Network-based interpretation of sequence variants

Sushmita Roy

sroy@biostat.wisc.edu

Computational Network Biology

Biostatistics & Medical Informatics 826

https://compnetbiocourse.discovery.wisc.edu

Nov 27th 2018

RECAP of problems in network biology

Biological problem

- ❑ Mapping regulatory network structure
- ❑ Dynamics and context specificity of networks
- ❑ Understanding design principles of biological networks
- ❑ **Interpretation of sequence variants/perturbations**
- ❑ Identification of important genes
- ❑ Integrating different types of molecular genomic data
- ❑ Smoothing noisy matrices

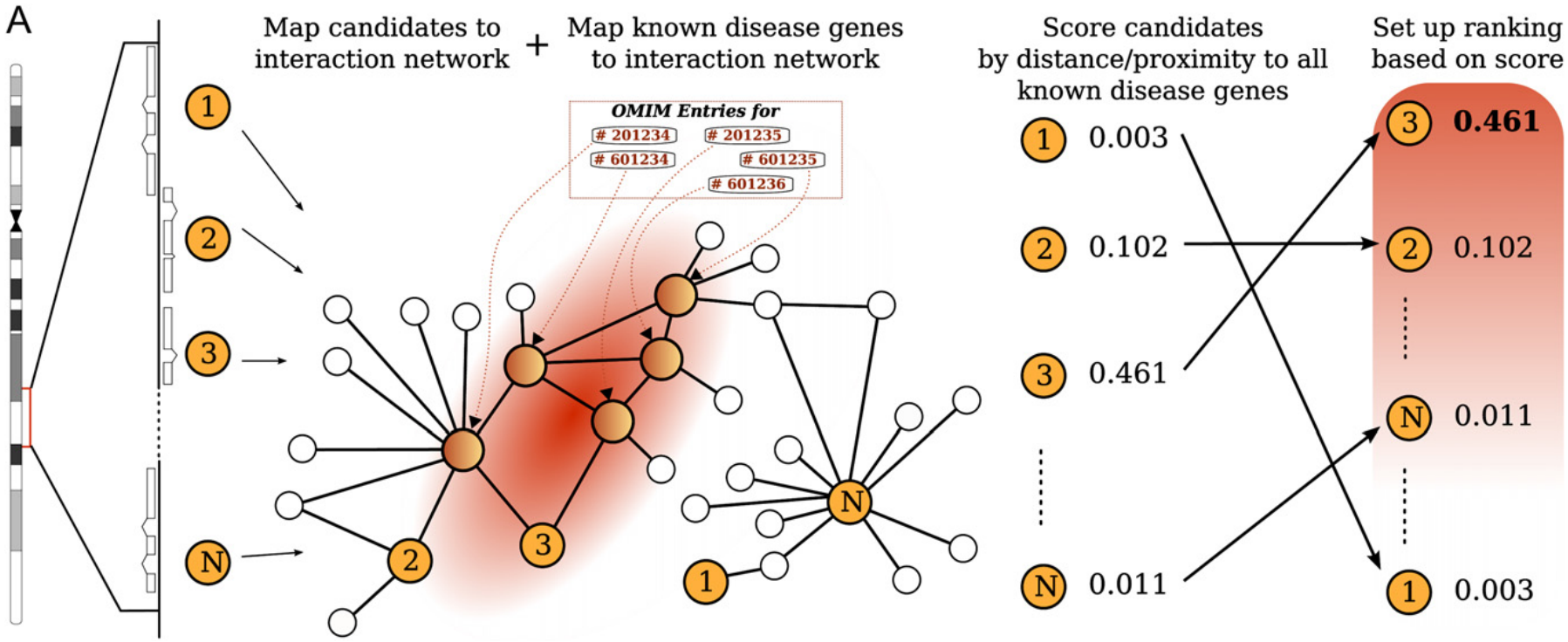
Computational approaches

- ❑ Probabilistic graphical models
- ❑ Graph structure learning
- ❑ Multiple network learning
- ❑ Topological properties of graphs
- ❑ Graph clustering
- ❑ Graph alignment
- ❑ **Diffusion on graphs**

Graph diffusion based methods

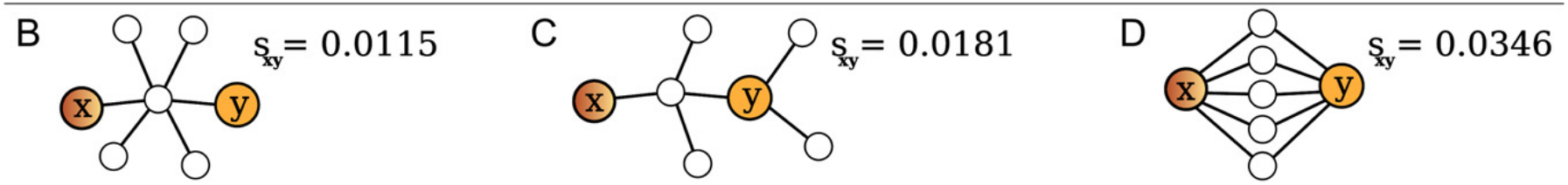
- Aim to use the graph structure to define similarity between nodes on the graph
- The similarity function is also sometimes referred to as a "kernel"
  - Random walk kernel, Diffusion kernel
- Many applications
  - Gene prioritization
  - Smoothing of count matrices
  - Interpreting variants on networks

Overview of network-based gene prioritization



This is for GeneWanderer, but other approaches are similar.

Motivation of using global distance



Known gene: x
Candidate gene: y

- Global similarity is more sensitive and different for each of the above cases
- In contrast, local shortest-path similarity is the same for all pairs
- Direct interactions will never select y as a candidate

RECAP of graph diffusion based gene prioritization

- We discussed the GeneWanderer method
- Focus on global rather than local graph distances
- Global distances can be obtained using random walks or a diffusion kernel
- Global distances were able to rank known disease genes much better than shortest path based methods
- Network-based prioritization can be used when we have a small number of known genes to start with

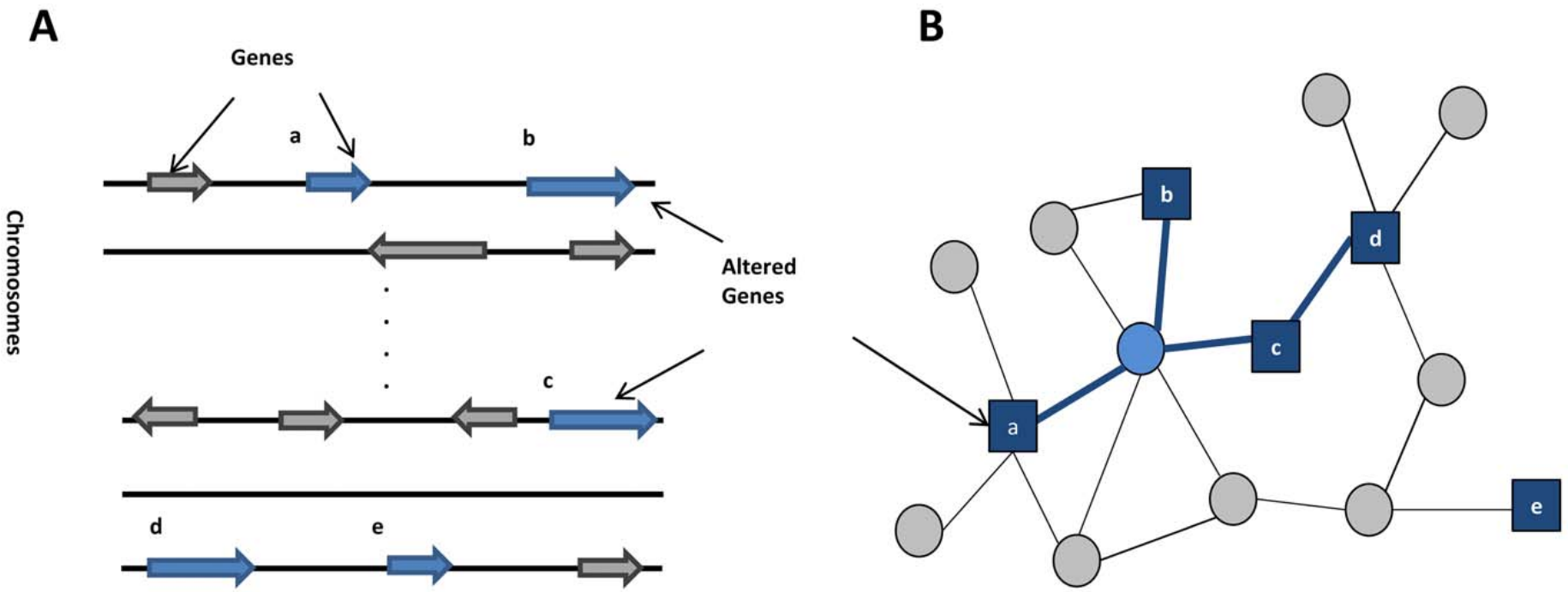
Perturbations in networks

- Understanding genetic perturbations are important in biology
- Genetic perturbations are useful to identify the function of genes
  - What happens if knock gene A down?
    - Measure some morphological phenotype like growth rate or cell size
    - Measure global expression signatures
- Perturbations can be artificial or natural
  - Artificial perturbations
    - Deletion strains
  - Natural perturbations
    - Single nucleotide polymorphisms
    - Natural genetic variation
- Perturbations in a network can affect
  - Nodes or edges
  - Edge perturbations
    - Mutations on binding sites

Types of algorithms used to examine perturbations in networks

- Graph diffusion followed by subnetwork finding methods
  - HOTNET
- Probabilistic graphical model-based methods
  - Factor graphs
  - Nested Effect Models (NEMs)
- Information flow-based methods (also widely used for integrating different types of data)
  - Prize collecting steiner tree
  - Min cost max flow

Identification of subnetworks perturbed in diseases



Cho D-Y, Kim Y-A, Przytycka TM (2012) Chapter 5: Network Biology Approach to Complex Diseases. PLoS Comput Biol 8(12): e1002820. doi:10.1371/journal.pcbi.1002820
http://www.ploscompbiol.org/article/info:doi/10.1371/journal.pcbi.1002820

Motivation of HOTNET

- Somatic mutations play a major role in cancer
- Mutation profiles of cancers is very heterogeneous
  - Tumors harbor on average approximately 80 somatic mutations, but two tumors rarely have the same complement of mutations
- Thousands of genes can be mutated in cancer
  - This makes it difficult to identify “driver” mutations vs “passenger” mutations
  - Mutations at the pathway level (group of genes) is likely responsible for a particular type of cancer
- Can we identify subnetworks (representing new pathways) that are significantly mutated?

F. Vandin, E. Upfal, and B. J. Raphael, "Algorithms for detecting significantly mutated pathways in cancer." *Journal of computational biology* 2011

HOTNET problem setup

- Given
  - A network of protein-protein interactions
  - A set of patient tumor mutation profiles
- Do
  - Find “significantly” mutated subgraphs
    - A subgraph that best connects these genetic alterations
    - Best means a subgraph that includes as many tumour samples with as few genes
- How?
  - Find the global influence of mutations on a particular gene
  - Search for a subnetwork in this global influence graph that is significantly mutated

HOTNET's approach vs ActiveSubgraphs

- HOTNET aims to find a subnetwork that has significantly many more mutations than random subnetworks
- This bears some resemblance to the ActiveSubgraph approach where we were trying to find subnetworks significantly up or down-regulated
- The key differences in HOTNET is that
  - we do not have (gene expression) measurement of mutations for all genes
  - only a small number of genes maybe mutated

Key steps of HOTNET algorithm

- Build an influence graph which specifies the influence of one node over another
  - Graph diffusion
  - Builds a network with the tested genes as well as their neighborhood
- Find significantly mutated subnetworks (two ways)
  - include genes mutated in a lot of samples
  - Enhanced influence model where the influence edges are weighted by the number of mutations
- Test for the significance of the number of subnetworks of a particular size

Diffusion kernel used in HOTNET

- HOTNET uses a specific type of kernel called the heat diffusion kernel
- L=D-A denotes the graph Laplacian, where A is the adjacency matrix and D is the degree matrix
- Let γ denote the constant rate at which heat is lost at any node
  - E.g. this could be proportional to the mean degree
- L_γ is L+γI
- Let s be a source node
- The influence of s on all n nodes at time t is denoted as

$$\mathbf{f}^s(t) = [f_1^s(t), \cdots, f_n^s(t)]$$

Influence of s on node 1

Diffusion kernel used in HOTNET

- Rate at which diffusion occurs from a source node on the graph is given by

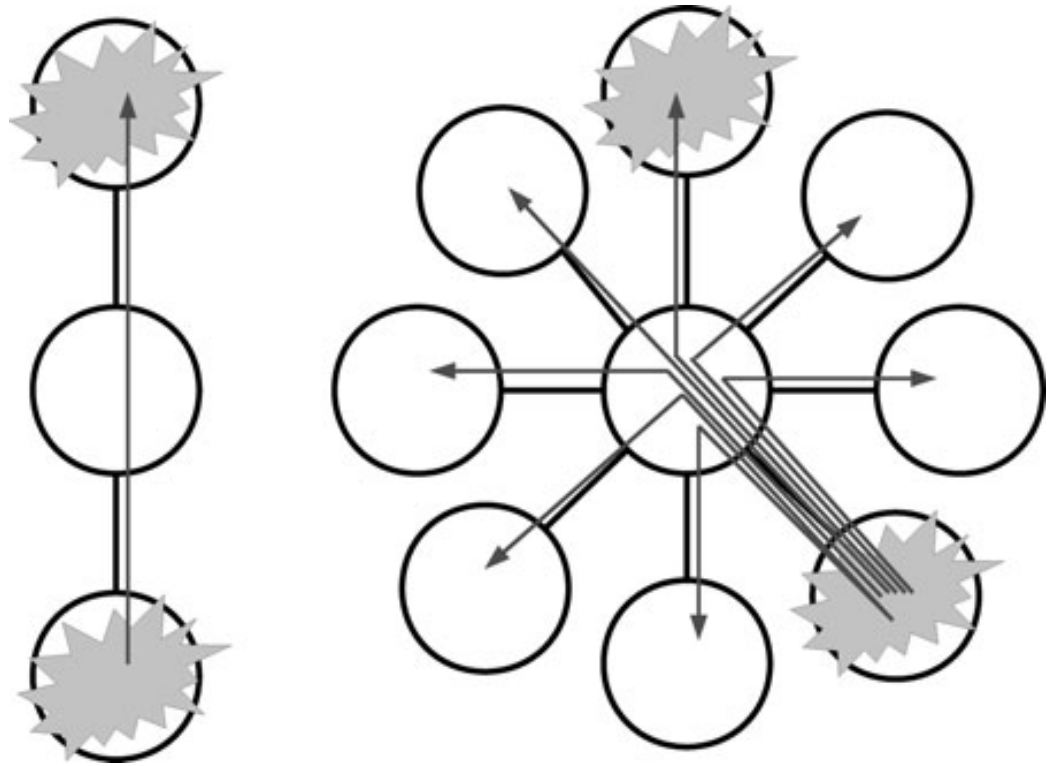
$$\frac{d\mathbf{f}^s(t)}{dt} = -L_\gamma \mathbf{f}^s(t) + \mathbf{b}^s$$

- where $\boldsymbol{b}^s$ be a unit vector which is 1 for node s and 0 otherwise
- The influence on all nodes at steady-state is given by

$$\mathbf{f}^s = L_\gamma^{-1}\mathbf{b}^s$$

- This diffusion process is very similar to what we had in the Diffusion kernel we used for ranking genes

Graph diffusion to downplay hub intermediate nodes



Mutations in a linear chain are more “interesting” than in a star graph

Computing the influence between vertex pairs

- Assume we have two vertices u and v
- Let $i(u,v)$ be the influence from u to v
- $w(u,v)$ is the influence between u and v and is $min[i(u,v),i(v,u)]$
- The influence graph is thus an $n \times n$ symmetric weighted graph, where n is the number of tested genes
- Further prune by removing edges with weight $w(u,v)<\delta$

Key steps of HOTNET algorithm

- Build an influence graph which specifies the influence of one node over another
  - Graph diffusion
  - Builds a network with the tested genes as well as their neighborhood
- Find significantly mutated subnetworks (two ways)
  - Set cover to include genes mutated in many samples
  - Enhanced influence model where the influence edges are weighted by the number of mutations
- Test for the significance of the number of subnetworks of a particular size

HOTNET's maximal connected cover approach

- Given
  - A weighted influence network $G=(V,E)$
  - T be a subset of V, comprising genes with a mutation
  - P_{v_i} : the set of samples with mutation in gene v_i
- Do:
  - Find a subnetwork over gene subset C that covers a maximal number of mutated samples
  - Formally, we want to find a $C=\{v_1,.. v_k\}$ such the following is maximized:

$$\left| \cup_{j=1}^{k} P_{v_j} \right|$$

Heuristic algorithm to find a maximal connected cover

- Finding the maximal connected cover set is computationally difficult
- A heuristic algorithm is used
- Add a vertex v such it is connected to the current vertex set via a node u, and maximizes the ratio of the number of new samples covered to the number of nodes between u and v

Heuristic algorithm to find a maximal connected cover

Combinatorial Algorithm

Input: Influence graph G_I and parameters δ and k
Output: Connected subgraph $\mathcal{C}$ of $G_I(\delta)$ with k vertices

1 Construct $G_I(\delta)$ by removing from G_I all edges with weight $< \delta$;
2 $\mathcal{C} \leftarrow \emptyset$;
3 **for** each node $v \in V$ **do**
4 $\quad \mathcal{C}_v \leftarrow \{v\}$;
5 $\quad$ **for** each $u \in V \setminus \{v\}$ **do** $p_v(u) \leftarrow$ shortest path from v to u in $G_I(\delta)$;
6 $\quad$ **while** $|\mathcal{C}_v| < k$ **do**
$\quad\quad$ *// $\ell_v(u)$ = set of nodes in $p_v(u)$; $P_v(u)$ = elements of I covered by $\ell_v(u)$; $P_{\mathcal{C}_v}$ = elements covered by $\mathcal{C}_v$; $P_{\mathcal{C}}$ = elements covered by $\mathcal{C}$*
7 $\quad\quad u \leftarrow \arg\max_{u \in V \setminus \mathcal{C}_v : |\ell_v(u) \cup \mathcal{C}_v| \le k} \left\{ \frac{|P_v(u) \setminus P_{\mathcal{C}_v}|}{|\ell_v(u) \setminus \mathcal{C}_v|} \right\}$;
8 $\quad\quad \mathcal{C}_v \leftarrow \ell_v(u) \cup \mathcal{C}_v$;
9 $\quad$ **if** $|P_{\mathcal{C}_v}| > |P_{\mathcal{C}}|$ **then** $\mathcal{C} \leftarrow \mathcal{C}_v$;
10 **return** $\mathcal{C}$;

Exploration

Keep adding a neighbor that has the maximal coverage with fewest additional vertices

Enhanced influence model

- The enhanced influence model was a more computationally efficient approach
- Enhance the influence measure between genes by the number of mutations observed in each of these genes
- Specifically, let v_j and v_k be two genes with a mutation

$$h(v_j, v_k) = w(v_j, v_k) \times max\{|S_j|, |S_k|\}$$

Enhanced influence

Set of samples with mutation in v_j

- Remove edges with influence $<\delta$
- Decompose the associated enhanced influence graph into connected components

Statistical analysis for determining significance of subnetworks

Null distribution of subnetworks

- Assign mutations uniformly at random
- Shuffle gene labels in mutation data (preserve mutation frequencies)
- Assess significance of the total number of subnetworks with s or more genes

Application of HOTNET to cancer dataset

- 453 mutations in 601 genes in 91 Glioblastoma (GBM) samples
- 1,013 mutations in 623 genes in 189 samples of lung adenocarcinoma
- Protein-protein interaction network with 18796 genes and 37,107 edges

HOTNET recovers pathways relevant to cancer

Table 1. Results of the Combinatorial Model

			p-*value*		*Pathway enrichment* p-*value*		
Dataset	k	*Samples*	H_0^{sample}	H_0^{gene}	*All*	*RTK/RAS/PI(3)K*	*p53*
GBM	10	67	$<10^{-10}$	4×10^{-3}	3×10^{-4}	8×10^{-4}	0.19
	20	78	$<10^{-10}$	$<10^{-3}$	10^{-5}	8×10^{-5}	0.05
Lung	10	140	$<10^{-10}$	0.02	8×10^{-6}	/	
	20	151	$<10^{-10}$	0.03	3×10^{-3}	/	

k is the number of genes in the subnetwork. *Samples* is the number of samples in which the subnetwork is mutated. *p-value* is the probability of observing a connected subgraph of size *k* mutated in a number of samples $\geq$ *samples* under the random model H_0^{sample} or H_0^{gene}. *enrichment p-value* is the *p*-value of the hypergeometric test for overlap between genes in the identified subgraph and genes reported significant pathways in TCGA (2008) or Ding et al. (2008). For GBM, *enrichment p-value* is the *p*-value of the hypergeometric test for RTK/RAS/PI(3)K and p53 pathways.

Application of HOTNET of pan-cancer mutation analysis

- More recently, an updated version of HOTNET (HOTNET2) was applied to mutation profiles of samples from 12 different cancers
- Dataset description
  - After data pre-processing there were 3,110 samples with mutations in 11,565 genes
  - Genes mutation frequency varied a lot: 1-1,291 samples

M. D. M. Leiserson, F. Vandin, H.-T. Wu, J. R. Dobson, J. V. Eldridge, J. L. Thomas, A. Papoutsaki, Y. Kim, B. Niu, M. McLellan, M. S. Lawrence, A. Gonzalez-Perez, D. Tamborero, Y. Cheng, G. A. Ryslik, N. Lopez-Bigas, G. Getz, L. Ding, and B. J. Raphael, "Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes," *Nature Genetics*, vol. 47, no. 2, pp. 106-114, Dec. 2014. [Online].

HOTNET versus HOTNET2 kernel

- Hotnet kernel

Rate of diffusing out

$$(L + \gamma I)^{-1}$$

- Hotnet2 kernel

Fraction of heat that stays on a node

$$\beta(I - (1 - \beta)W)^{-1}$$

The Hotnet2 kernel was specifically designed to further avoid “star” subnetworks.

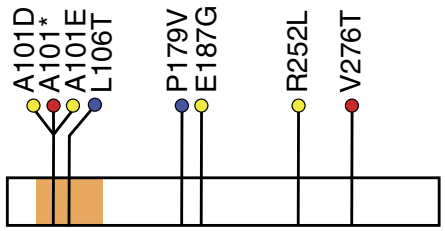
HOTNET2 for pan-cancer mutation analysis

a

Input mutation data

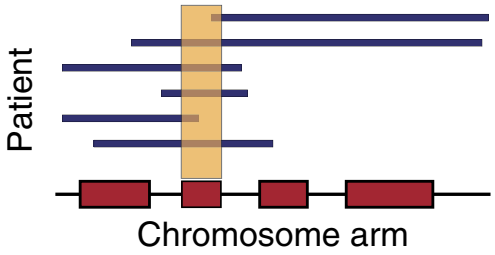
SNVs and small indels

293,863 SNVs in 20,473 genes



CNAs

19,773 CNAs in 539 genes



3,281 samples from 12 cancer types

Remove unexpressed genes

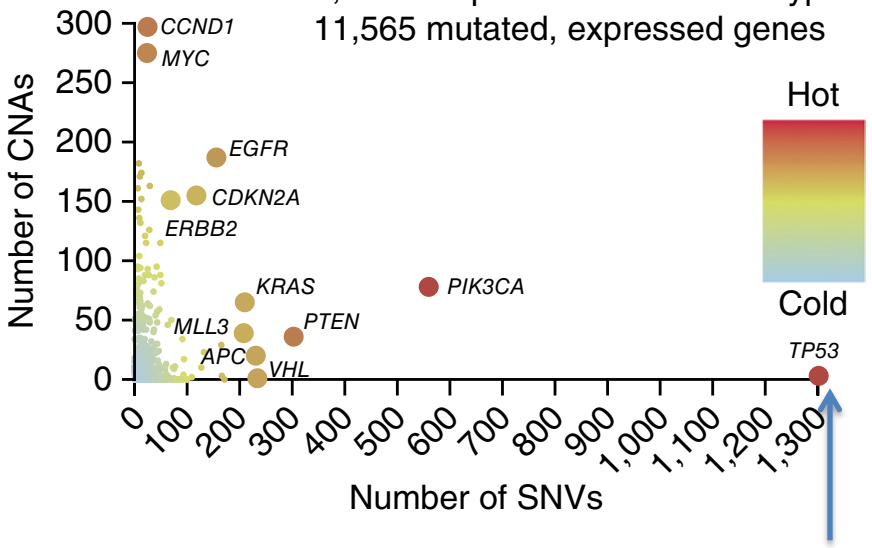
Remove hypermutator samples

b

Filtered mutation data

3,110 samples from 12 cancer types

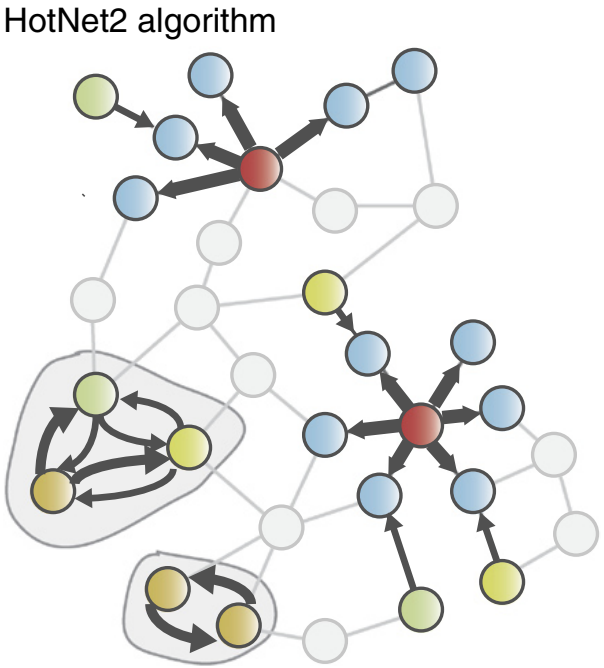
11,565 mutated, expressed genes

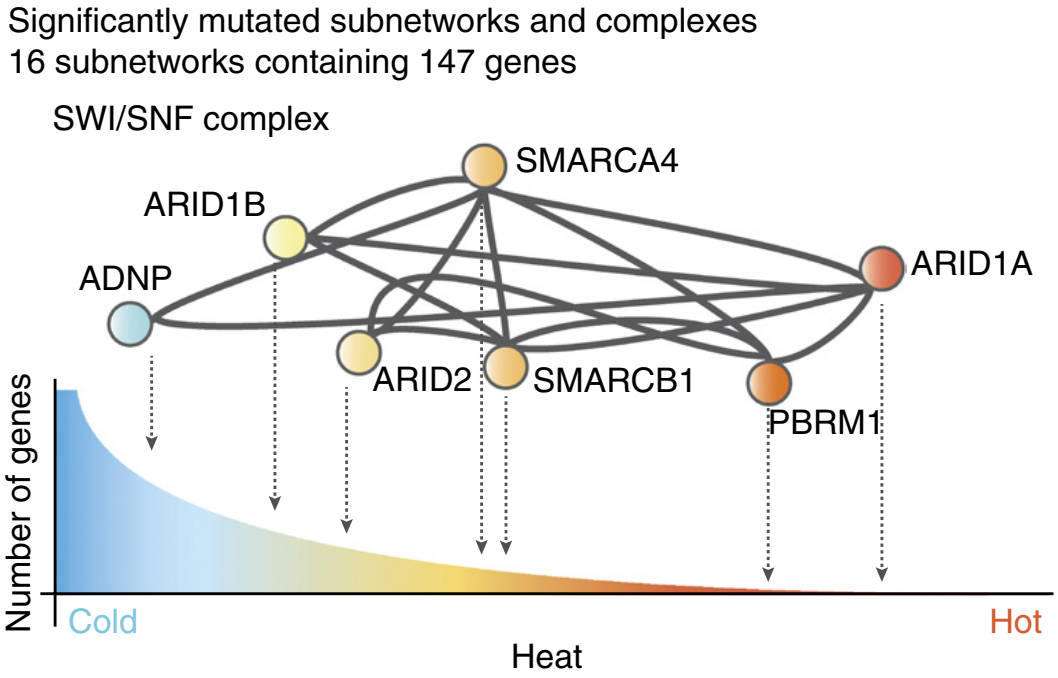


Very hot genes

Leiserson et al . 2014, Nature Genetics

HOTNET2 for pan-cancer mutation analysis



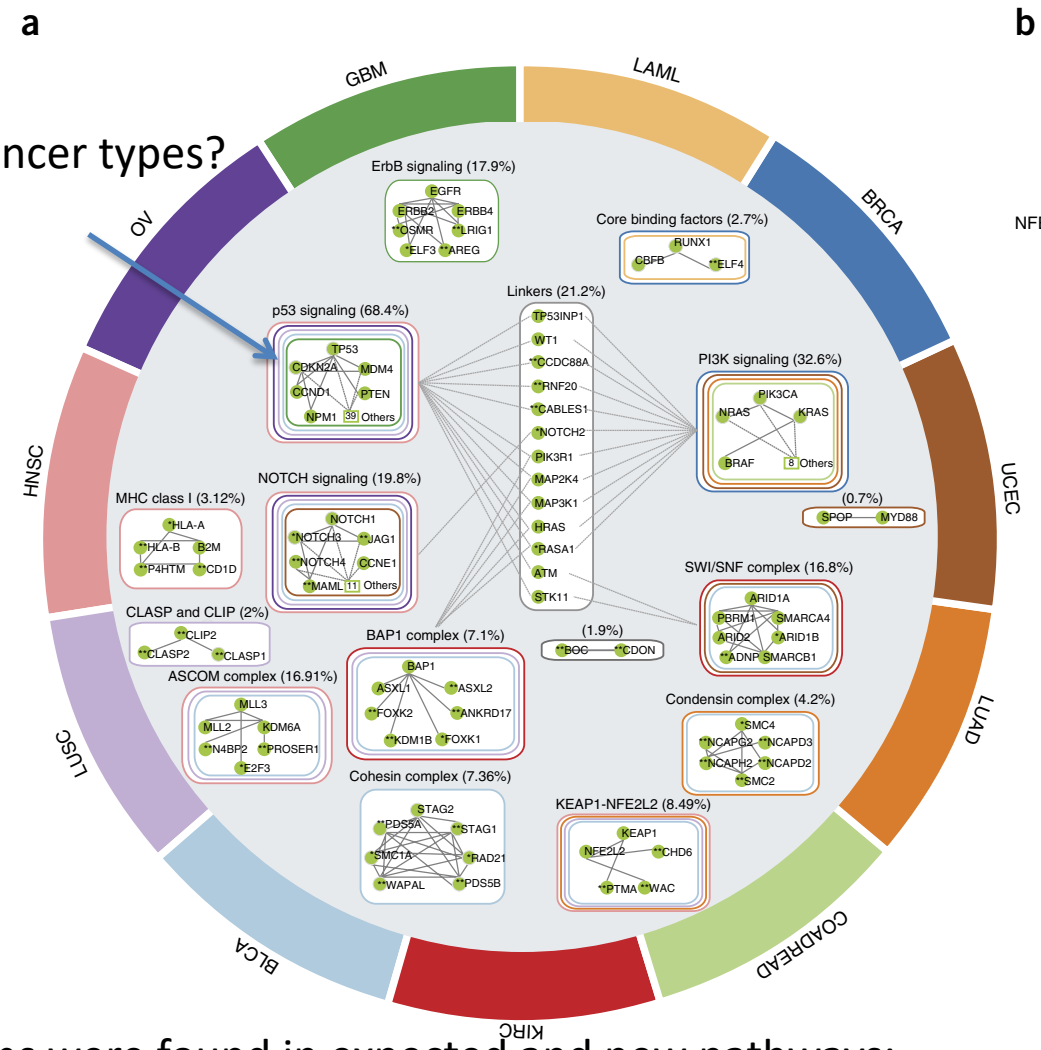


HOTNET2 subnetworks include genes with a wide range of mutation frequency

Leiserson et al . 2014, Nature Genetics

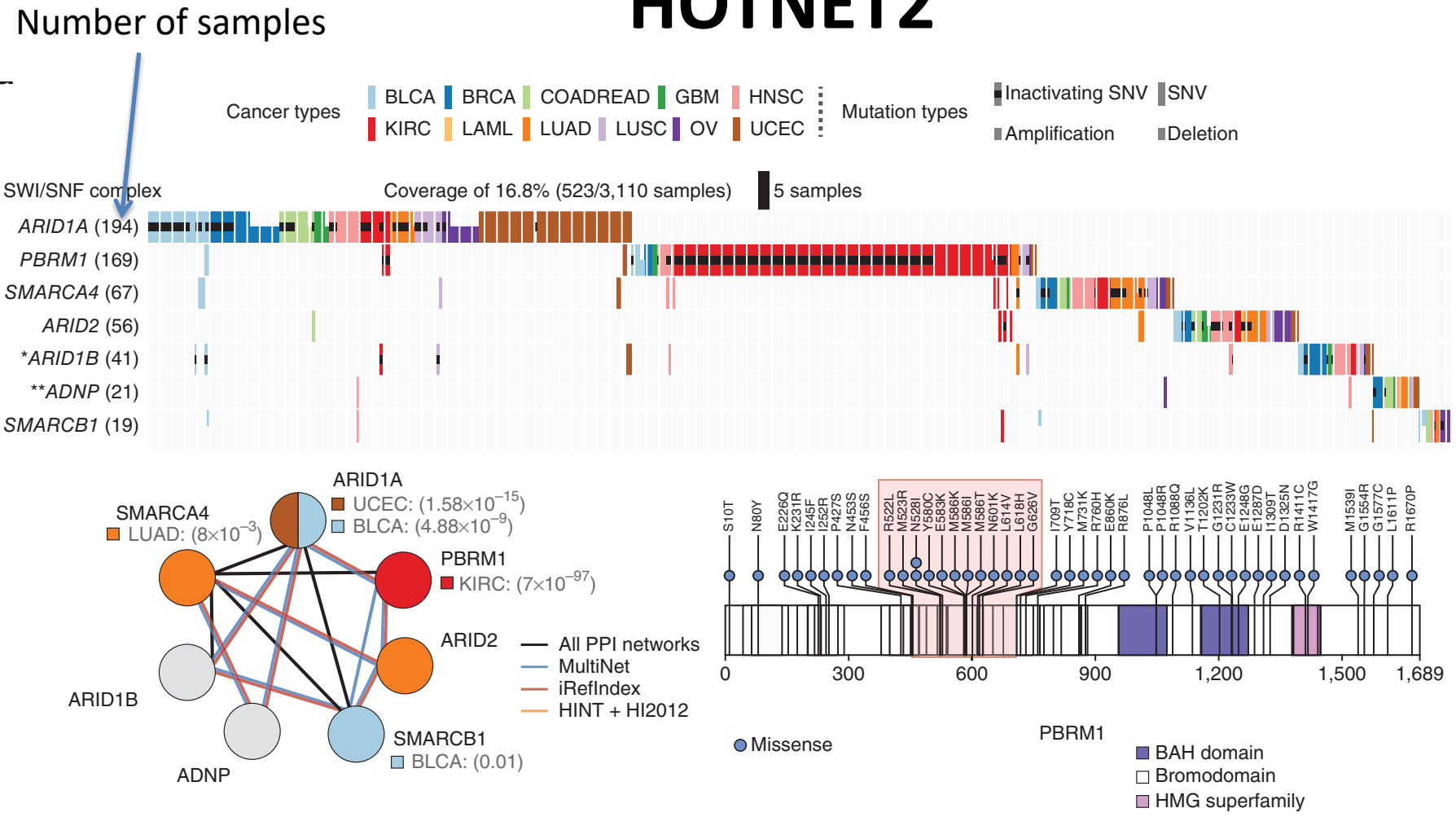
Overview of HOTNET2 results

Which cancer types?



Mutations were found in expected and new pathways:
PI3K, TP53, cohesin

SWI/SNF complex pathways identified by HOTNET2



Sixth most mutated Hotnet2 subnetwork.

HOTNET summary

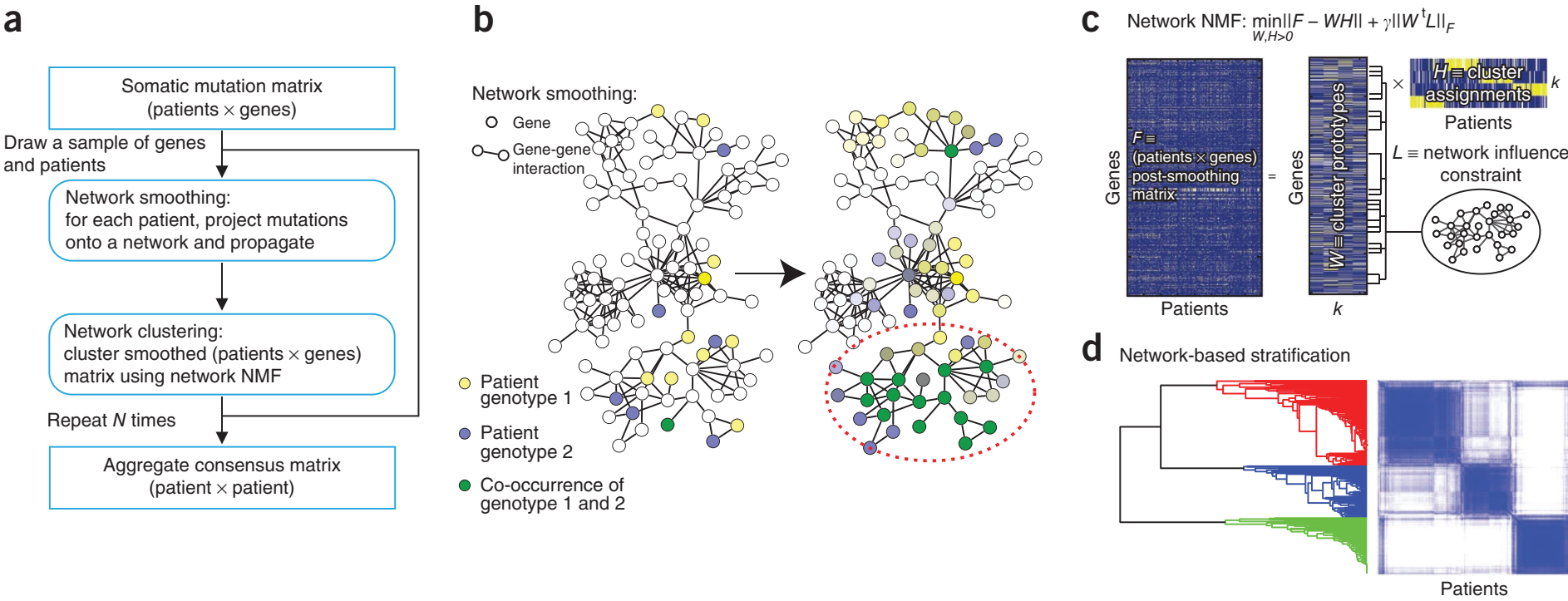
- An algorithm to find significantly mutated subnetworks
- Based on creating an “influence graph”, followed by identification of “interesting” subnetworks
- Non-local, less sensitive to network hubs
- Note: the subgraph detection component could be also addressed using module detection algorithms
- Post diffusion the graph could be used also for network-based stratification

Network-based stratification of patient samples

Input: Patient tumour mutation profiles, skeleton network

Output: Patient groups

How: (1) Smooth mutation profile using network smoothing; (2) Use Non-negative Matrix Factorization to cluster samples



Hofree et al. 2013, Nature methods

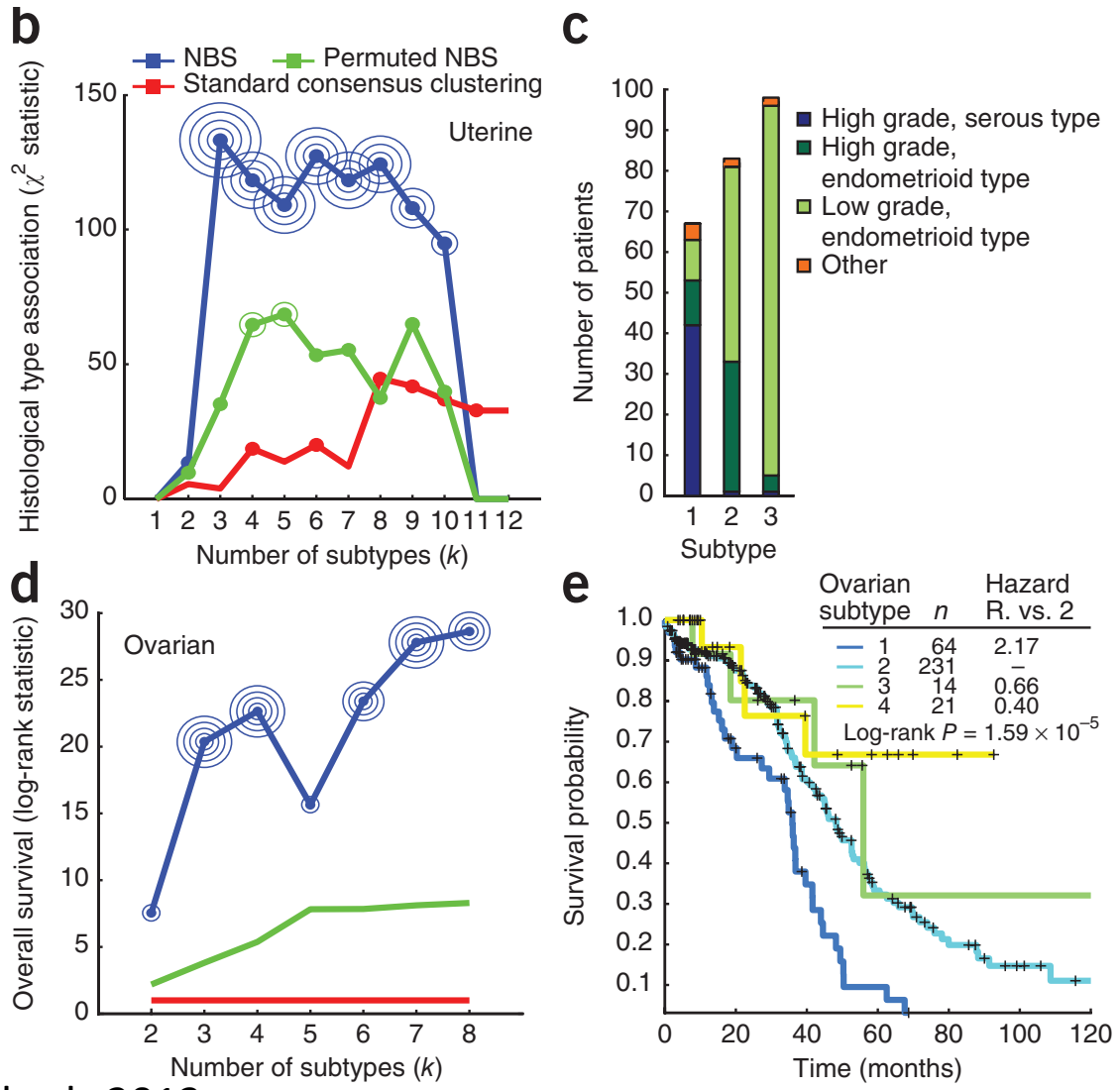
Network-based stratification of patient tumor samples

Uterine cancer

NBS subtypes associated with different histological types

Ovarian cancer

NBS subtypes associated with survival



Hofree et al., Nature Methods 2013

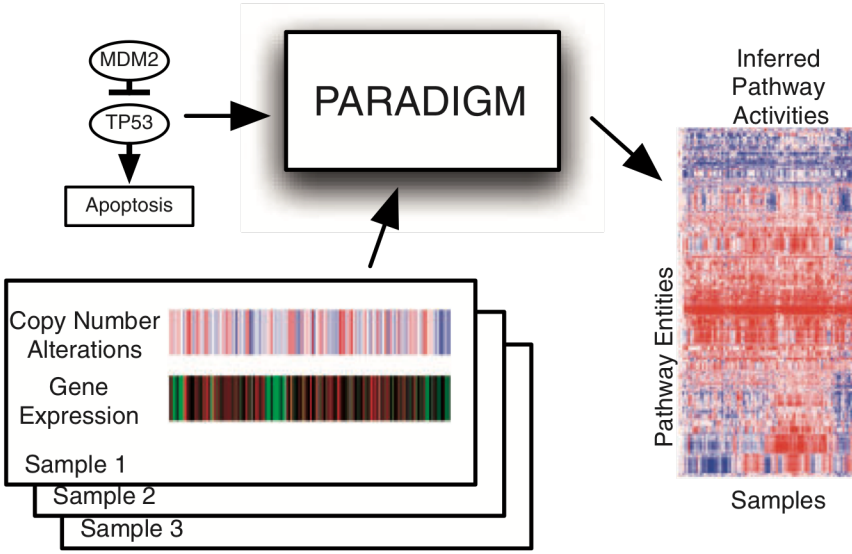
Types of algorithms used to examine perturbations in networks

- Graph diffusion followed by subnetwork finding methods
  - HOTNET
- Probabilistic graphical model-based methods
  - Factor graphs
  - Nested Effect Models (NEMs)
- Information flow-based methods (also widely used for integrating different types of data)
  - Min cost max flow
  - Prize collecting steiner tree

Probabilistic graphical models for interpreting network perturbations

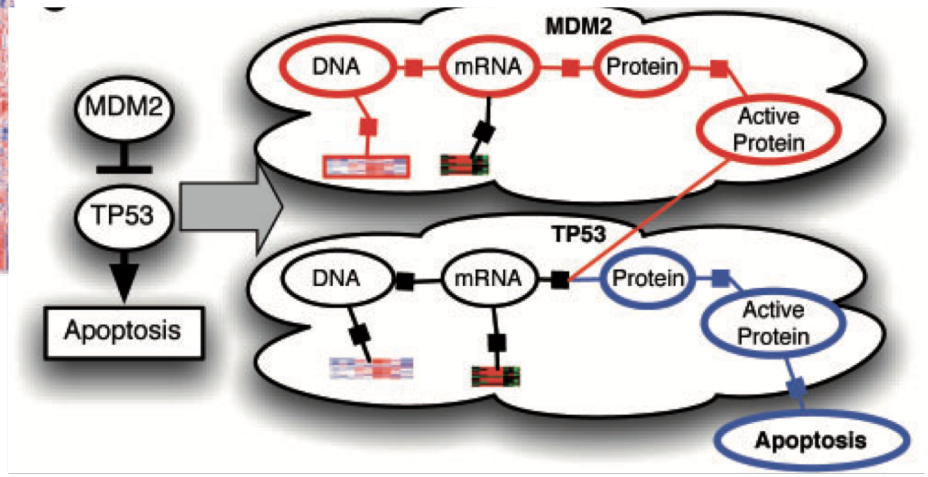
- “Inference of Patient-Specific Pathway Activities from Multi-Dimensional Cancer Genomics Data Using PARADIGM. Bioinformatics” https://academic.oup.com/bioinformatics/article/26/12/i237/282591
- C.-H. H. Yeang, T. Ideker, and T. Jaakkola, "Physical network models." *Journal of computational biology : a journal of computational molecular cell biology*, vol. 11, no. 2-3, pp. 243-262, Mar. 2004.
- F. Markowetz, D. Kostka, O. G. Troyanskaya, and R. Spang, "Nested effects models for high-dimensional phenotyping screens," *Bioinformatics*, vol. 23, no. 13, pp. i305-312, Jul. 2007.
- C. J. Vaske, C. House, T. Luu, B. Frank, C.-H. H. Yeang, N. H. Lee, and J. M. Stuart, "A factor graph nested effects model to identify networks from genetic perturbations." *PLoS computational biology*, vol. 5, no. 1, pp. e1 000 274+, Jan. 2009.

PARADIGM for detecting pathway activities



Overview of paradigm

Encode pathways as factor graphs



Factor graphs

- A type of graphical model
- A bi-partite graph with variable nodes and factor nodes
- Edges connect variables to potentials that the variables are arguments of
- Represents a global function as product of smaller local functions
- Perhaps the most general graphical model
  - Bayesian networks and Markov networks have factor graph representations

Example factor graph

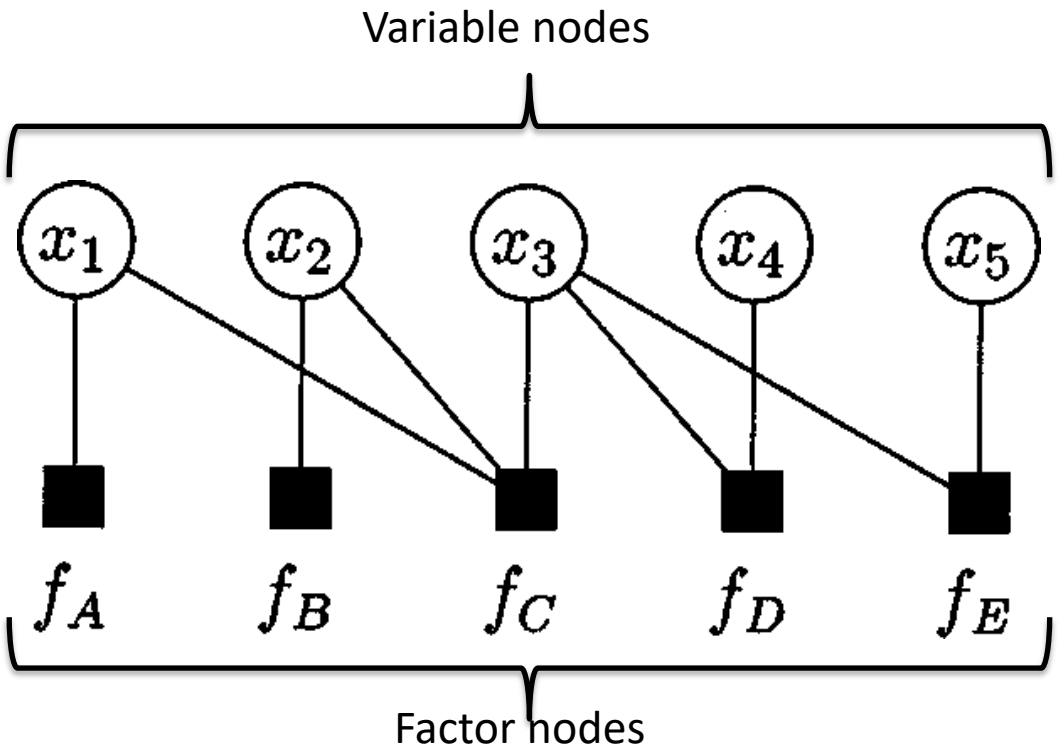


Fig. 1. A factor graph for the product $f_A(x_1)f_B(x_2)f_C(x_1, x_2, x_3) \cdot f_D(x_3, x_4)f_E(x_3, x_5)$.

From Kschischang, Frey, Loeliger 2001

Probabilistic graphical models for interpreting network perturbations

- “Inference of Patient-Specific Pathway Activities from Multi-Dimensional Cancer Genomics Data Using PARADIGM. Bioinformatics” https://academic.oup.com/bioinformatics/article/26/12/i237/282591
- C.-H. H. Yeang, T. Ideker, and T. Jaakkola, "Physical network models." *Journal of computational biology : a journal of computational molecular cell biology*, vol. 11, no. 2-3, pp. 243-262, Mar. 2004.
- F. Markowetz, D. Kostka, O. G. Troyanskaya, and R. Spang, "Nested effects models for high-dimensional phenotyping screens," *Bioinformatics*, vol. 23, no. 13, pp. i305-312, Jul. 2007.
- C. J. Vaske, C. House, T. Luu, B. Frank, C.-H. H. Yeang, N. H. Lee, and J. M. Stuart, "A factor graph nested effects model to identify networks from genetic perturbations." *PLoS computational biology*, vol. 5, no. 1, pp. e1 000 274+, Jan. 2009.

Motivation of nested effect models

- Perturbation of genes followed by high-throughput profiling of different phenotypes can be used to characterize functions of genes
- However, most genes do not function independently but interact in a network to drive a particular function
- Phenotypic measurements (e.g. mRNA levels) are indirect measurements of the underlying network structure
  - Includes direct and indirect effects
- Given perturbation data from multiple genes, can we more systematically identify the functions of these genes and how they interact at a pathway level?

Problem overview

- Given
  - global measurements of gene expression after single gene deletions of multiple genes
- Do
  - Infer interactions between genes with deletions to enable further characterization of these genes
- Nested Effect Models are probabilistic model-based approaches to solve this problem

Nested Effect Models

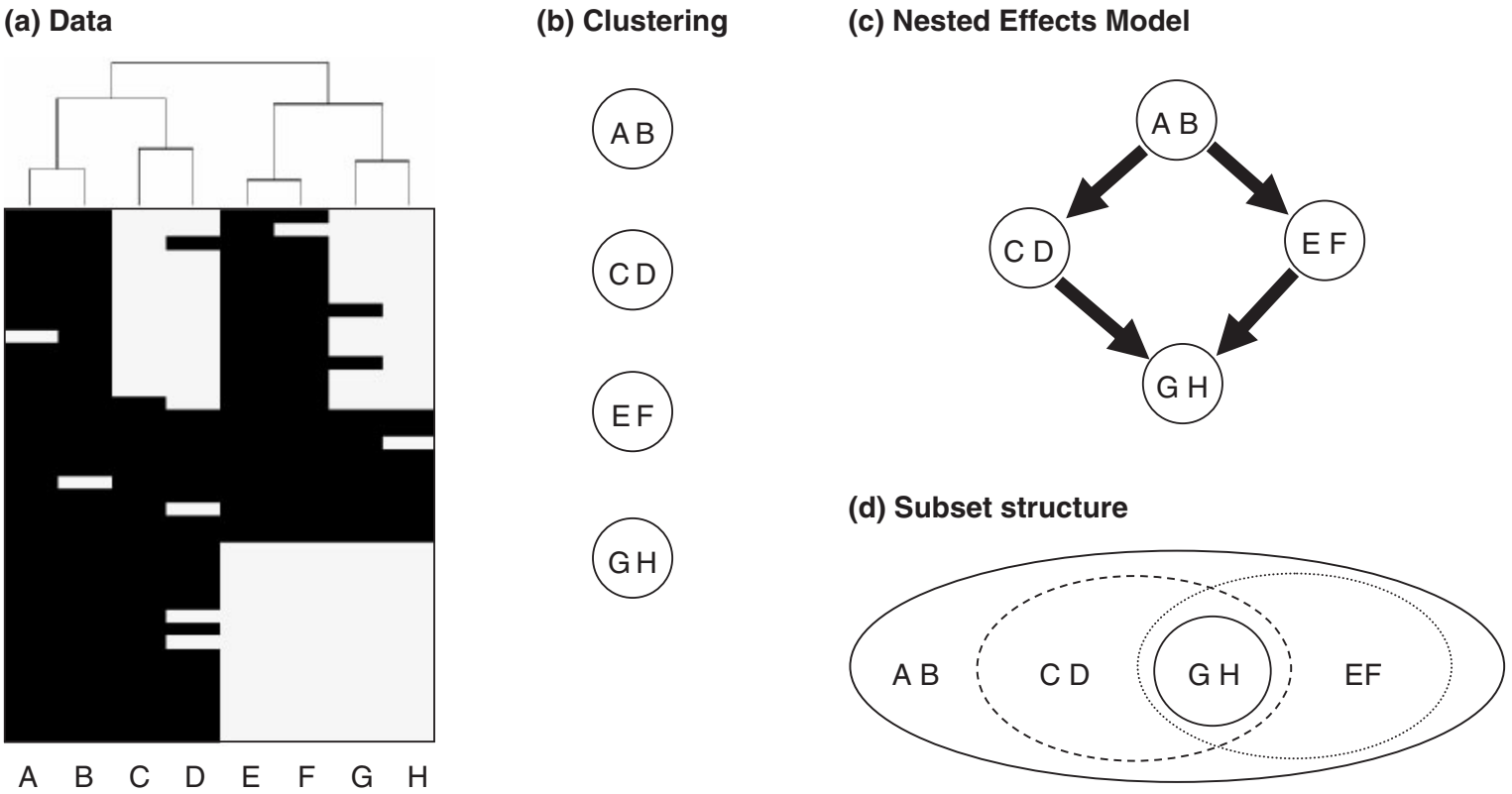


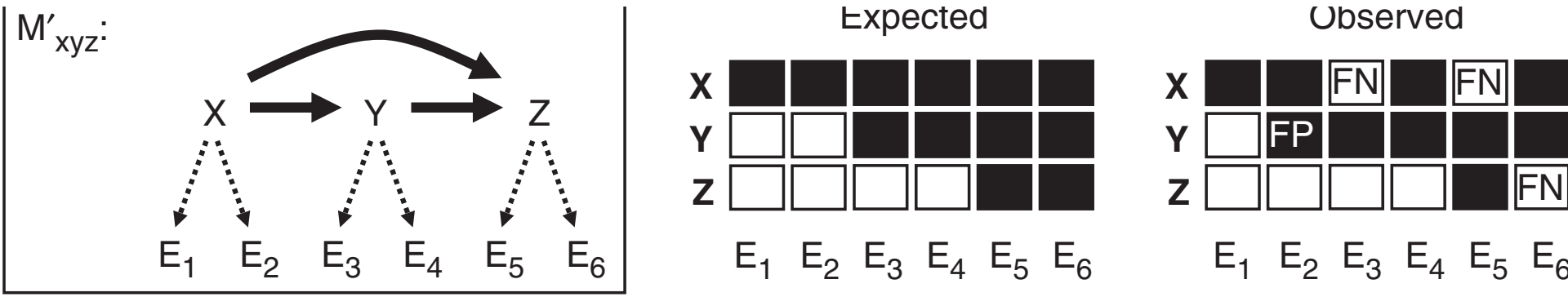
Fig. 1. An introduction to Nested Effects Models. Plot (**a**) shows a toy dataset consisting of phenotypic profiles for eight perturbed genes ($A, \ldots, H$). Each profile is binary with *black* coding for an observed effect and *white* for an effect not observed. The eight profiles are hierarchically clustered, showing that they fall into four pairs of genes with almost identical phenotypic profiles: (A, B), (C, D), (E, F) and (G, H), as shown in plot (**b**). An important feature of the data missed by clustering is the subset structure visible between the profiles in the data set: the effects observed when perturbing genes A or B are a superset to the effects observed for all other genes. The effects of perturbing G or H are a subset to all other genes' effects. The pairs (C, D) and (E, F) have different but overlapping effect sets. The directed acyclic graph (DAG) shown in plot (**c**) represents these subset relations, which are shown in plot (**d**). Compared to the clustering result in plot (b) the NEM additionally elucidates relationships between the clusters and thus describes the dominant features of the data set better.

Markowetz et al, 2007

Nested Effect Models Key properties

- A generalization of similarity based clustering
- Orders the clusters according to subset relationships
  - A gene A is upstream of another gene B if B's effects are a subset of A's effects
- Build a hierarchy of all perturbed genes by constructing from smaller sub-models of pairs and triplets of genes

Subset relationships to order genes



A complete model. The left part of the figure shows a complete model M'_{xyz} consisting of a transitively closed graph between genes and assignments of genes to specific effects (the dashed arrows). Given the complete model, we can formulate a prediction of what effects to expect: perturbing x should cause all effects, while perturbing y should only cause E3–E6, and perturbing z only E5 and E6 (middle plot). In reality, our observations will be noisy: there can be false positive (FP) and false negative (FN) effect observations (right plot).

Probabilistic graphical models for interpreting network perturbations

- "Inference of Patient-Specific Pathway Activities from Multi-Dimensional Cancer Genomics Data Using PARADIGM. Bioinformatics" https://academic.oup.com/bioinformatics/article/26/12/i237/282591
- C.-H. H. Yeang, T. Ideker, and T. Jaakkola, "Physical network models." *Journal of computational biology : a journal of computational molecular cell biology*, vol. 11, no. 2-3, pp. 243-262, Mar. 2004.
- F. Markowetz, D. Kostka, O. G. Troyanskaya, and R. Spang, "Nested effects models for high-dimensional phenotyping screens," *Bioinformatics*, vol. 23, no. 13, pp. i305-312, Jul. 2007.
- C. J. Vaske, C. House, T. Luu, B. Frank, C.-H. H. Yeang, N. H. Lee, and J. M. Stuart, "A factor graph nested effects model to identify networks from genetic perturbations." *PLoS computational biology*, vol. 5, no. 1, pp. e1 000 274+, Jan. 2009.

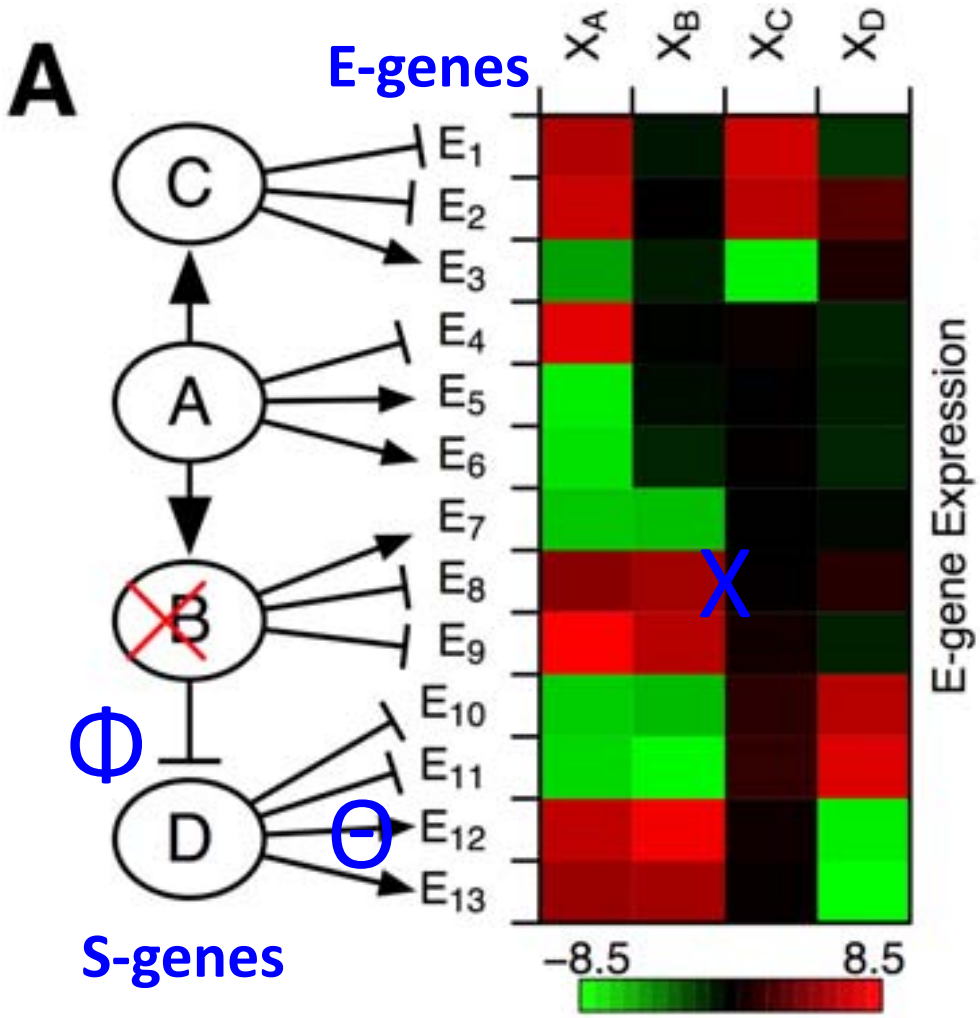
Key properties of Factor Graph-NEMs (FG-NEMs)

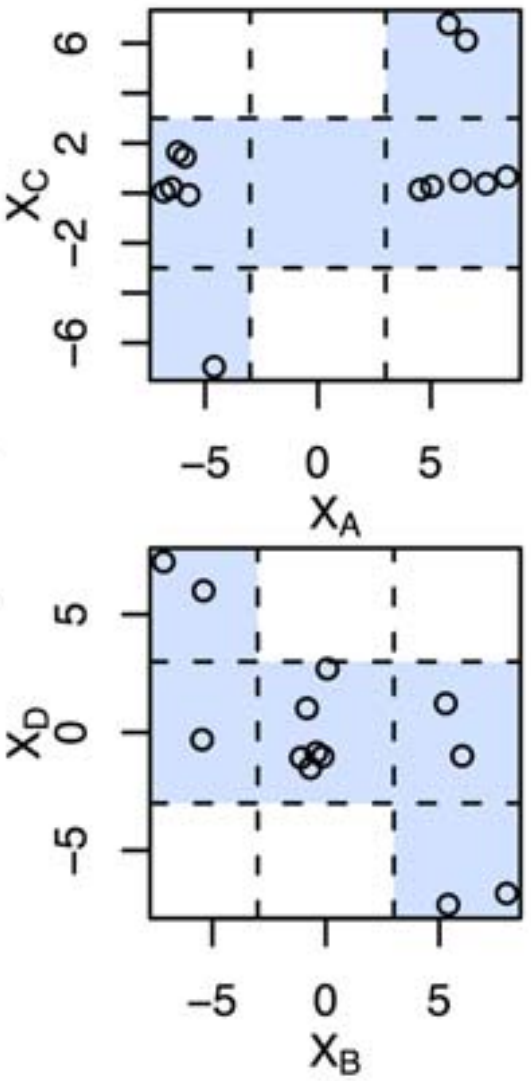
- NEMs assume the genes that are perturbed interact in a binary manner
- But many interactions have sign
  - inhibitory or stimulating action
- FG-NEMs capture a broader set of interactions among the perturbed genes
- Formulation based on a Factor Graph
  - Provide an efficient search over the space of NEMs

Notation

- S-genes: Set of genes that have been deleted individually
- E-genes: Set of effector genes that are measured
- Θ: The attachment of an effector gene to the S-gene network
- Φ: The interaction matrix of S-genes
- X: The phenotypic profile, each column gives the difference in expression in a knockout compared to wild type
  - Rows: E-genes
  - Columns: S-genes
- Y: Hidden effect matrix, each entry is {-1, 0, +1} which specifies whether an S-gene affects the E-gene

An example of 4 S-genes and 13 E-gens

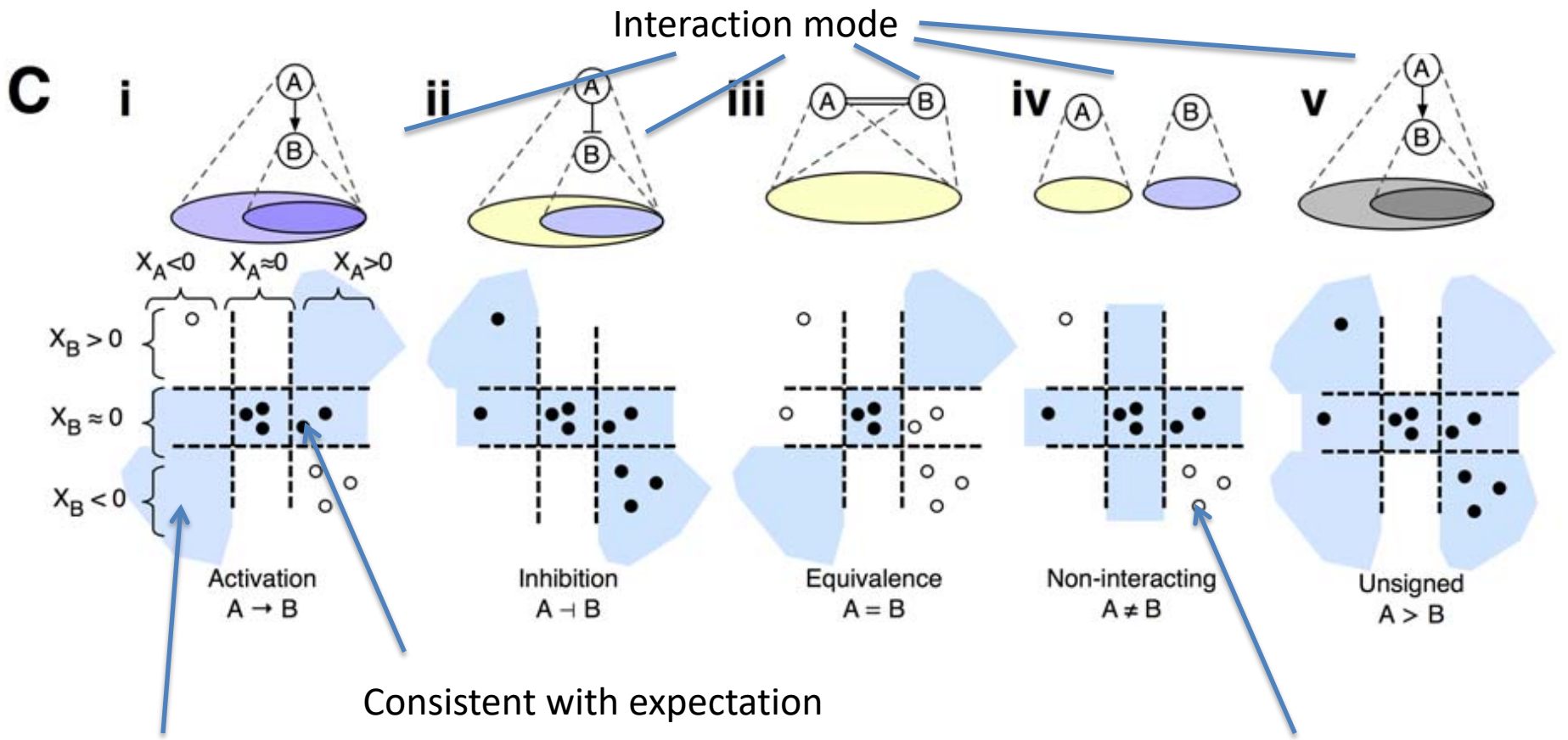




A->C is reflected in the scatter plot.
When X_A is up, X_C is up. When X_A is down, X_C is down or no change

B-ıD is also reflected in the scatter plot.
X_D is a subset of opposite changes from X_B

S-gene interaction modes and their expression signatures

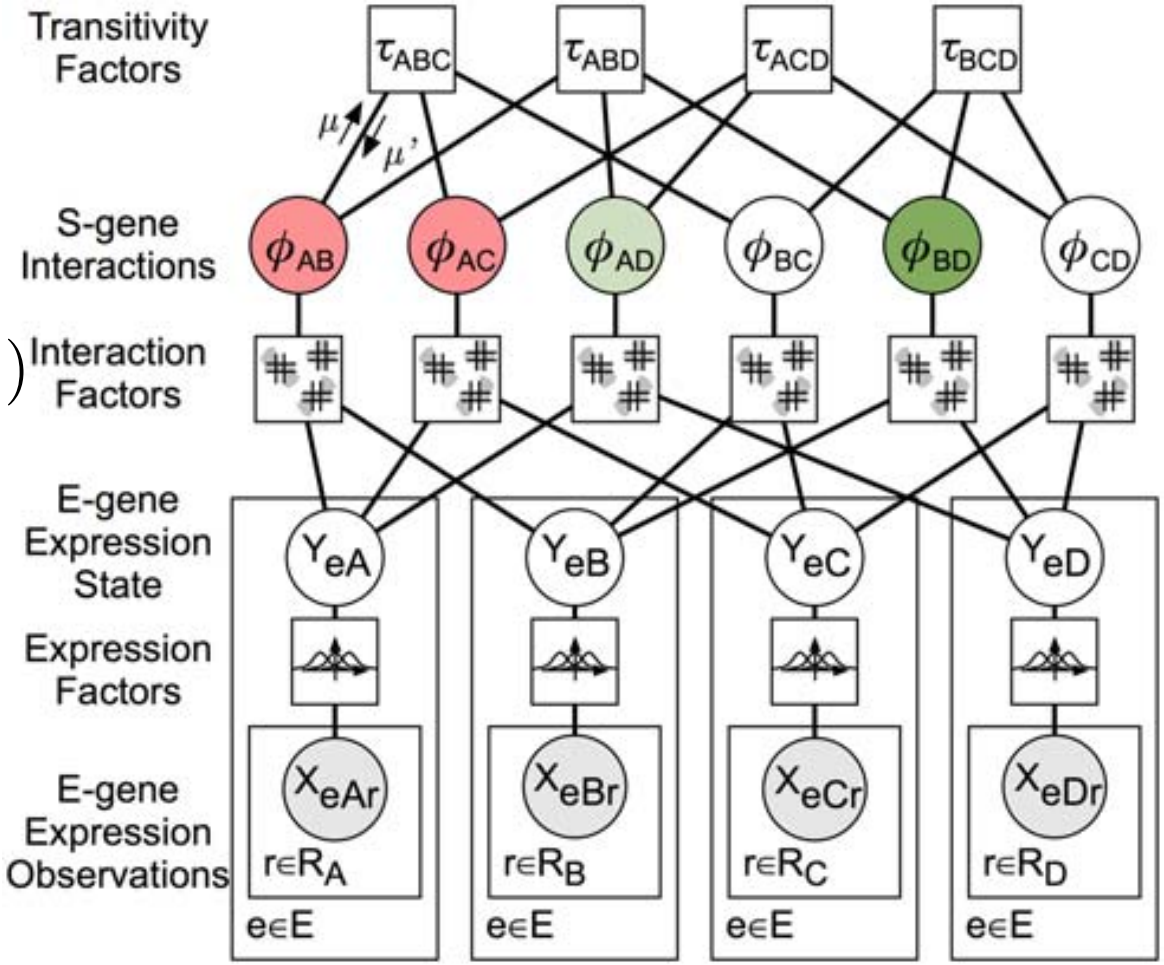


Factor graph representation of NEMs

Prior

$$P(Y_{eA}, Y_{eB} | \phi_{A,B}, \theta_{eAB})$$

$$P(X_{eA} | Y_{eA})$$



Probabilistic model for NEMs

- Goal is to find a network, Φ and Θ that best fit the observed data (*X*)
- This is an inference problem
- Use a Maximum a posterior (MAP) approach

$$J(X) = \max_{\phi,\theta} P(\phi, \theta | X)$$

$$J(X) = \max_{\phi,\theta} \sum_{Y} P(\phi, \theta, Y | X)$$

- Makes use pairwise potentials to make the computation tractable

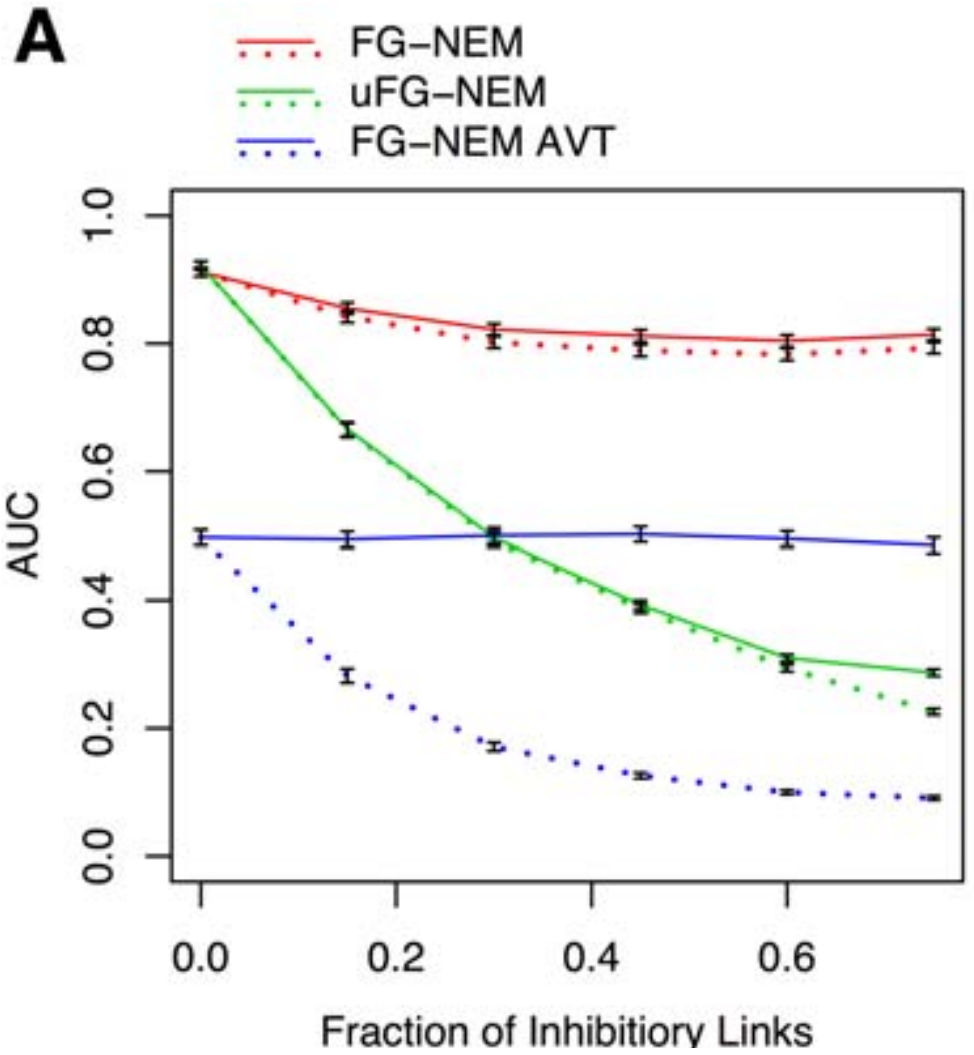
X is a noisy measurement of Y. Y is the quantity we need to sum over

Inference on the factor graph

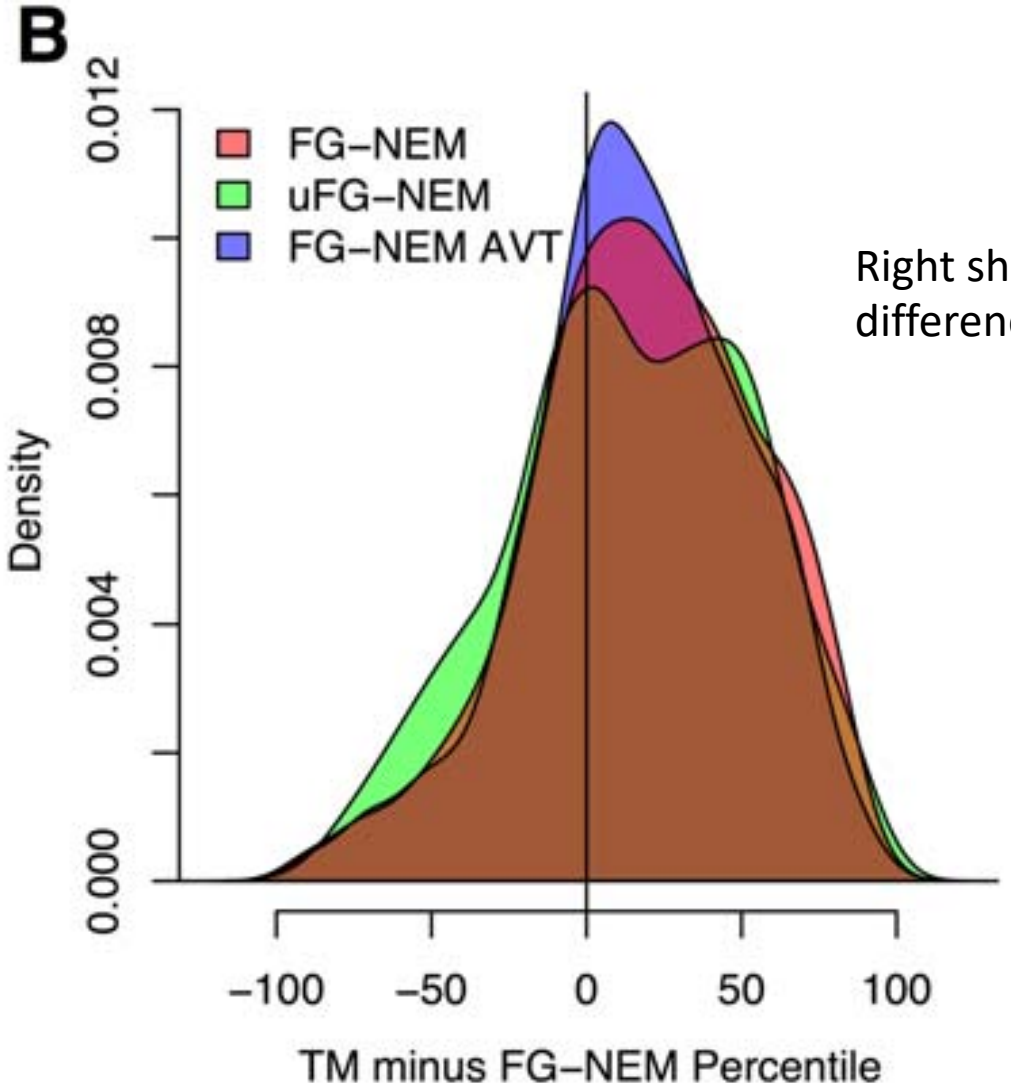
- Find most likely configurations for $\phi_{A,B}$
- Use a message passing algorithm (standard for factor graphs)
- Called the Max-Product algorithm
- Message passing happens in two steps
  - Messages are passed from observations X_{eA} to the $\phi_{A,B}$
  - Messages are passed between the interaction and transitivity factors until convergence

Does FG-NEM capture activating and inhibitory relationships?

FG-NEM: capture inhibitory and activating relationships
uFG-NEM: capture only unsigned interactions
FG-NEM AVT: FG-NEM run on absolute value data
Solid lines: structure recovery
Dashed lines: sign recovery



Does FG-NEM expand pathways better than the baseline approach



Right shift in percentile rank difference of NEM methods

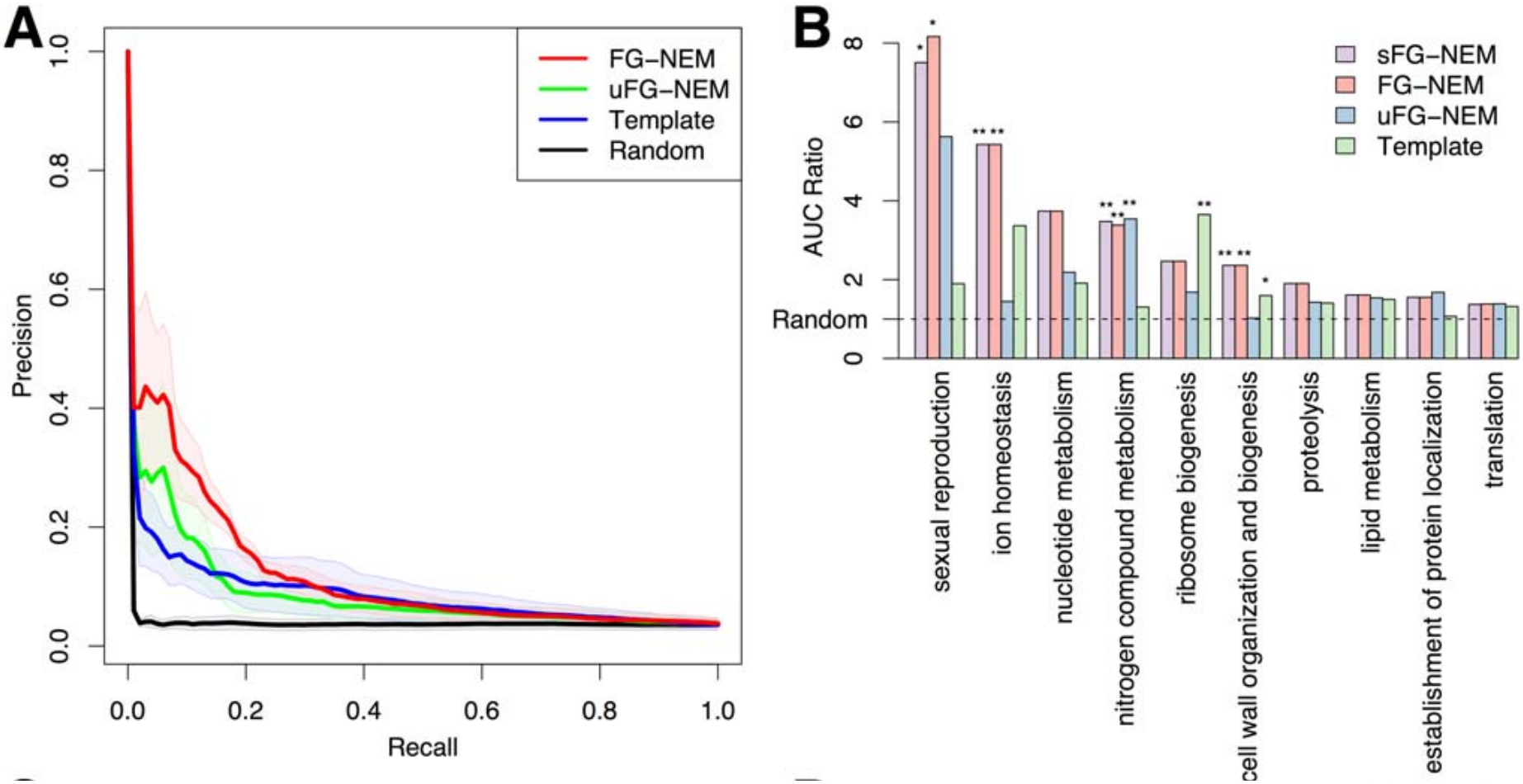
Pathway expansion

- Attach new E-genes to S-gene network
- An attached gene e to S-gene s asserts that e is directly downstream of s
- All E-genes attached to the S-gene network are called frontier genes
- An E-gene's connectivity is examined based on the Log-likelihood Attachment Ratio

One of the S genes

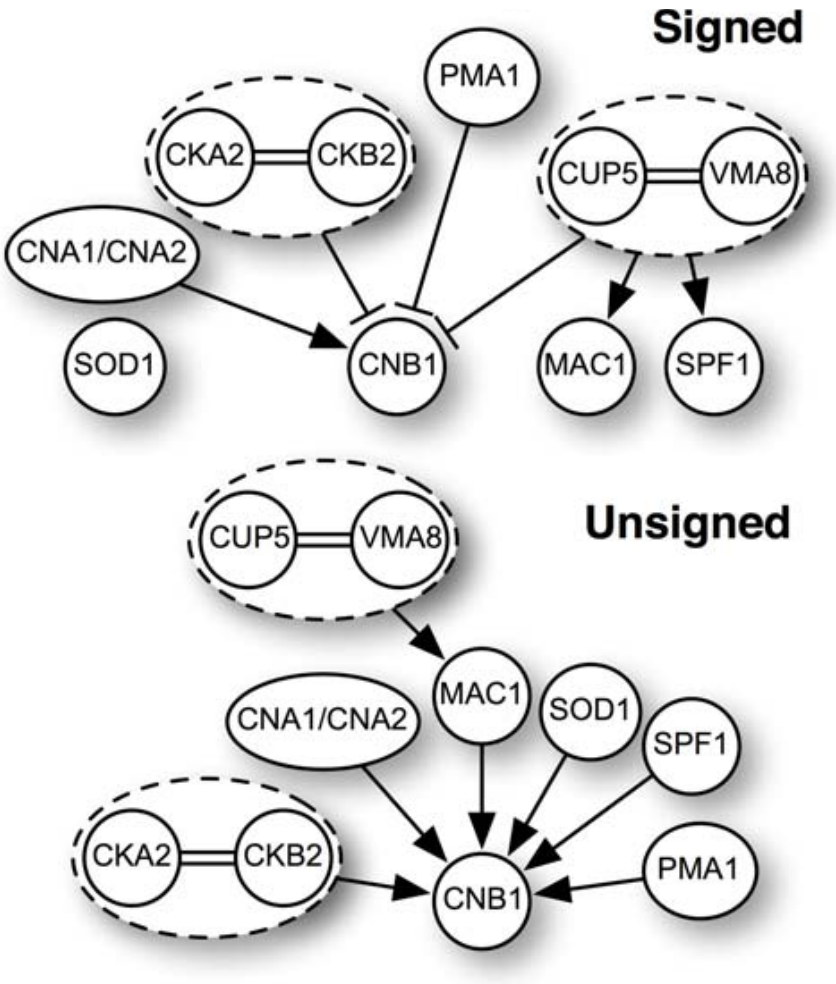
$$LAR(e) = \log\left(\frac{\max_{i \neq 0} P(X_e|\Phi,\theta_e = i)}{P(X_e|\Phi,\theta_e = 0)}\right)$$

FG-NEM based pathway expansion in yeast



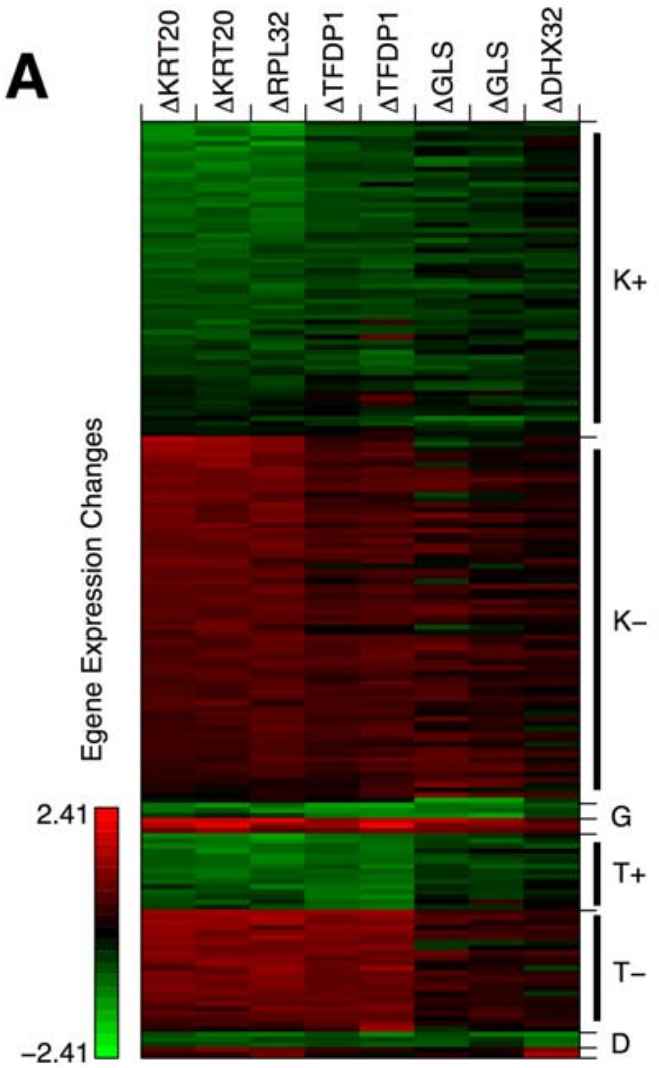
Template matching: rank E genes based on similarity in expression to an “idealized template”

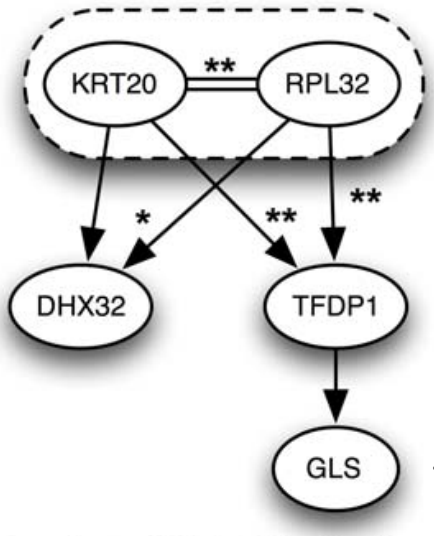
FG-NEM infers a more accurate network than the unsigned version in yeast



- FG-NEM and uFG-NEM networks inferred in the ion-homeostasis pathway
- FG-NEM inferred more genes associated with ion homeostatis compared to uFG-NEM

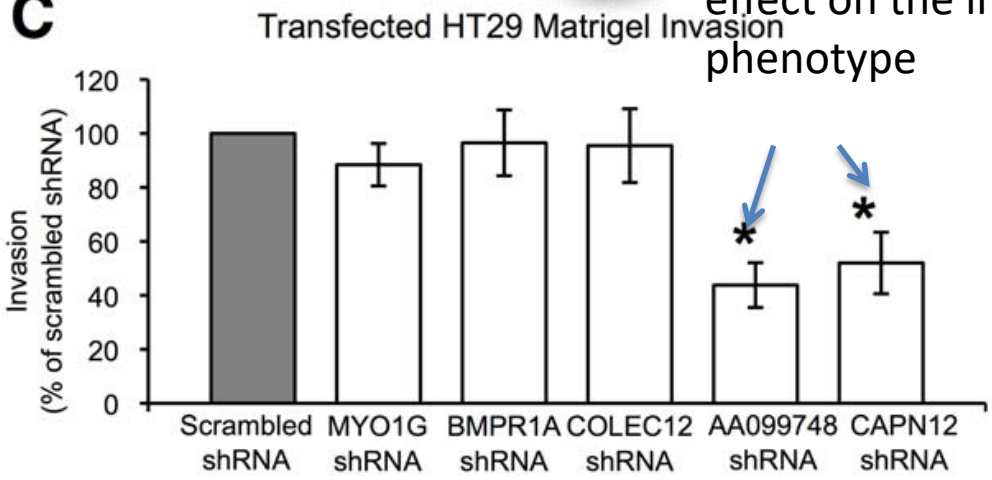
FG-NEM application to colon cancer





Most interactions in S-gene network are activating

Novel expanded genes that have significant effect on the invasive phenotype



Summary

- FG-NEMs: A general approach to infer an ordering of genes from knock-down phenotypes
- Strengths
  - FG-NEMs could be used in an iterative computational-experimental framework
  - Handles signed interactions between S-genes
- Weaknesses
  - Computational complexity of the inference procedure might be high
    - Required independence among E-genes
    - Model pairs of S-genes at a time

Overall conclusion

- Networks are powerful models for interpreting sequence variants or genetic perturbations as such
- We have see two classes of methods
  - Extract a weighted graph based on the influence of a mutation on one node to another
  - Probabilistic approaches
- A systematic comparison of these two classes of methods has not been done so far.