

Network-based interpretation and integration

Sushmita Roy

sroy@biostat.wisc.edu

Computational Network Biology

Biostatistics & Medical Informatics 826

<https://compnetbiocourse.discovery.wisc.edu>

Nov 29th 2018

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)
 - 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.

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

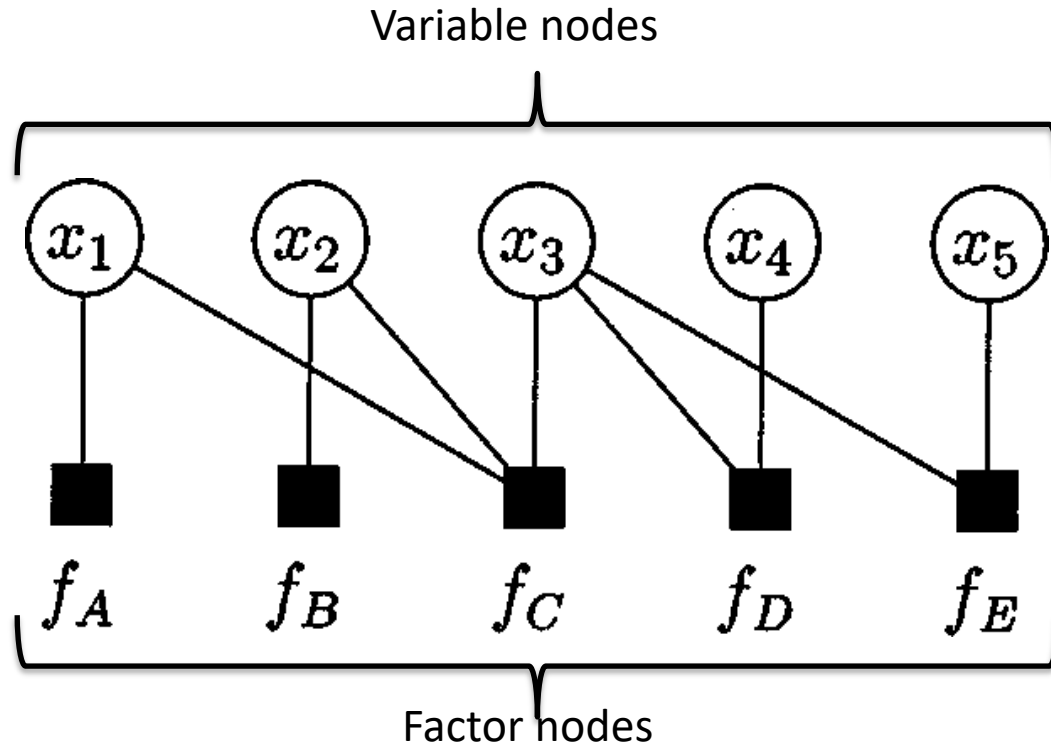


Fig. 1. A factor graph for the product $f_A(x_1)f_B(x_2)f_C(x_1, x_2, x_3) \cdot f_D(x_3, x_4)f_E(x_3, x_5)$.

Probabilistic graphical models for interpreting network perturbations

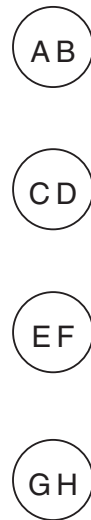
- “Inference of Patient-Specific Pathway Activities from Multi-Dimensional Cancer Genomics Data Using PARADIGM. *Bioinformatics*” <https://academic.oup.com/bioinformatics/article/26/12/i237/282591>
- C.-H. H. Yeang, T. Ideker, and T. Jaakkola, "Physical network models." *Journal of computational biology : 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.

Nested Effect Models

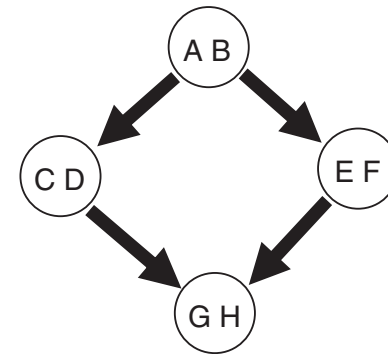
(a) Data



(b) Clustering



(c) Nested Effects Model



(d) Subset structure

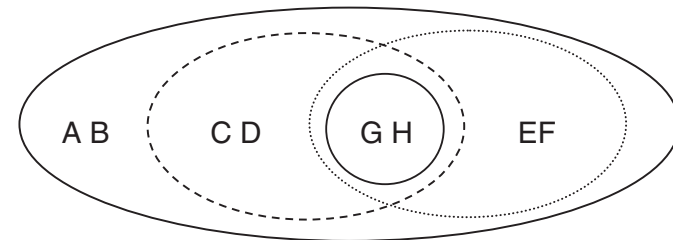


Fig. 1. An introduction to Nested Effects Models. Plot (a) shows a toy dataset consisting of phenotypic profiles for eight perturbed genes (A, \dots, H). Each profile is binary with *black* coding for an observed effect and *white* for an effect not observed. The eight profiles are hierarchically clustered, 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.

Markowitz et al, 2007

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

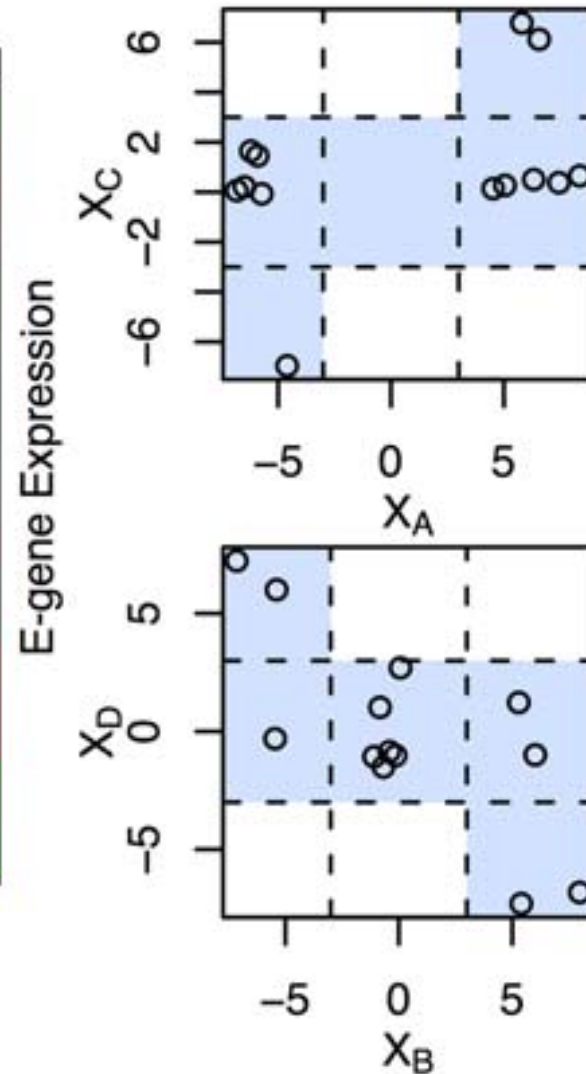
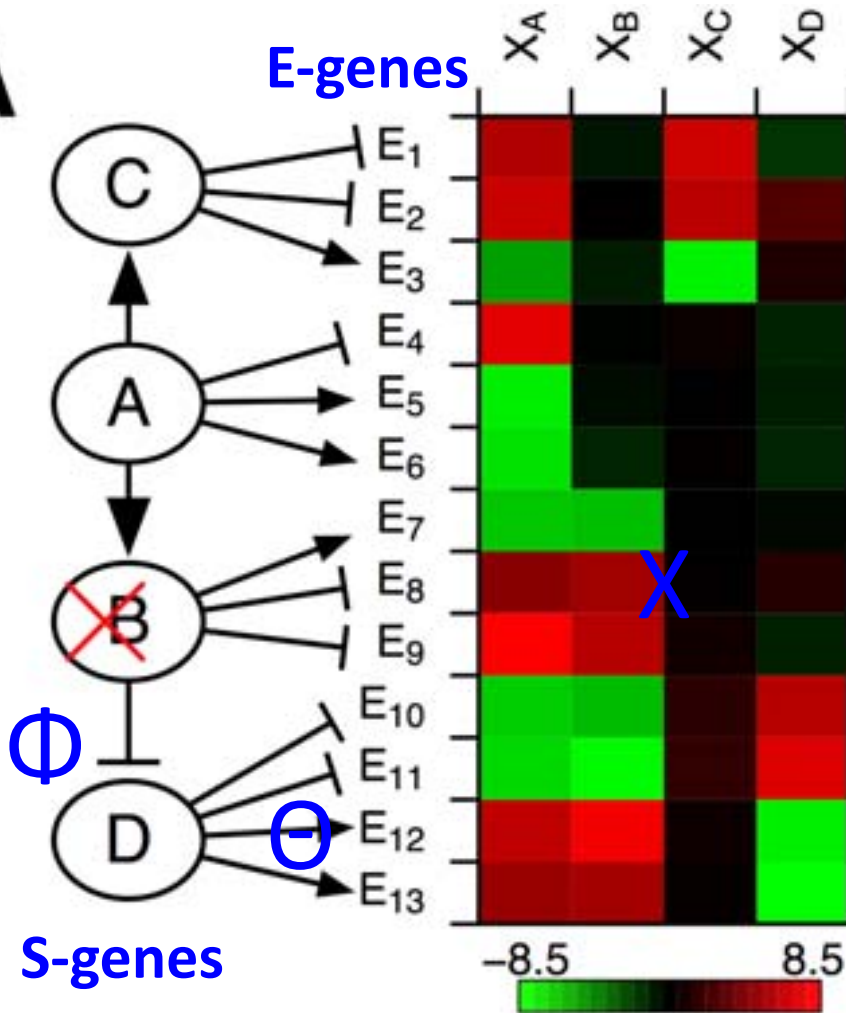
- NEMs assume the genes that are perturbed interact in a binary manner
- But many interactions have sign
 - inhibitory or stimulating action
- FG-NEMs capture a broader set of interactions among the perturbed genes
- Formulation based on a Factor Graph
 - Provide an efficient search over the space of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

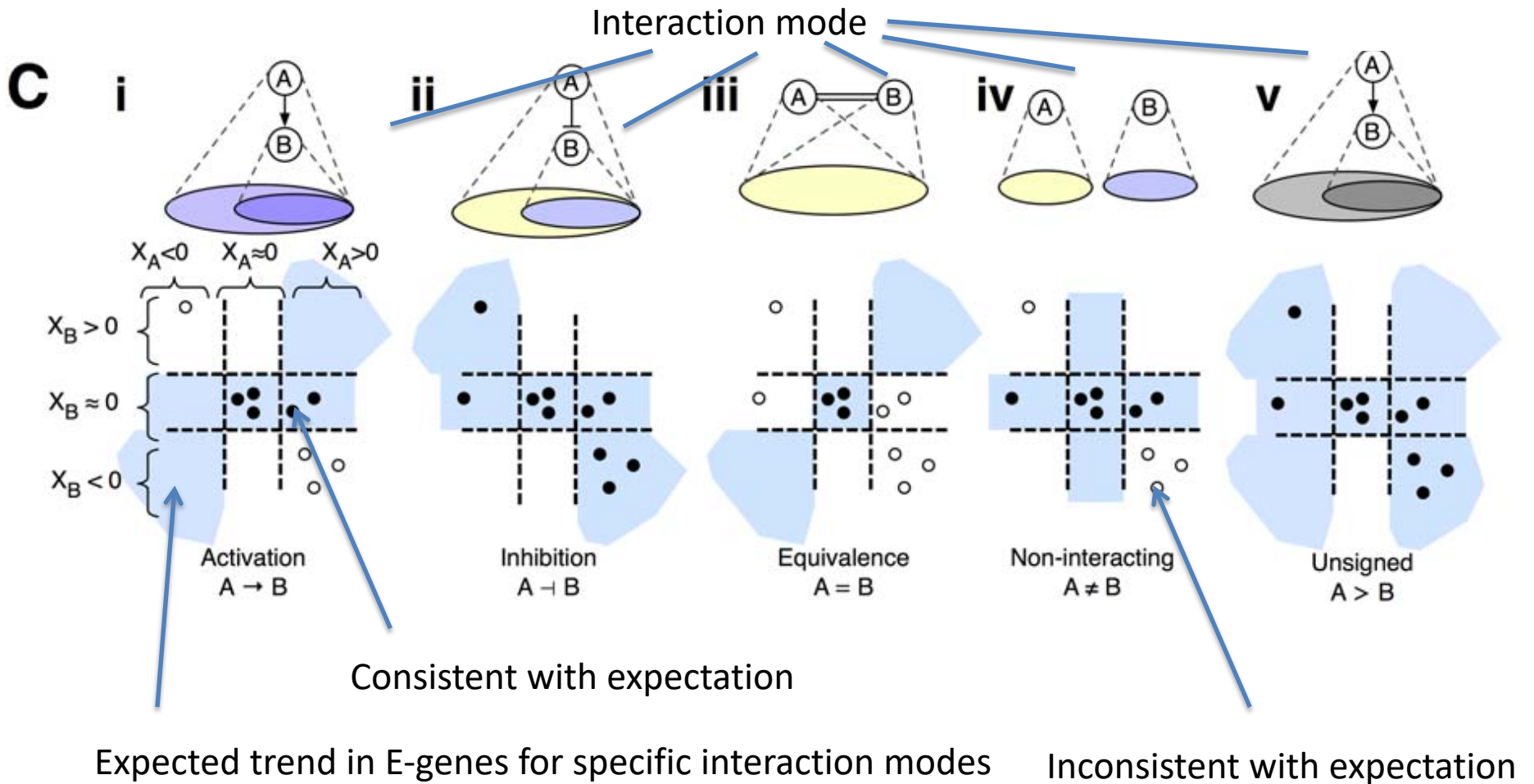
A



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→D 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



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

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

- Y encodes the “true” expression state of effector genes (X).
- X is a noisy measurement of Y . Y is the quantity we need to sum over

Probabilistic model continued

$$J(X) = \max_{\Phi, \Theta} \left\{ P(\Phi) \sum_Y \prod_{e \in E} P(Y_e | \Phi, \theta_e) P(X_e | Y_e) \right\}$$

Independence over all E-genes


$$= \max_{\Phi, \Theta} \left\{ P(\Phi) \prod_{e \in E} \sum_Y P(Y_e | \Phi, \theta_e) P(X_e | Y_e) \right\}$$

Re-arranging the terms

$$= \max_{\Phi, \Theta} \left\{ P(\Phi) \prod_{e \in E} L_e \right\}$$

Digging inside the L_e term

- Note: $Y_e = \{Y_{eA}, Y_{eB}, Y_{eC} \dots Y_{eN}\}$, where N is the total number of S-genes
- Define L'_e , proportional to L_e using a set of pairwise potentials

$$L'_e = \sum_{A, B \in S} \prod_{\substack{Y_{eA}, \\ Y_{eB}}} P(Y_{eA}, Y_{eB} | \phi_{AB}, \theta_{eAB}) P(X_{eA} | Y_{eA}) P(X_{eB} | Y_{eB})$$


- $\phi_{A,B}$ The S- gene interaction
- θ_{eAB} Attachment of gene e with respect to A or B

Digging inside the L_e term


- $\phi_{A,B}$ The S- gene interaction
- θ_{eAB} Attachment of gene e with respect to A or B

Now the joint can be written in a more tractable way

$$J(X) = \max_{\Phi} \left\{ P(\Phi) \prod_{\substack{e \in E, \\ A, B \in S}} \max_{\theta_{eAB}} \right. \\ \left. \sum_{Y_{eA}, Y_{eB}} P(Y_{eA}, Y_{eB} | \phi_{AB}, \theta_{eAB}) P(X_{eA} | Y_{eA}) P(X_{eA} | Y_{eA}) \right\}$$

Each of these conditional distributions will correspond to a factor

Defining the factors

$$P(Y_{eA}, Y_{eB} | \phi_{AB}, \theta_{eAB}) P(X_{eA} | Y_{eA}) P(X_{eA} | Y_{eA})$$


Four variable factor, over discrete variables

Y_{eA} : binary variables

Modeled as Gaussian distributions

ϕ_{AB} Four values for each possible type of interaction: inhibitory, activating, equivalent, no interaction

θ_{eAB} Interaction of e with A or B : inhibited or activated by A or B or no action

This factor has value=1 if the E-gene e is attached to either A or B and e 's state is consistent with the interaction mode between A and B .

The prior over S-gene graph

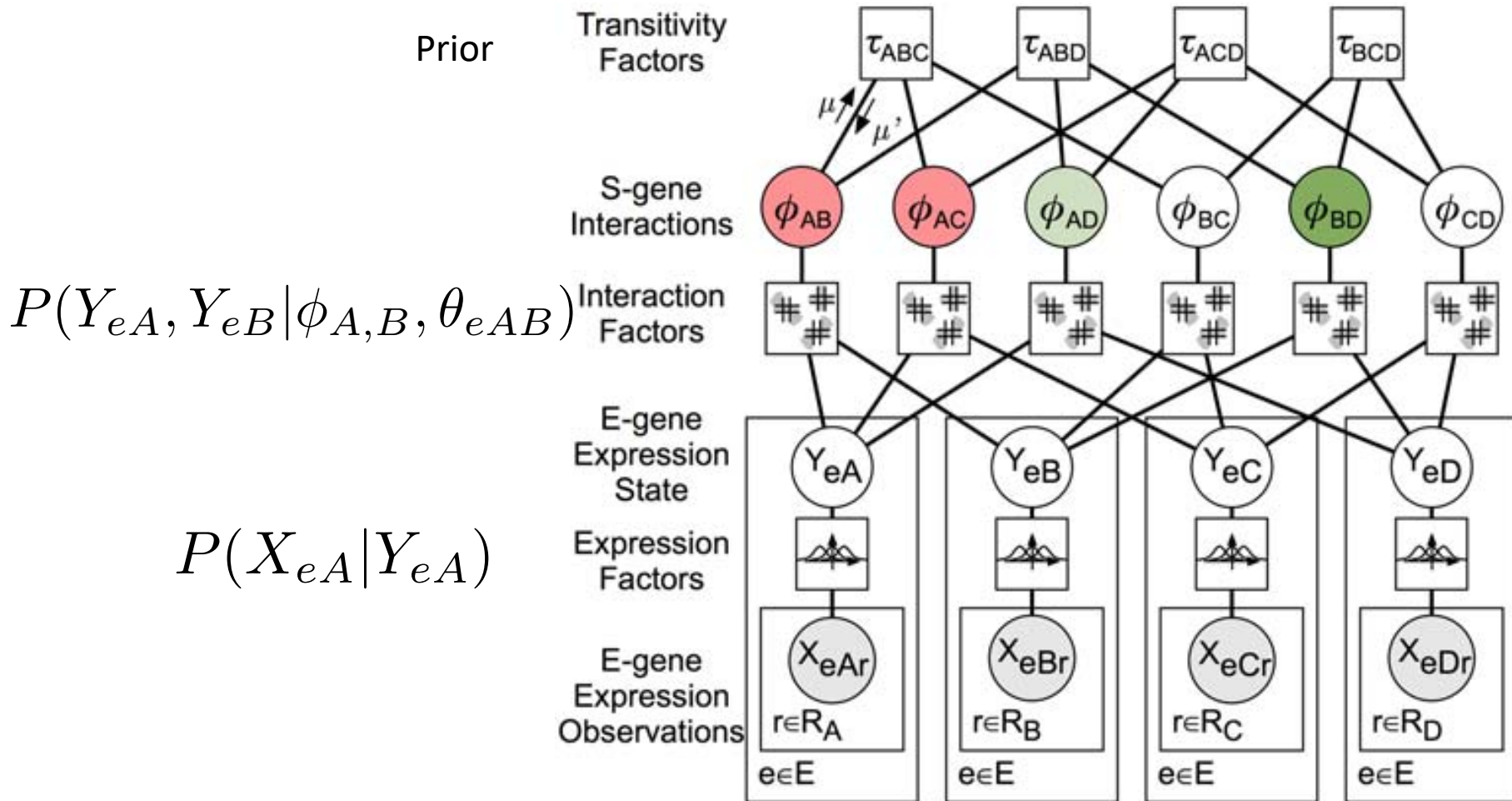
- The prior $P(\Phi)$ can incorporate prior knowledge of interactions among genes in pathways
- At its simplest, it should encode a transitivity relationship to force all pairwise interactions to be consistent among all triples

$$P(\Phi) \propto \left(\prod_{A,B,C \in S} \tau_{ABC}(\phi_{AB}, \phi_{BC}, \phi_{AC}) \right) \left(\prod_{A,B \in S} \rho_{AB}(\phi_{AB}) \right)$$

Transitivity constraint for triples Physical network constraints

Example transitivity: If $A \rightarrow B$, $B \rightarrow C$, Then, $A \rightarrow C$

Factor graph representation of NEMs

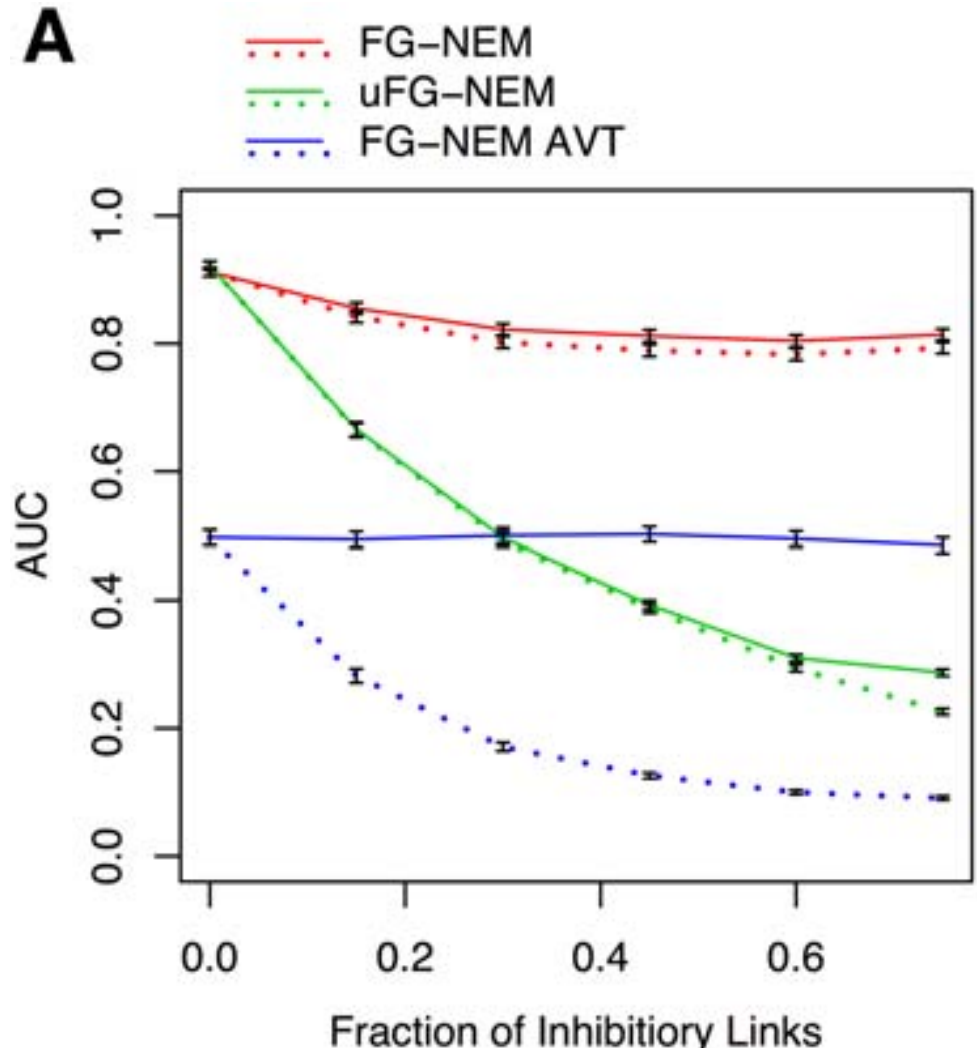


Inference on the factor graph

- Find most likely configurations for $\phi_{A,B}$
- Use a message passing algorithm called the Max-Product algorithm (standard for factor graphs)
- Message passing happens in two steps
 - Messages are passed from observations X_{e_A} 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




Pathway expansion

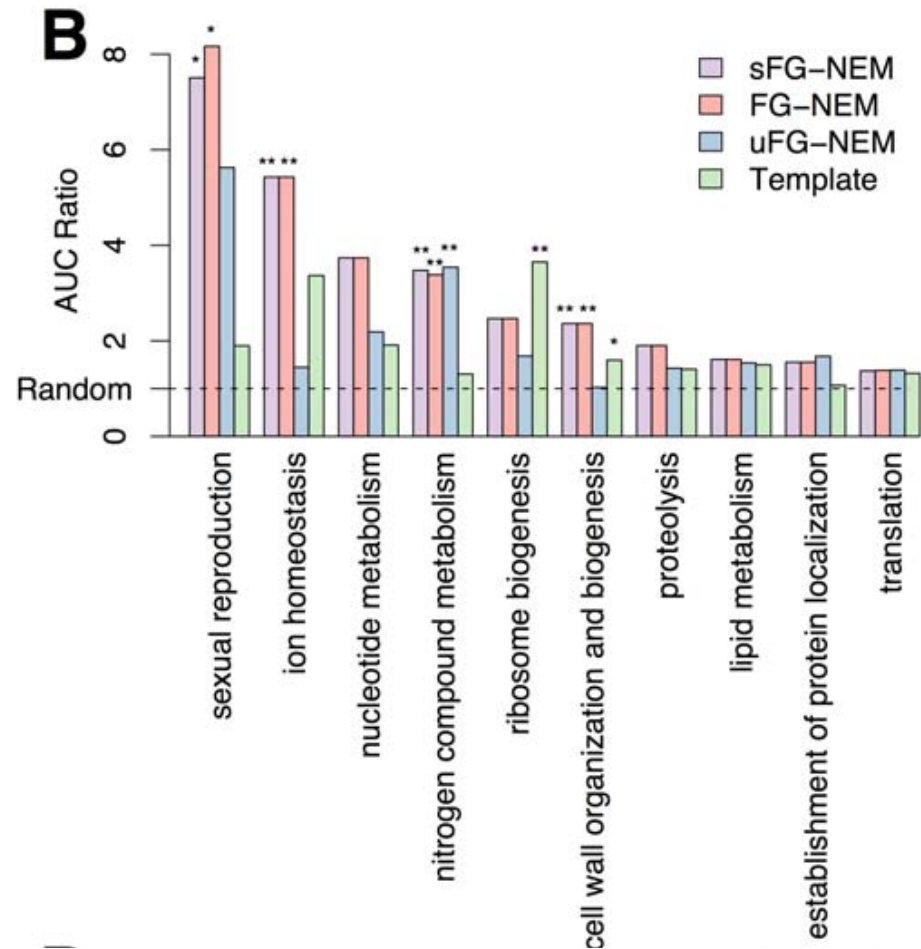
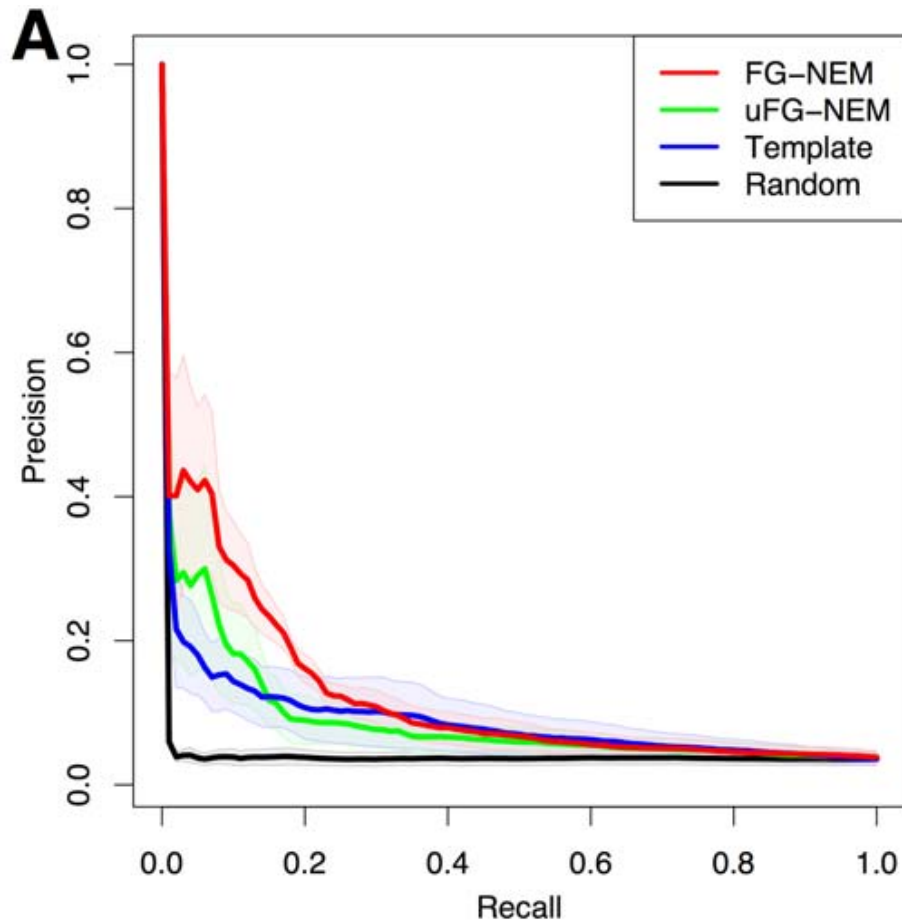
- Attach new E-genes to S-gene network
- An attached gene e to S-gene s asserts that e is directly downstream of s
- All E-genes attached to the S-gene network are called frontier genes
- An E-gene's connectivity is examined based on the Log-likelihood Attachment Ratio

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

One of the S genes

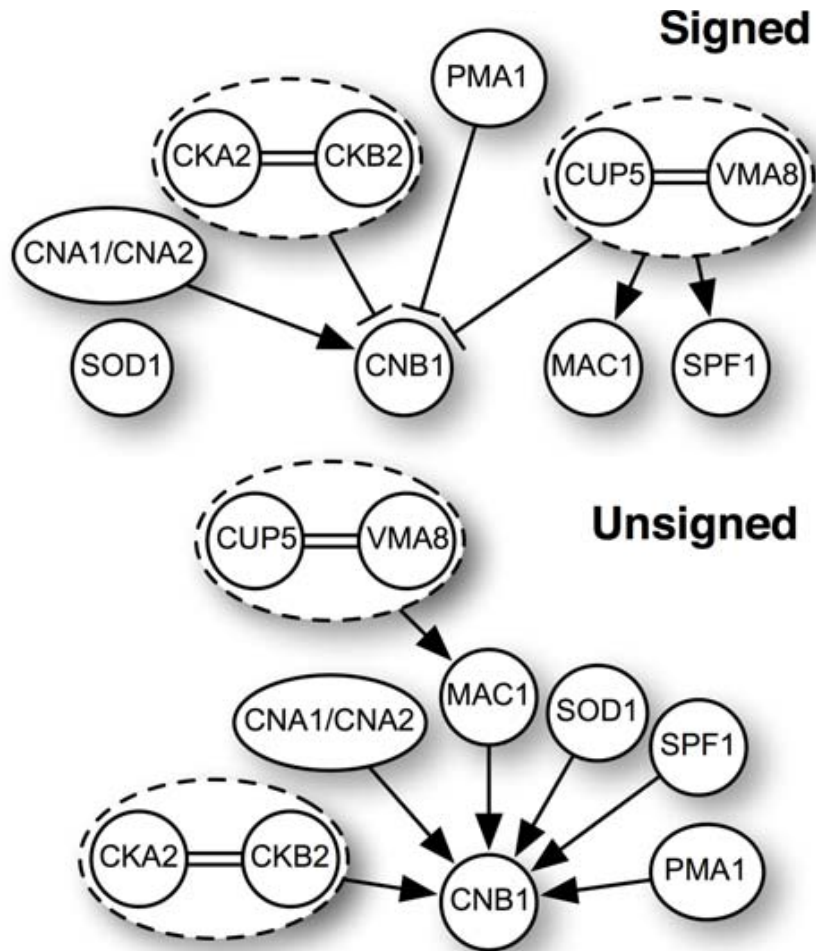


FG-NEM based pathway expansion in yeast



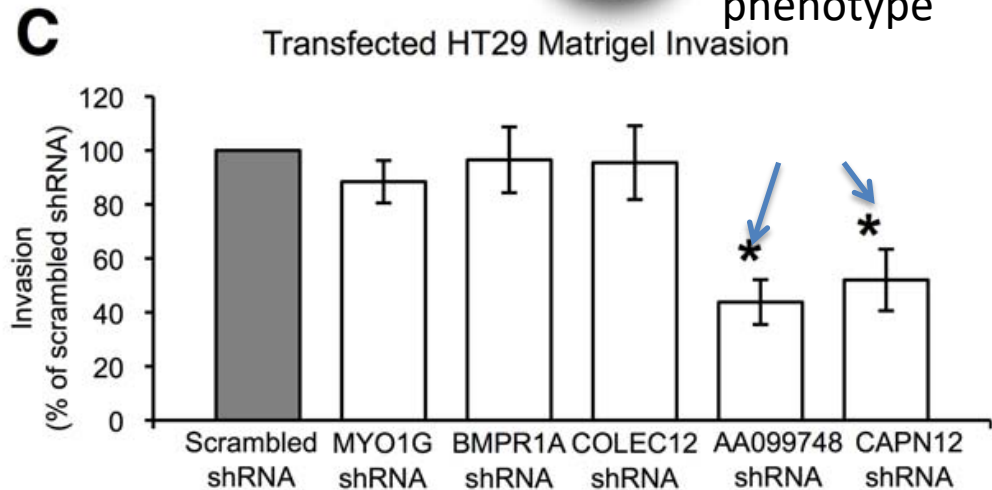
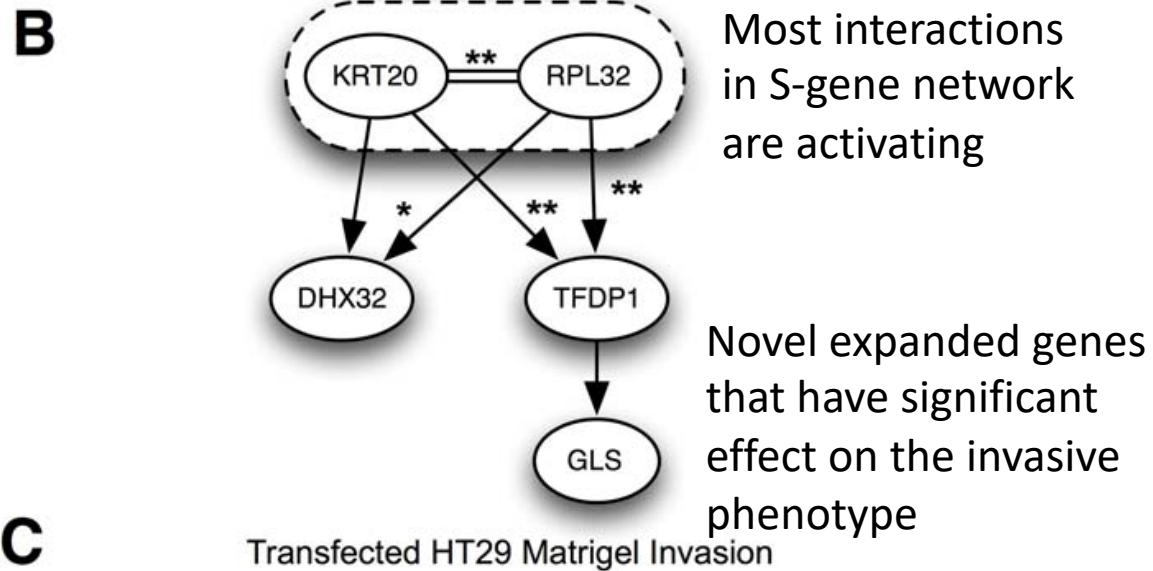
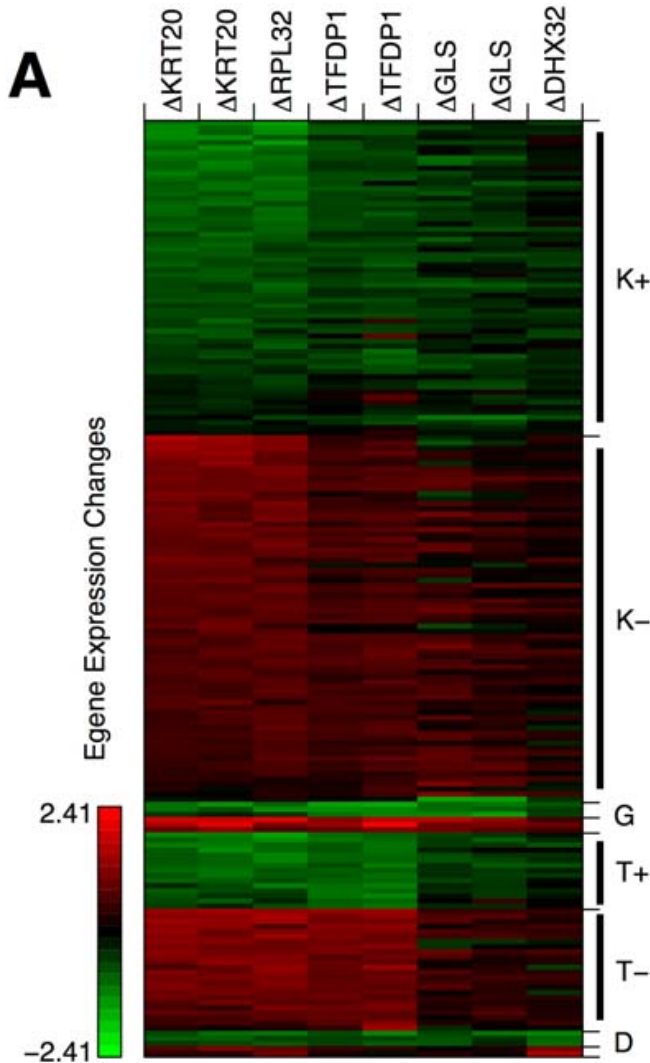
Template matching: rank E genes based on similarity in expression to an “idealized template”

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 homeostasis compared to uFG-NEM

FG-NEM application to colon cancer



Summary

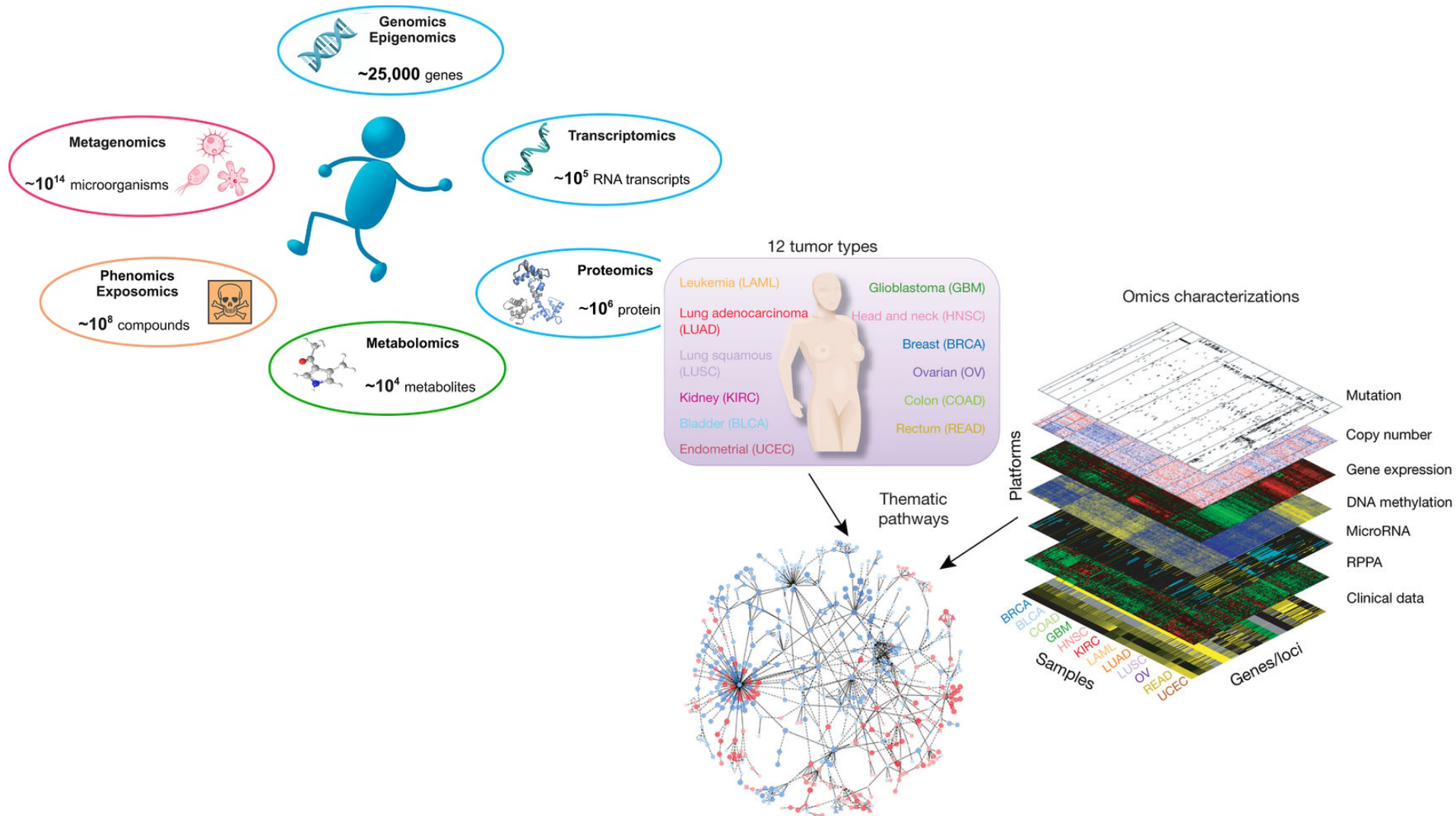
- FG-NEMs: A general approach to infer an ordering of genes from knock-down phenotypes
- Strengths
 - FG-NEMs could be used in an iterative computational-experimental framework
 - Handles signed interactions between S-genes
- Weaknesses
 - Computational complexity of the inference procedure might be 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 seen 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.

Data integration strategies

Biological data is of many different types



We are getting better at collecting lots of different types of biological datasets

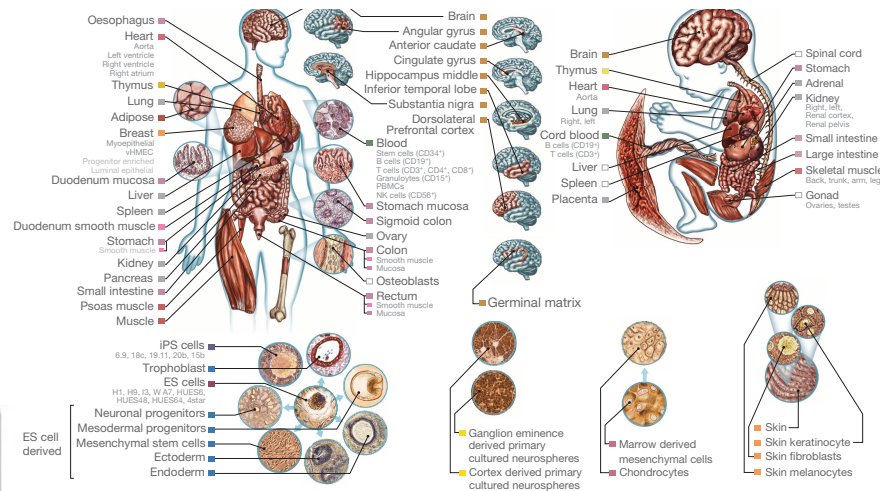
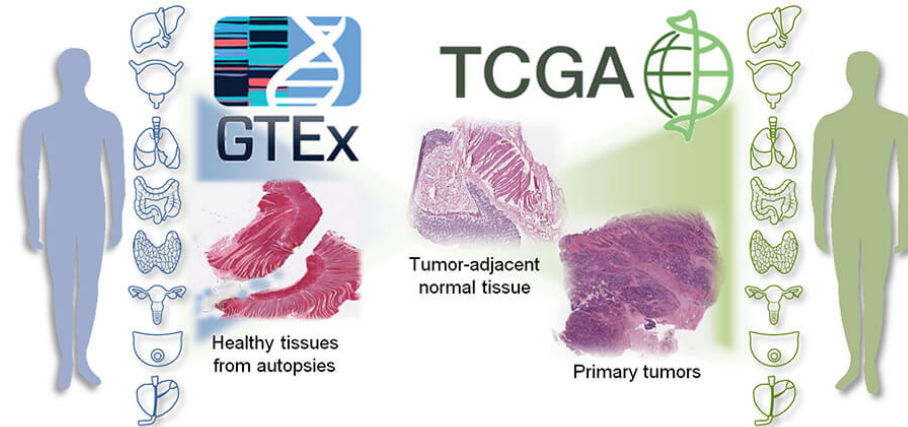
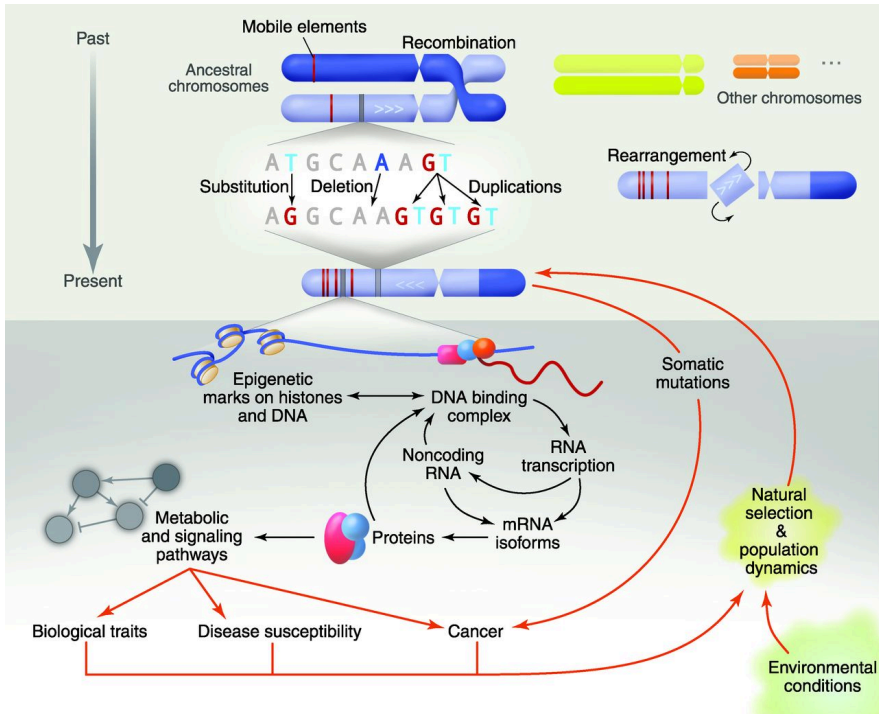


Image Credit: Dvir Aran, Ph.D., University of California, San Francisco

Need for systematic approaches for data integration

- The approach to integrate different data types depends upon the end goal and the types of data available
- Three considerations
 - Number of samples per data type
 - Supervised or unsupervised
 - Types of measurements
 - Gene sets versus quantitative profiles

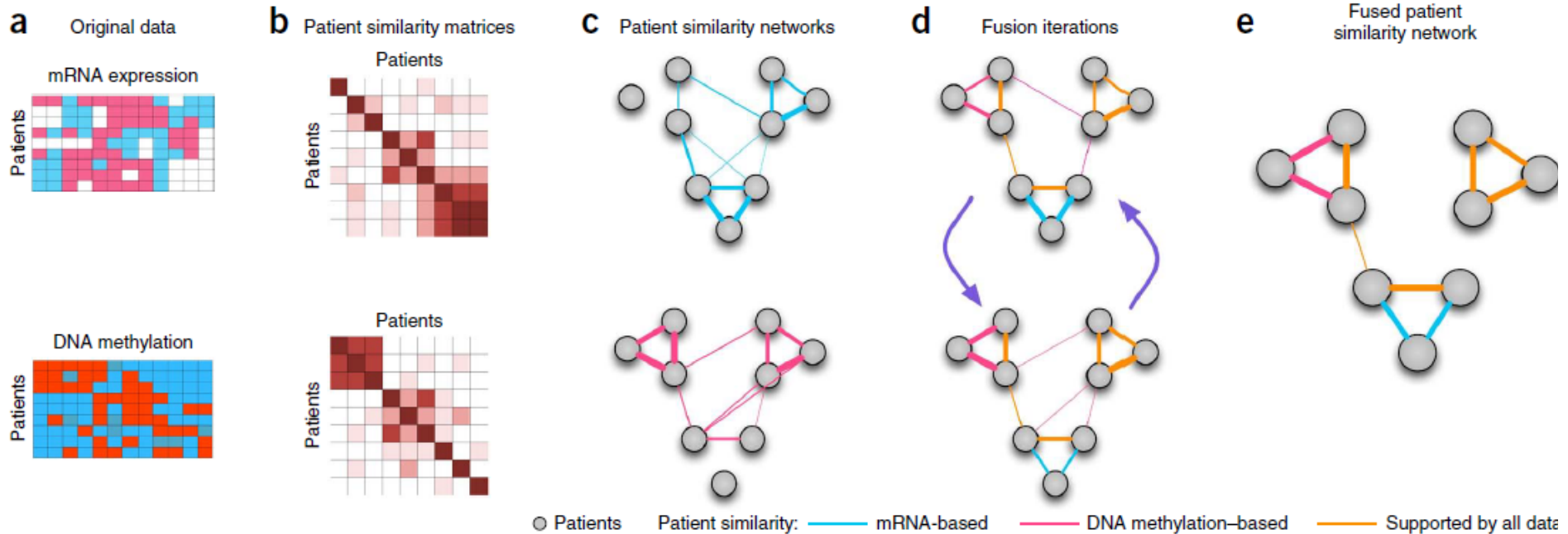
Network-based approaches for integrating data

- Network-inference based
 - Learning mixed graphical models where different variable types (different probability distribution families) represent different omic data types
- Diffusion based
 - Similarity Network Fusion (Wang et al., Nature Methods 2014)
 - MASHUP (Cho et al., Cell Systems 2016)
 - GeneMania (Mostafavi et al, Genome Biology 2008)
- Information flow based methods
 - Especially suited if we have a small number of samples
 - Max flow
 - Steiner tree

Similarity Network Fusion

- Given N different types of measurements for different individuals
- Do
 - Construct a similarity matrix of individuals for each data type
 - Integrate the networks using a single similarity matrix using an iterative algorithm
 - Cluster the network into a groups of individuals

Similarity network fusion with two data types

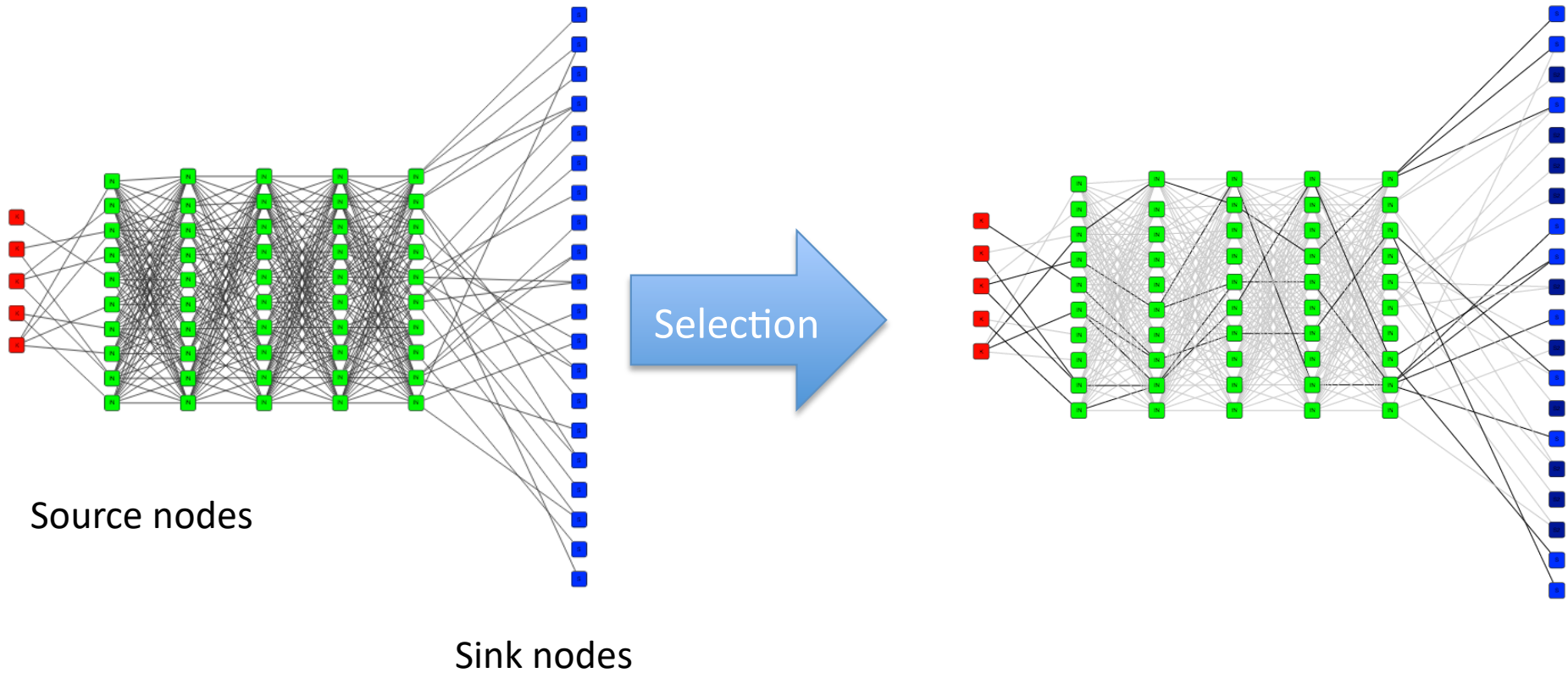


Similarity network fusion (Nodes are patients, edges represent similarities).

Problem definition of information flow

- Given
 - two node sets and a weighted directed network with edge weights corresponding to the flow between two nodes
- Do
 - Find the subnetwork that maximizes the flow between the two node sets

Information flow between sink to source nodes



Source nodes

Sink nodes

Skeleton Graph:

Potential gene network derived from a given confidence level

Selected Signaling Pathway:

Selected by different set of algorithms to predict the most significant and relevant paths

Information flow-based methods

- Used for integrating different types of data, as well as for examining perturbations and their effect
- Integration of different types of “omics” data
 - Min cost max flow (ResponseNet; Yeger-Lotem et al 2008)
 - Prize-collecting Steiner tree variants (Huang & Fraenkel 2009, OmicsIntegrator)

Notation

- A flow network is defined as directed graph $G=(V,E)$, with capacities for each edge
 - V : vertex set
 - E : edge set
- s : source node
- t : sink node
- $c(u,v)>0$: Capacity of edge (u,v)

Flow in a graph G

- A flow in G is defined by a function f that has the following properties for each edge (u,v) :

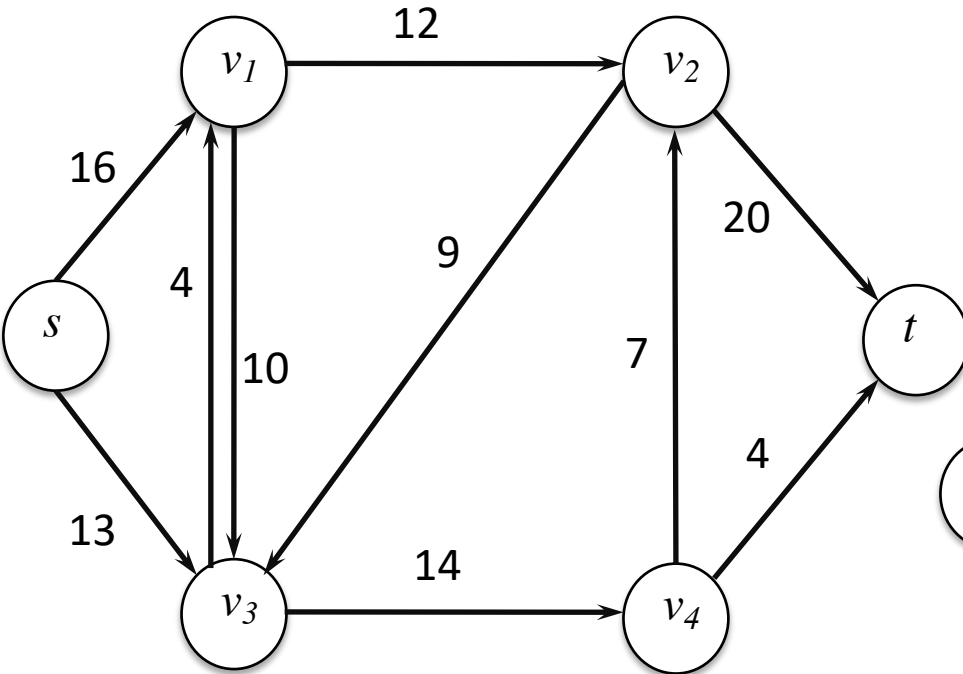
$$f(u, v) \leq c(u, v) \quad \text{Capacity constraint}$$

$$\sum_{v \in V, v \neq s, t} f(u, v) = 0 \quad \text{Conservation of flow}$$

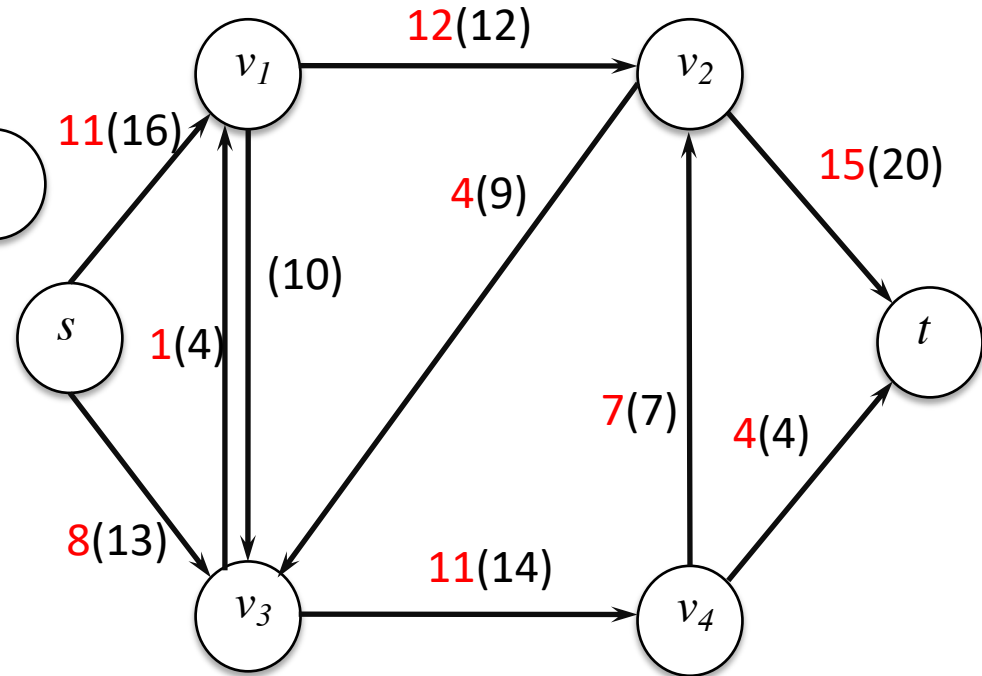
- The value of a *flow* is defined as

$$|f| = \sum_{v \in V} f(s, v)$$

An example flow network



Flow network G



A flow of 19 on G

Only positive flows are shown

Max-flow problem

- Given
 - A flow network G , source s and sink t
- Do
 - find a flow f with maximum value
- How
 - Ford-Fulkerson algorithm

Variation: Min cost max flow

- Often the question is not to maximize flow, but to find the most efficient/least expensive way of doing this
- In addition to the flow, there is also a cost associated with each edge
 - For example, the cost might be inversely proportional to the edge confidence
- So we would try to maximize the overall flow at the smallest cost

Min cost max flow

- Define cost of each edge as $a(u,v)$

- Overall cost:

$$\sum_{(u,v) \in E} a(u,v) f(u,v)$$

- Minimize cost while maximize flow as follows:

$$\sum_{(u,v) \in E} a(u,v) f(u,v) - \gamma \sum_{v \in V} f(s,v)$$

- This idea was used in ResponseNet tool
 - E. Yeger-Lotem, L. Riva, L. J. J. Su, A. D. Gitler, A. G. Cashikar, O. D. King, P. K. Auluck, M. L. Geddie, J. S. Valastyan, D. R. Karger, S. Lindquist, and E. Fraenkel, "Bridging high-throughput genetic and transcriptional data reveals cellular responses to alpha-synuclein toxicity." *Nature genetics*, vol. 41, no. 3, pp. 316-323, Mar. 2009.

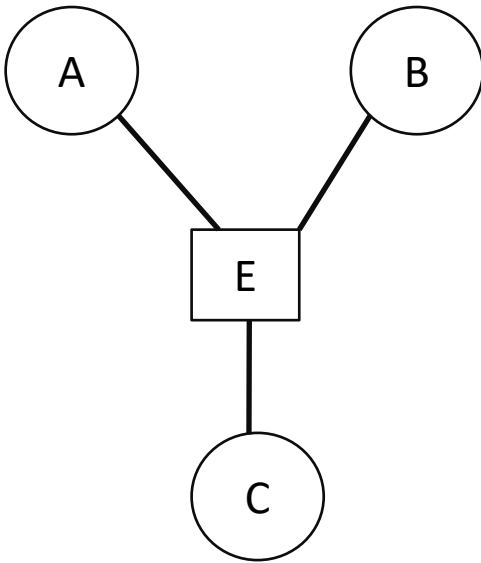
Alternate problem definition of information flow

- Given
 - A node set
 - A weighted network
- Do
 - Find the minimal graph connecting the nodes, where minimal is defined by the graph with the lowest total weight
- We will use a Steiner tree approach to address this problem

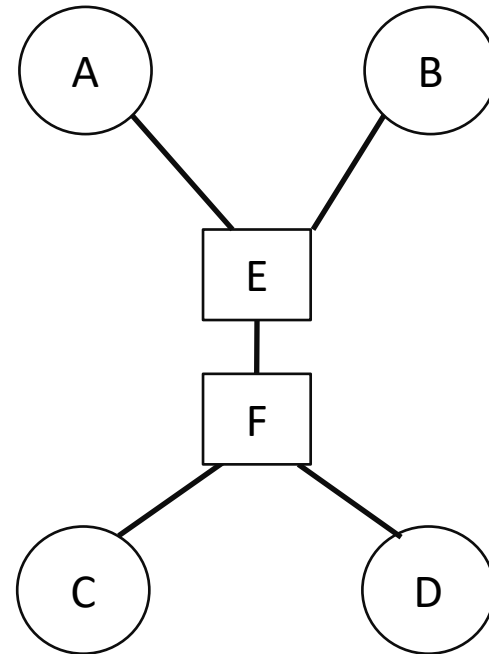
Steiner tree

- Let's start by defining a Steiner tree
- Given
 - edge-weighted graph $G=\{V, E, w\}$
 - A subset S of V
- A Steiner tree is a minimal length tree connecting S , including potentially intermediate nodes
- This problem is NP-complete

Steiner tree examples



$S = \{A, B, C\}$



$S = \{A, B, C, D\}$

Prize-collecting Steiner tree objective function

- $p(i)$: Define prize of node i as
- $y(i)$: include a node i
- $a(i,j)$: Define cost of edge (i,j)
- $x(i,j)$: include an edge
- Constrain that subnetwork must be a tree
- PCST objective

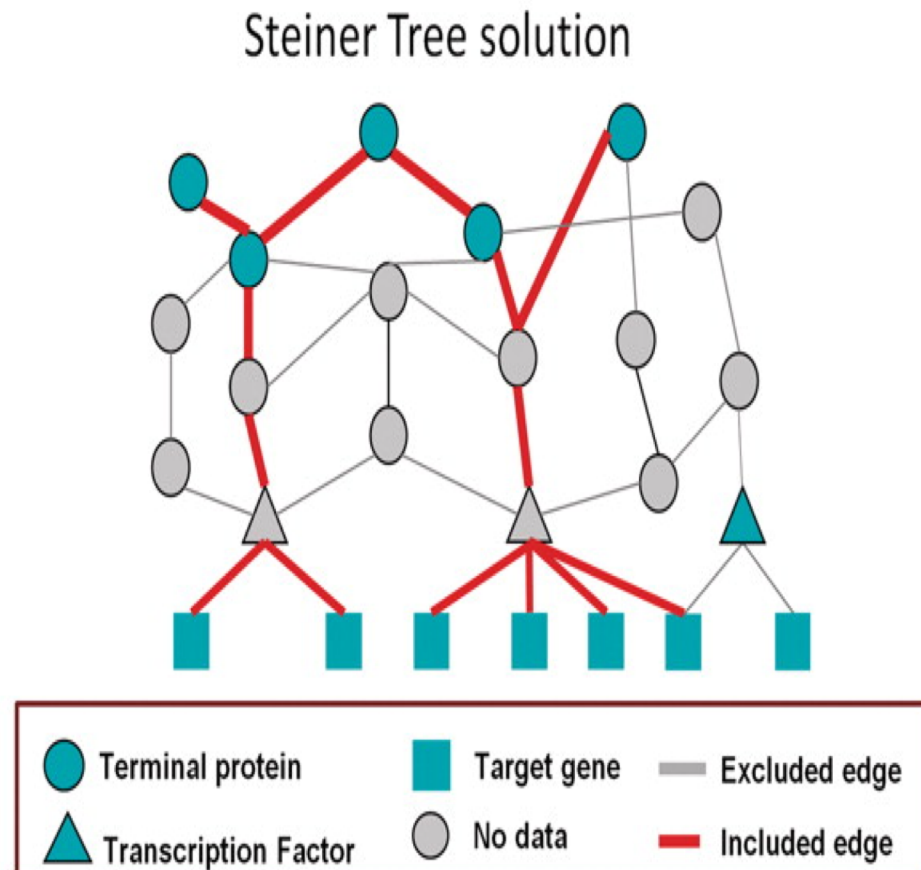
$$\max \sum_i p(i)y(i) - \lambda \sum_{(i,j)} a(i,j)x(i,j)$$

Trade-off between cost and prize

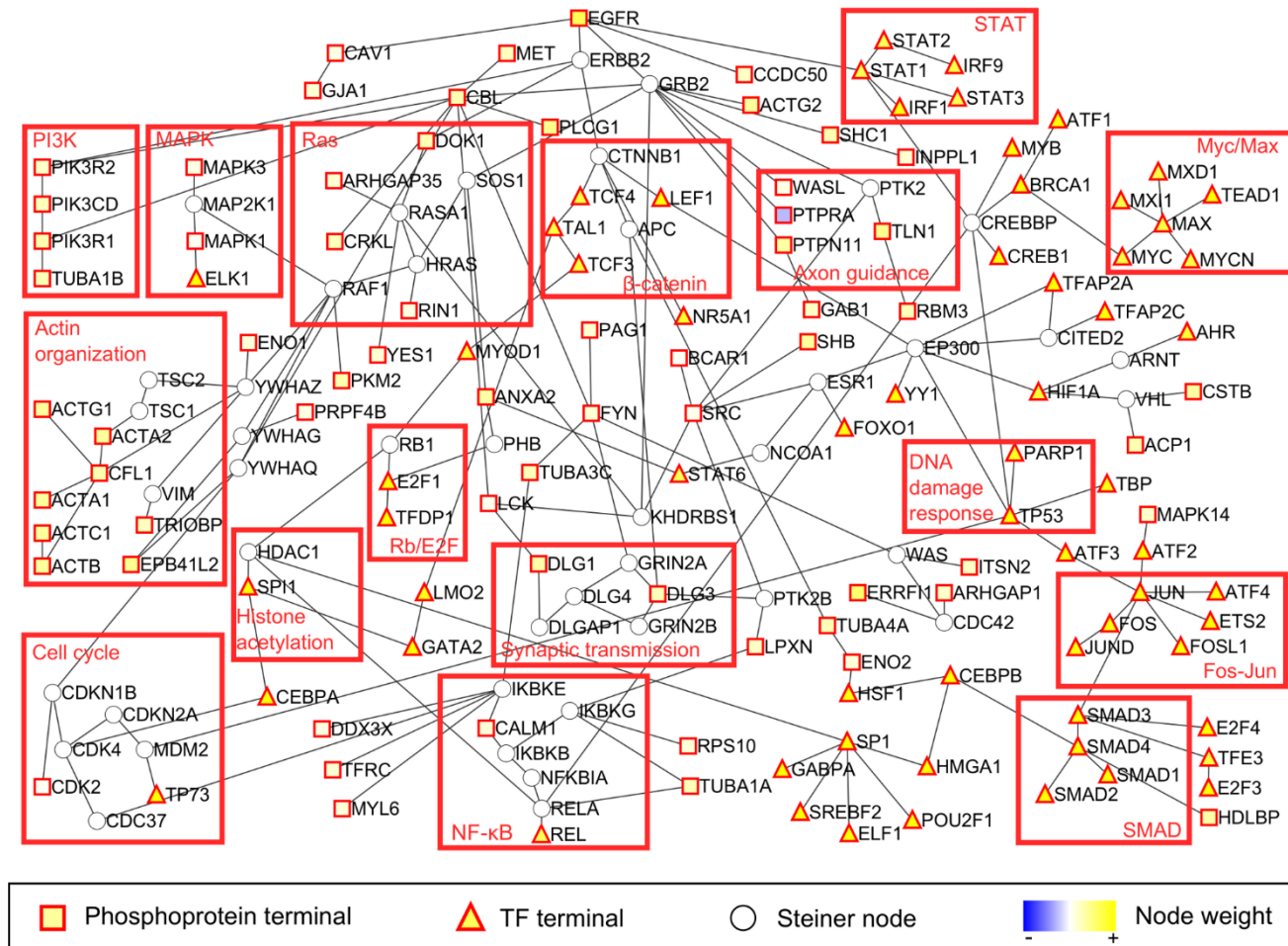
- Solve using variety of optimization techniques
 - E.g. integer linear programming-based method (Ljubic et al, 2006)

Prize-collecting Steiner trees (PCST) connect signaling proteins to gene regulation

- Top: functional screen hits, bottom: mRNA response
- Predicts relevant nodes, paths, transcription factors
- Cannot directly predict transcriptome effect from perturbations; edges are not oriented



PCST to phosphoproteomic and transcriptomic data to find genes relevant to glioblastoma multiforme

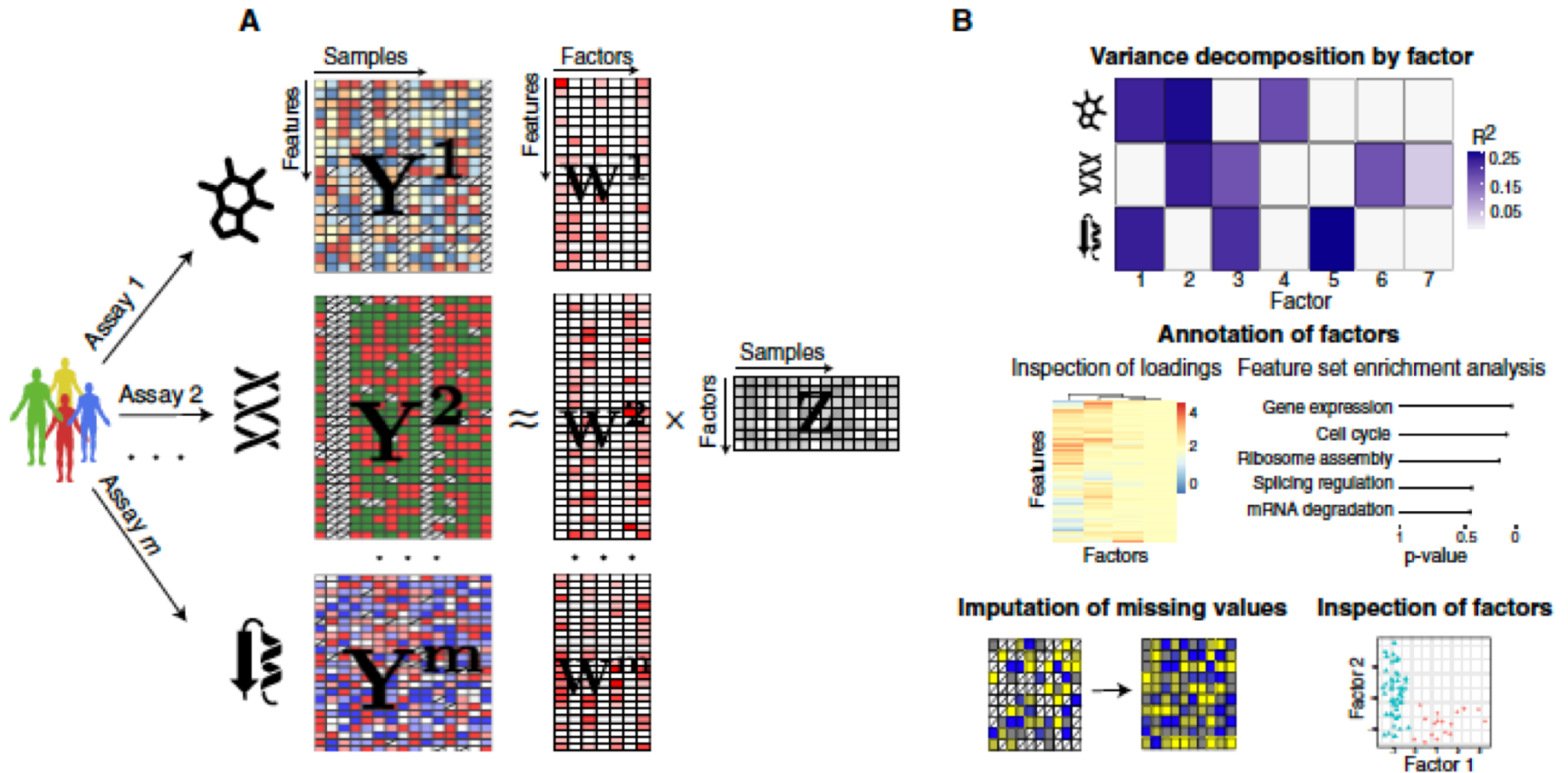


Huang *et al*, 2013, Figure 2

Types of approaches

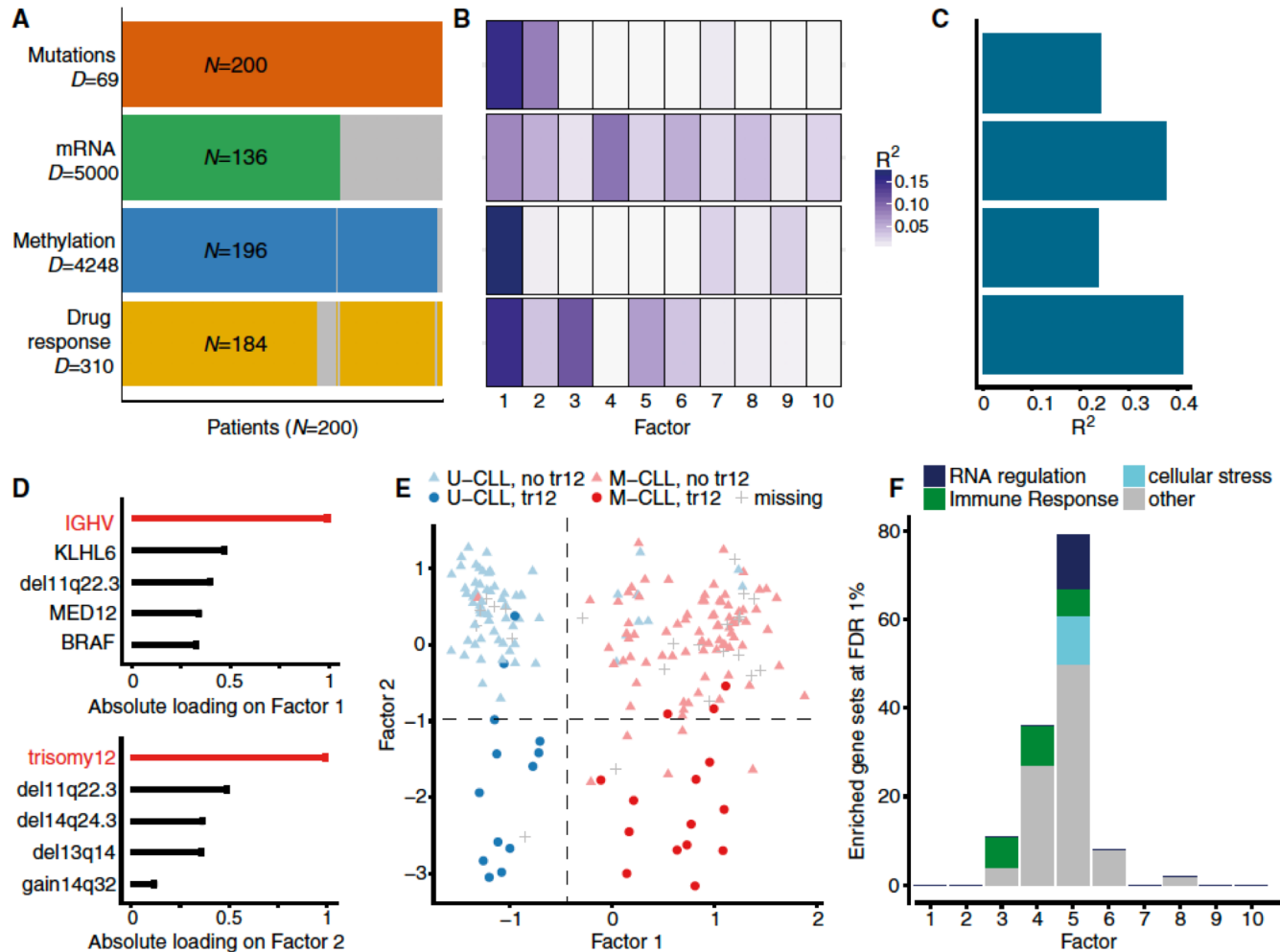
- Network-based approaches
 - Network inference
 - Similarity network fusion
 - Information flow based methods
- Matrix factorization based approaches
 - Also known as clustering/dimensionality reduction based approaches
 - Multi-omics factor analysis
 - Non-negative matrix tri-factorization

Multi-omics Factor Analysis (MOFA)

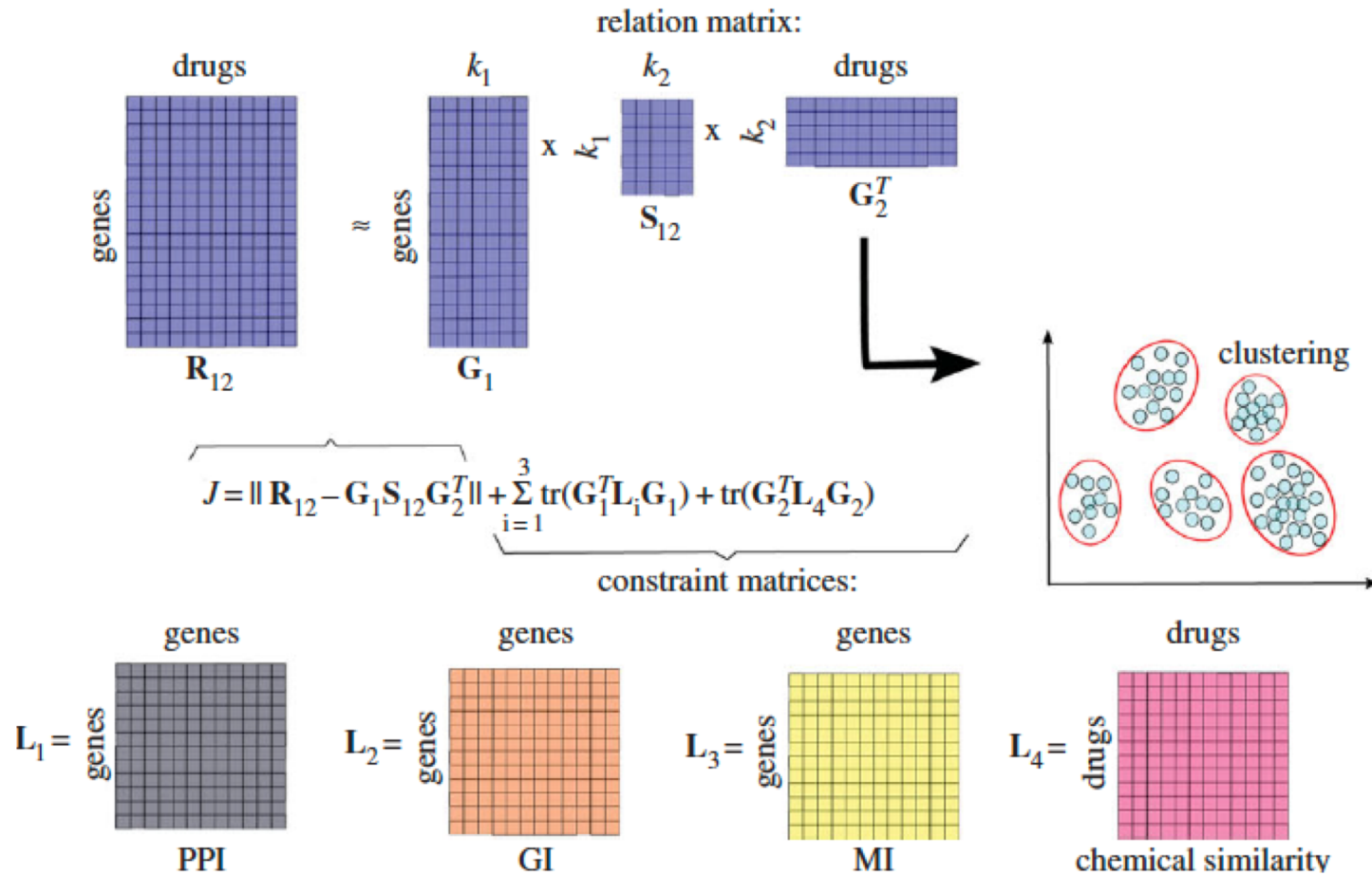


(1) Allows for missing partially overlapping datasets, (2) Based on a probabilistic model, (3) Learns sparse factors

Using MOFA for Chronic Lymphocytic Leukaemia



Non-negative Matrix Tri Factorization for predicting gene drug interactions



Web and software resources

- GeneMANIA (Network integration and diffusion-based subnetworks)
 - <http://www.genemania.org>
- HOTNET (Diffusion-based subnetworks)
 - <http://compbio.cs.brown.edu/projects/hotnet2/>
- ResponseNet (flow network)
 - <http://netbio.med.ad.bgu.ac.il/respnet/>
- OmicsIntegrator (PCST)
 - <http://fraenkel-nsf.csbi.mit.edu/omicsintegrator/>

Concluding remarks

- We have seen a suite of problems, algorithms and applications in a real setting
- These ranged from network inference, dynamic network inference, network modules, network alignment and network-based interpretation
- We saw less of
 - Integration of different types of networks
 - Experimental design for better learning of networks
- If you remain interested in these topics or would like to learn more, feel free to reach out to me.