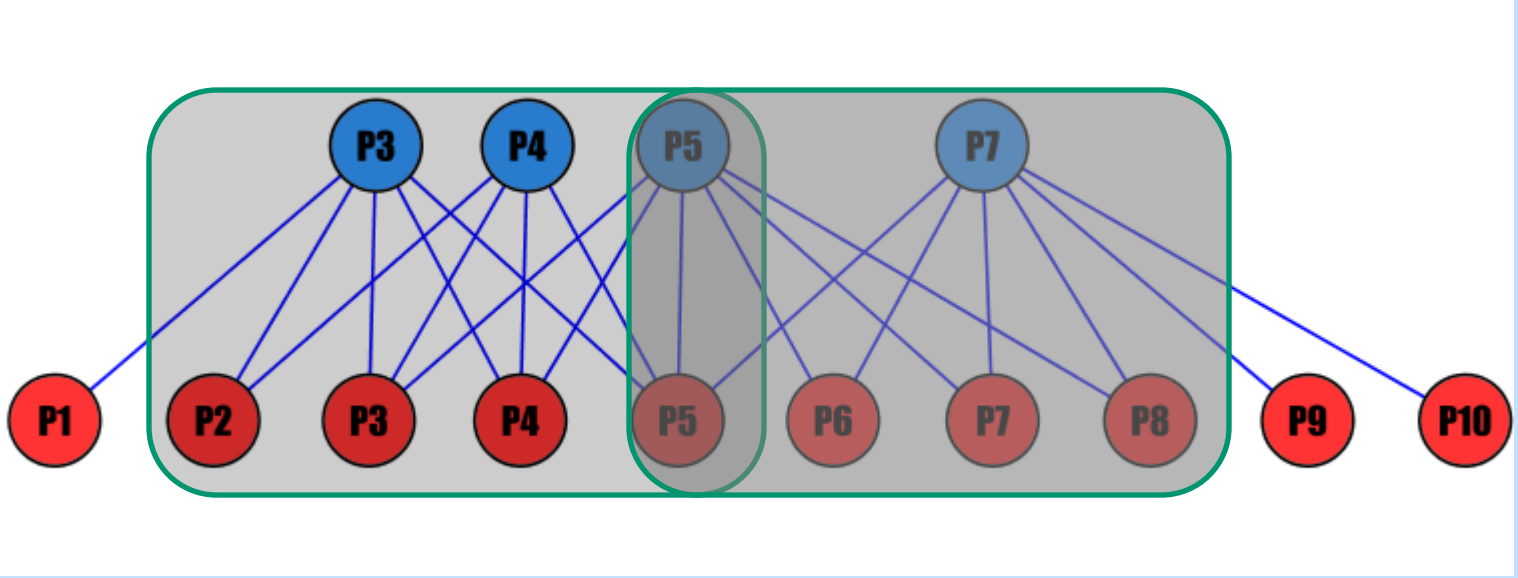
Song and Singh. Bioinformatics 2009

Outline

- **Protein complex**: local prediction strategies
- **Protein complex**: global (clustering) strategies
- **Protein complex**: biclustering
- **Pathway** inference
- Network **integration**

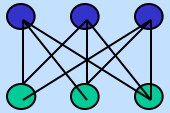
Going back to the sources

- Data are not binary interactions! (Scholtens et al.'05)
- Construct a bait-prey graph.
- Use biclustering to detect sets of preys that co-occur with the same baits.

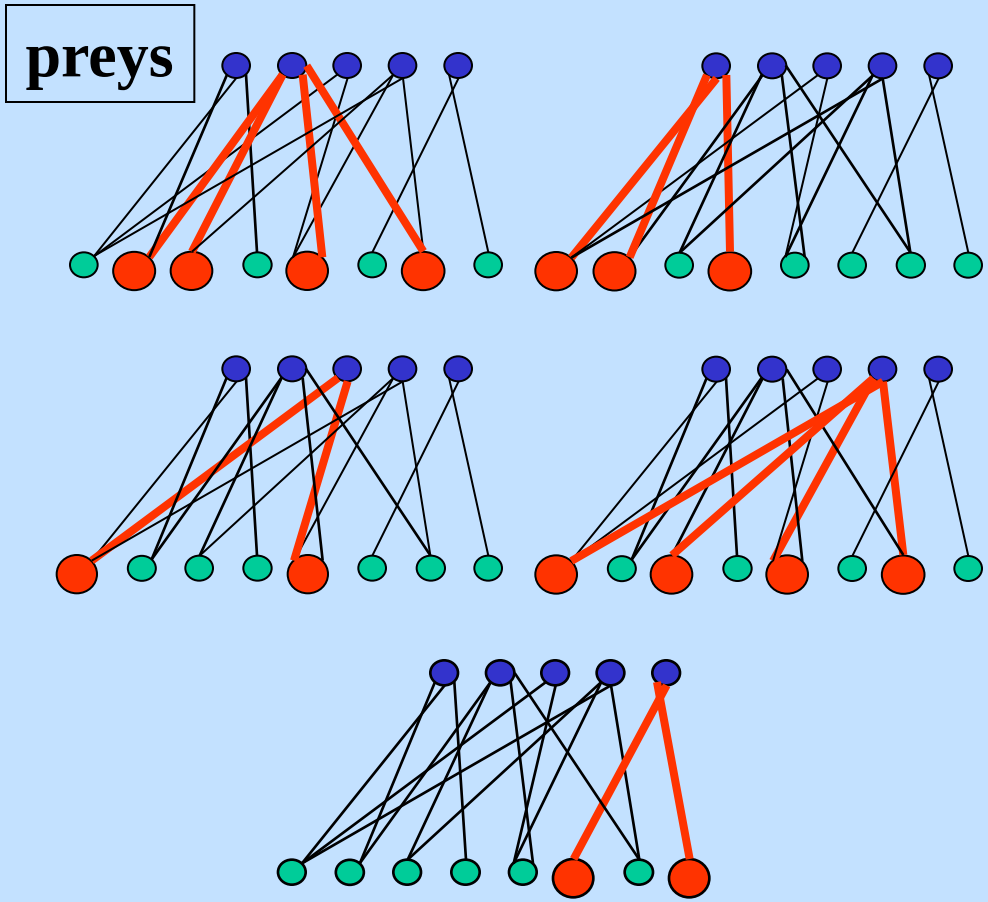


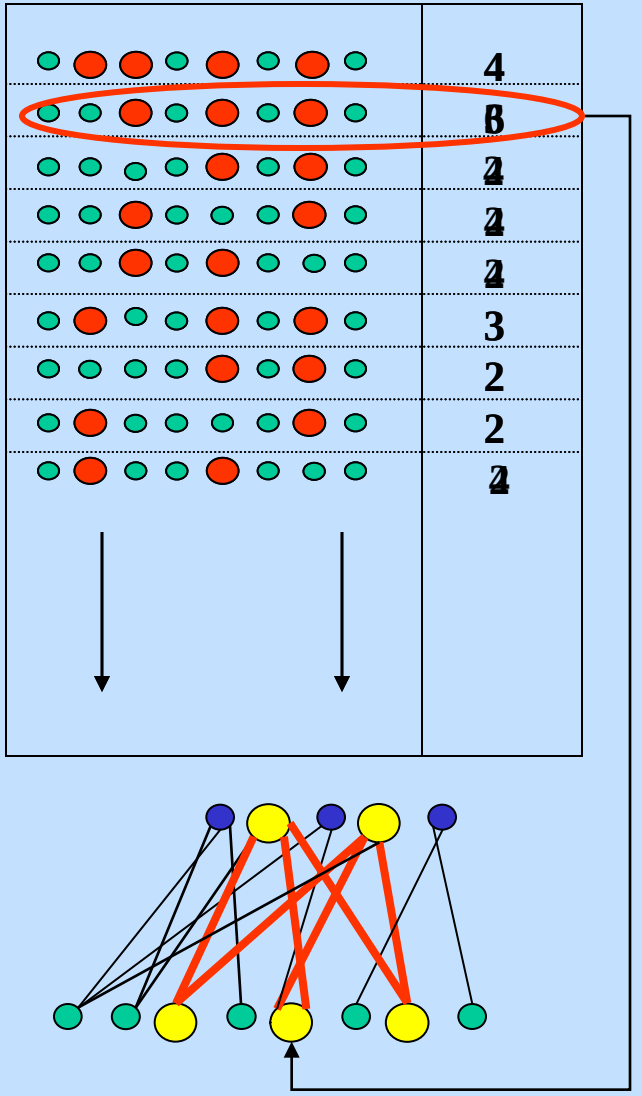
Geva et al. Bioinformatics 2011

Maximum Bounded Biclique



Assumption: Prey degrees $\leq d$.





$O(n2^d)$-time

Tanay et al. Bioinformatics 2002

Outline

- **Protein complex**: local prediction strategies
- **Protein complex**: global (clustering) strategies
- **Protein complex**: biclustering
- **Pathway** inference
- Network **integration**

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.
- Trivial algorithm runs in $O(n^k)$.
- First application to PPI networks by Steffen et al. Bioinformatics 2002
- We will be interested in a *fixed parameter* algorithm, i.e., time is exponential in k but polynomial in n.

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).

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

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

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

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:

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

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.

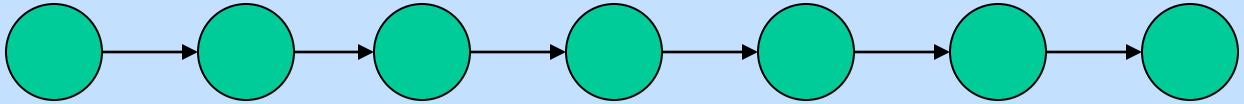
Scott et al. JCB 2005

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, e.g. membrane to TF (a la Steffen et al.)
- Can force the inclusion of a specific protein on the path by giving it a unique color
- …

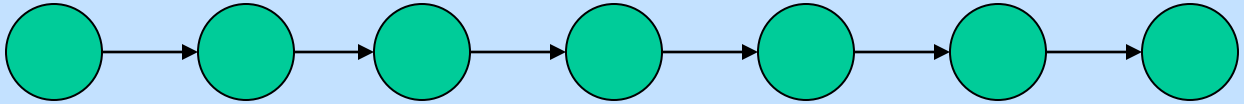
A) Cell wall integrity pathway in yeast

MID2 **RHO1** **PKC1** BCK1 **MKK1**/2 **SLT2** RLM1



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

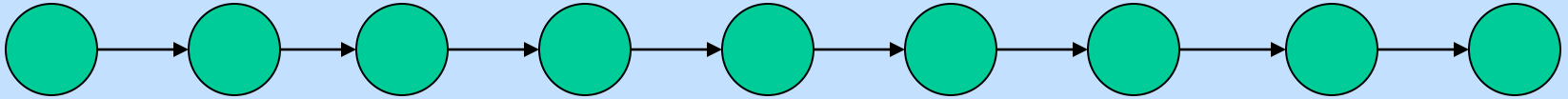
MID2 ROM2 **RHO1** **PCK1** **MKK1** **SLT2** RLM1



Appl. to yeast

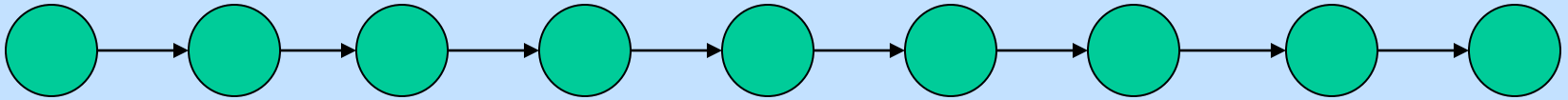
C) Pheromone response pathway in yeast

STE2/3 **STE4**/18 CDC42 STE20 STE11 **STE7** FUS3 DIG1/2 STE12

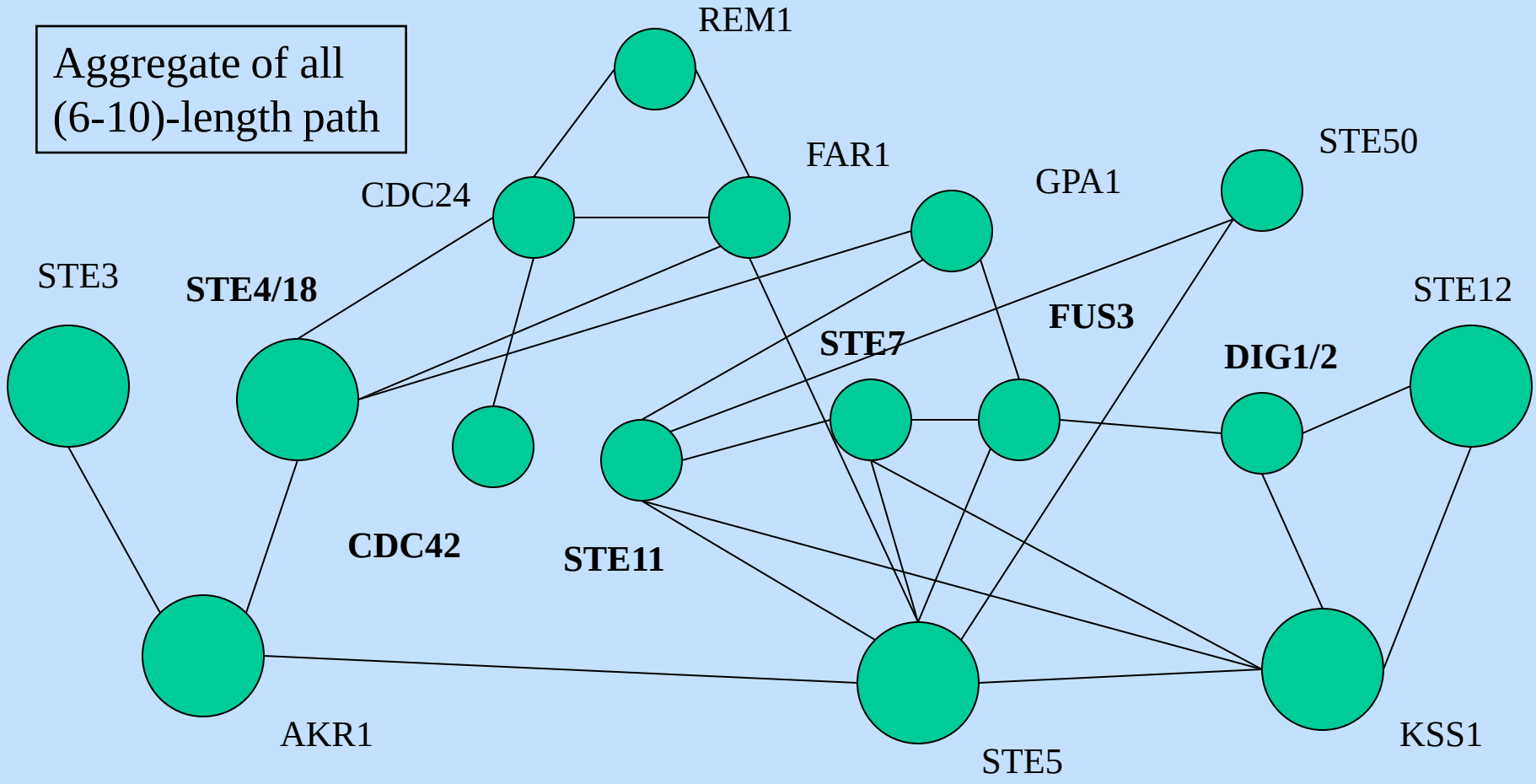


D) Best path of length 9 found from STE2/3 to STE12

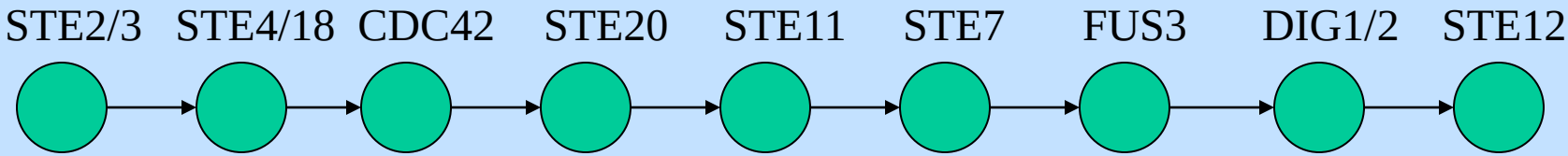
STE3 AKR1 **STE4** CDC24 BEM1 STE5 **STE7** KSS1 STE12



A Closer Look at Pheromone Response



The real pathway (main chain):



Outline

- **Protein complex**: local prediction strategies
- **Protein complex**: global (clustering) strategies
- **Protein complex**: biclustering
- **Pathway** inference
- Network **integration**

Integration: main idea

• Overcomes noise and incomplete information problems.

• Provides a more complete information on the module's activity or cross-talk or regulation.

Two common integration schemes:

• Identical modeling of all data types – commonly looking for cliques (e.g. Gunsalus et al.'05).

• **Different models for different data types**

Genetic interactions

A *genetic interaction* is the interaction of two genetic perturbations in determining a phenotype.

Synthetic lethality: Two genes A,B are synthetic lethal if knockouts of A or B separately are viable, but knocking out both is lethal.

1 + 1 = 0

- Can be systematically assayed by a Synthetic Geneic Array (SGA): query vs. all non-essentials.
- There are workarounds also for essential genes.

Integrating PPI & GI (Kelley & Ideker ’05)

Two common models for genetic interactions:

1. Between-pathway: bridging genes operating in two parallel pathways. When either pathway is active the cell is viable.
2. Within-pathway: occur between protein sub-units within a single pathway. A single gene is dispensable for the function of the pathway.

Scoring schemes

- Apply likelihood ratio scoring for physical and genetic networks separately and combine the scores.

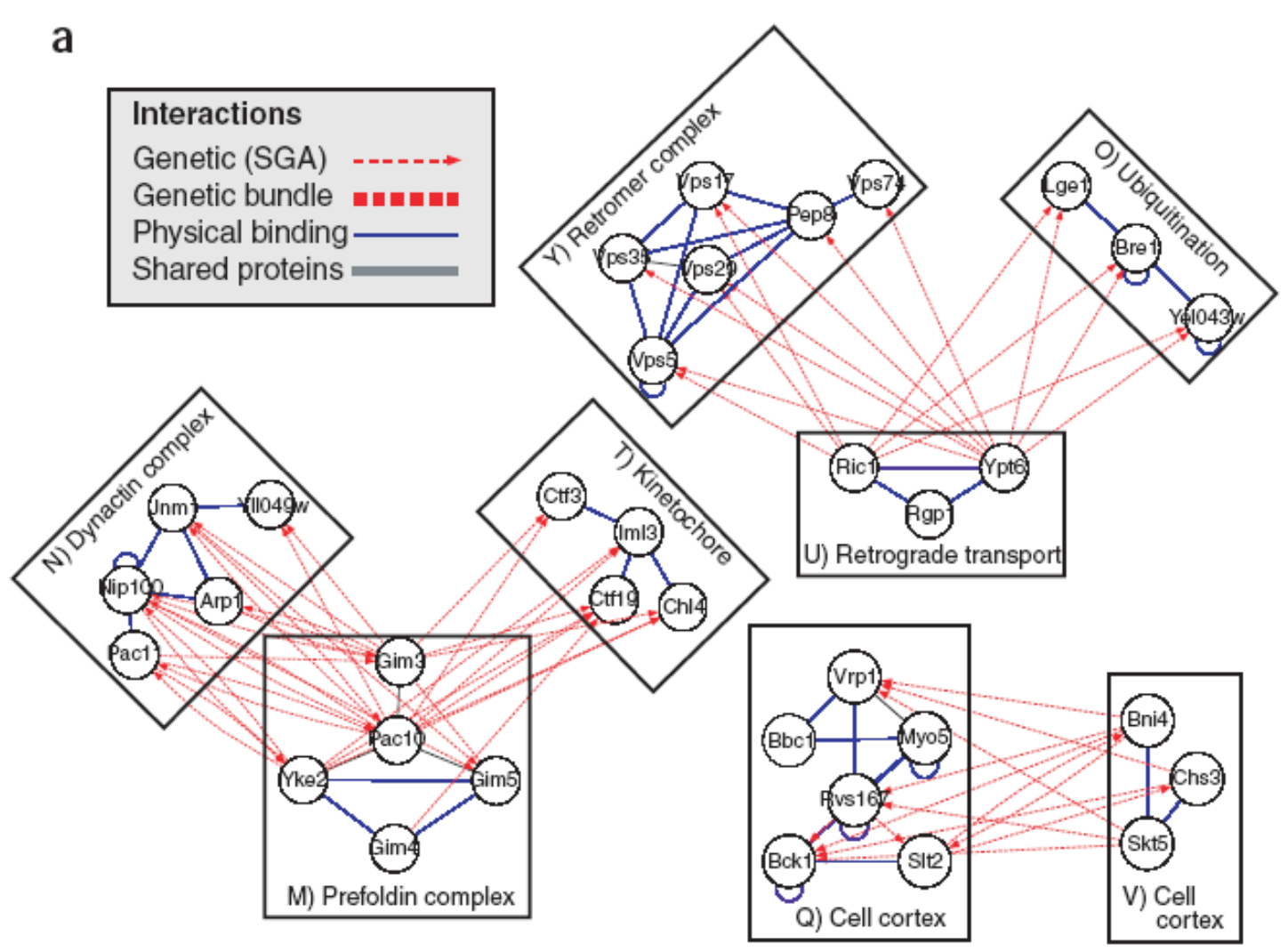
$$C = (V', E')$$

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

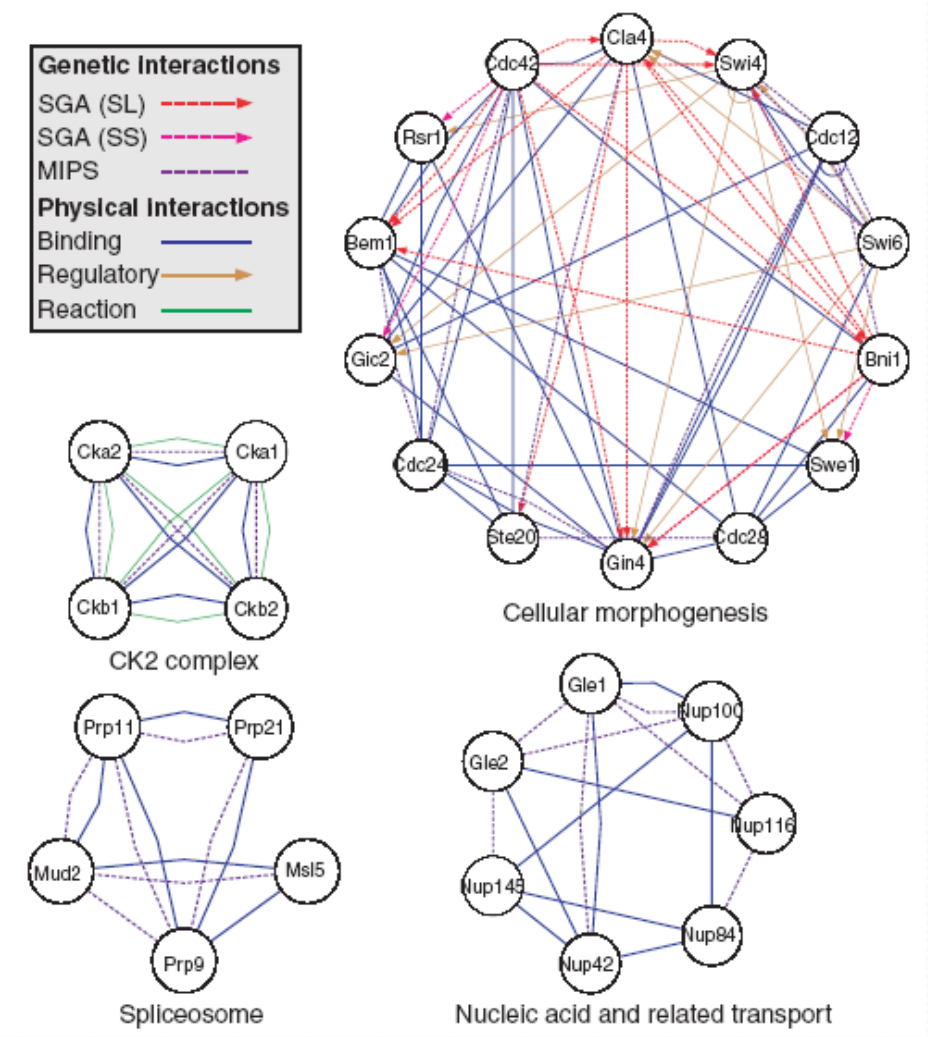
$$L(C_{within}) = L(C_{physical}) L(C_{genetic})$$

$$L(C_{between}) = L_{physical}(C_1) L_{physical}(C_2) L_{genetic}(C_1, C_2)$$

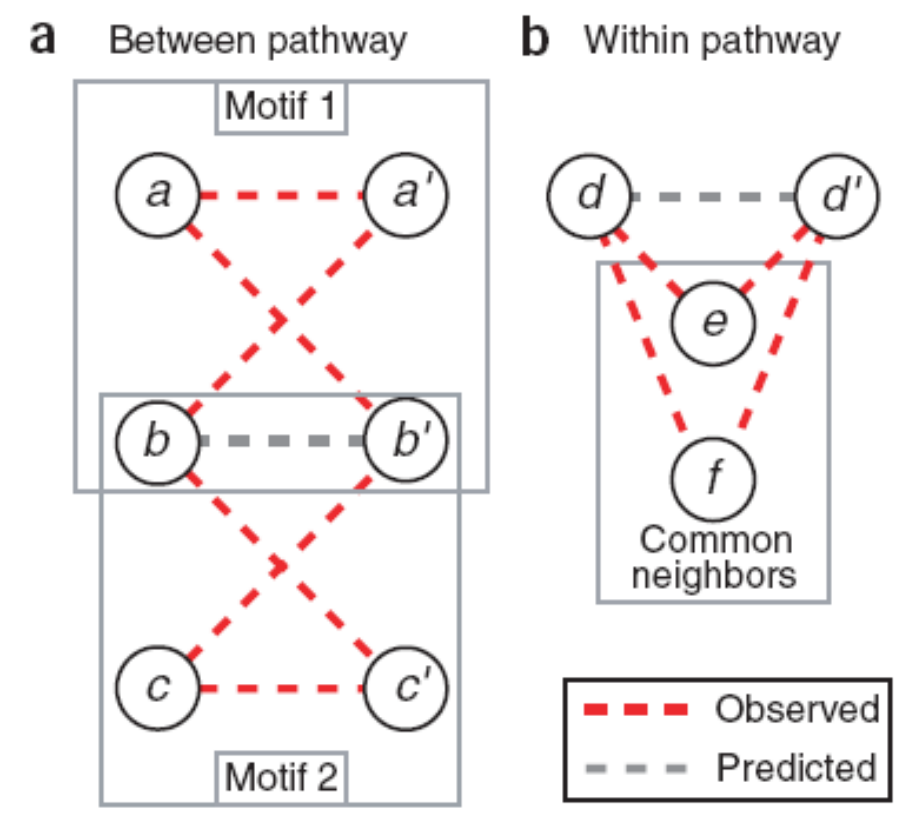
Between-Pathway Results



Within-Pathway Results

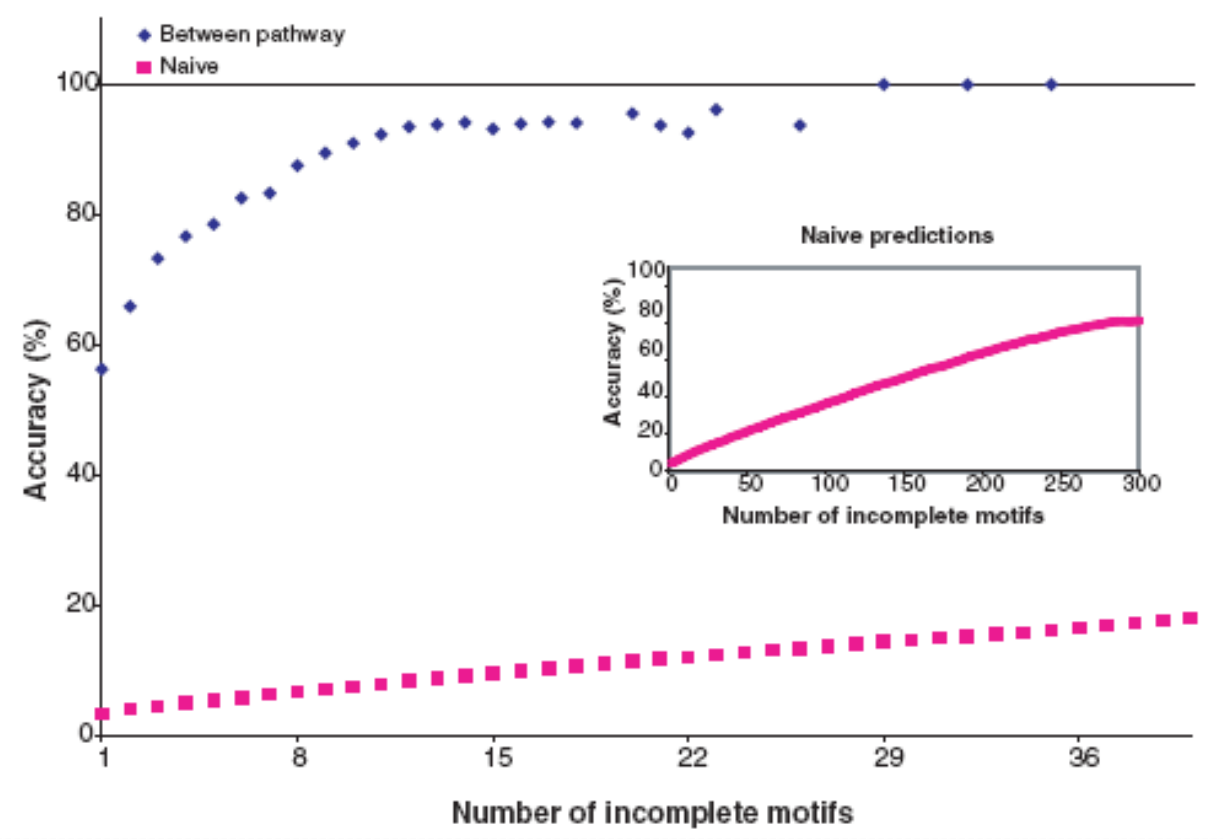


GI Prediction



- Prediction is based on incomplete motifs, as shown here.
- Two strategies: motif genes are unconstrained (naïve) or, alternatively, forced to be within a model.

GI Prediction – Between-Pathway



- Predicted 43 GIs with 87% estimated accuracy (5-fold CV).
- Physical data greatly improves accuracy (from 5%).

GIs mostly occur between pathways

- 1377 interactions are associated with between-pathway models; only 394 within-pathway ones.

(These statistics account for only ~40% of GIs.)

- ~63% of between-pathway models show enriched function, while ~57% within-pathway models are enriched.

- Higher accuracy of between-pathway in GI prediction: only 38% accuracy attained for within-pathway model.

Summary

• Modules take different shapes, most focus is on protein complexes that are modeled as heavy subgraphs

• Local, global and biclustering strategies

• Integration of different networks enhances prediction accuracy

• The field is moving toward module prediction from multiple information types such as disease modules, drug response pathways etc.