

Identifying Signaling Pathways

BMI/CS 776

www.biostat.wisc.edu/bmi776/

Spring 2016

Anthony Gitter

gitter@biostat.wisc.edu

Goals for lecture

- Challenges of integrating high-throughput assays
- Connecting relevant genes/proteins with interaction networks
- ResponseNet algorithm
- Related signaling pathway prediction methods

High-throughput screening

- Which genes are involved in which cellular processes?
- Hit: gene that affects the phenotype
- Phenotypes include:
 - Growth rate
 - Cell death
 - Cell size
 - Intensity of some reporter
 - Many others

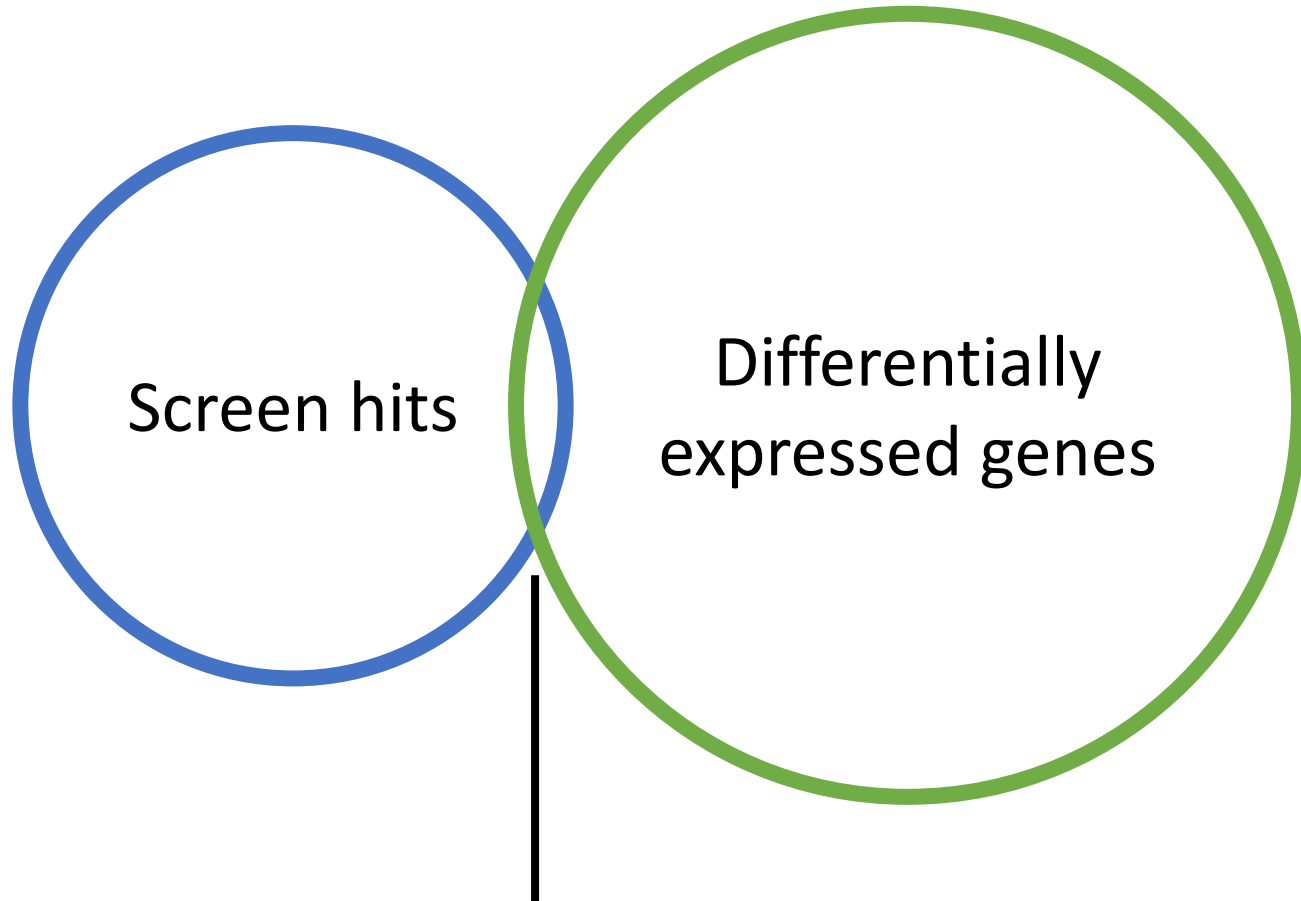
Types of screens

- Genetic screening
 - Test genes individually or in parallel
 - Knockout, knockdown (RNA interference), overexpression, CRISPR/Cas genome editing
- Chemical screening
 - Which genes are affected by a stimulus?

Differentially expressed genes

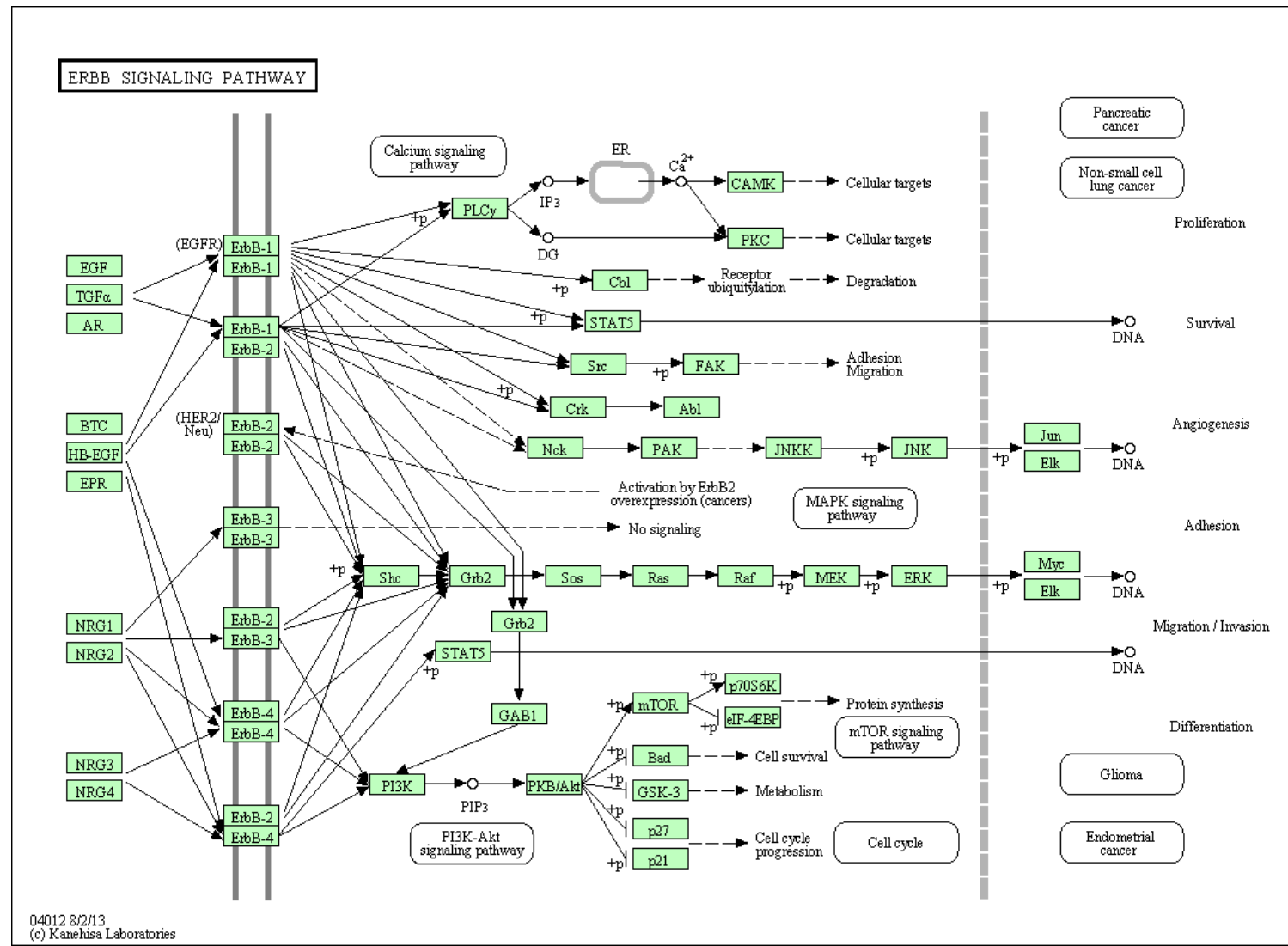
- Compare mRNA transcript levels between control and treatment conditions
- Genes whose expression changes significantly are also involved in the cellular process

Interpreting screens



Very few genes detected in both

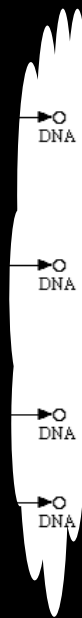
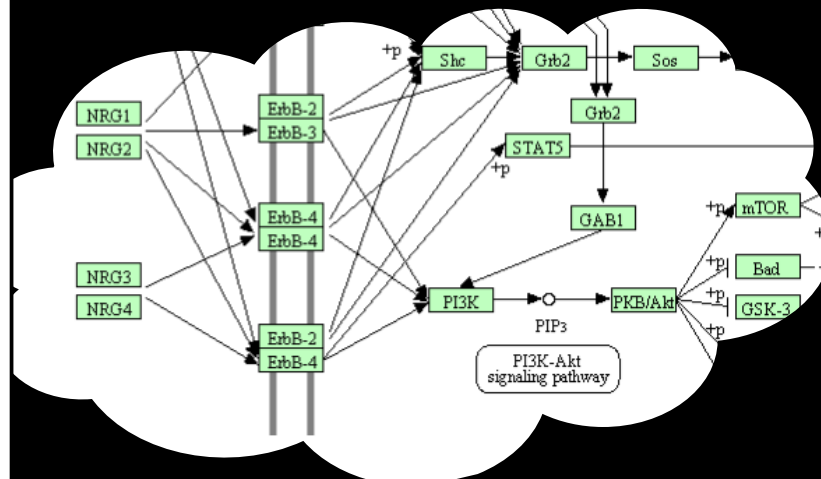
Assays reveal different parts of a cellular process



Assays reveal different parts of a cellular process

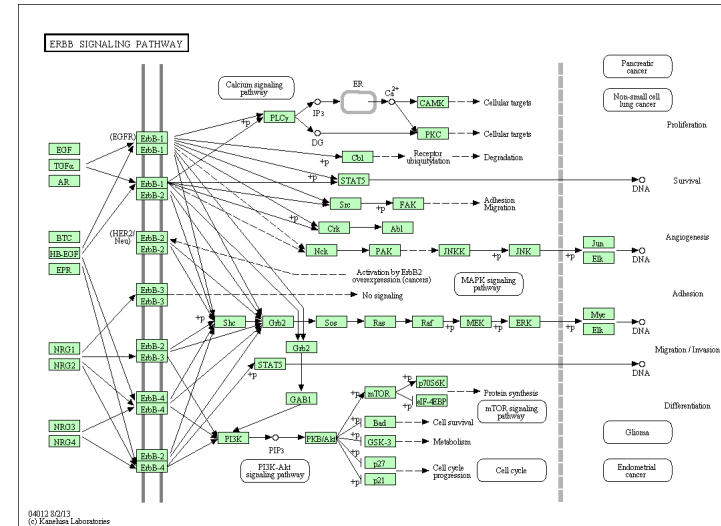
Differentially expressed genes

Genetic screen hits



Pathways connect the disjoint gene lists

- Can't rely on pathway databases
- High-quality, low coverage

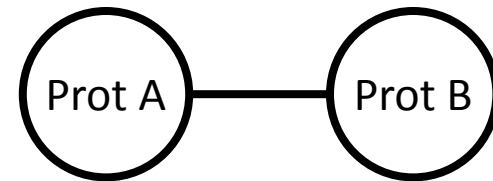
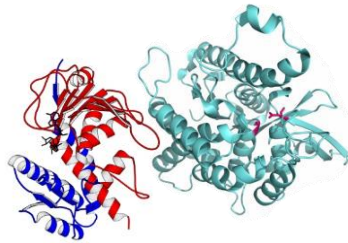


- Instead learn condition-specific pathways
- Combine data with generic physical interaction networks

Physical interactions

- Protein-protein

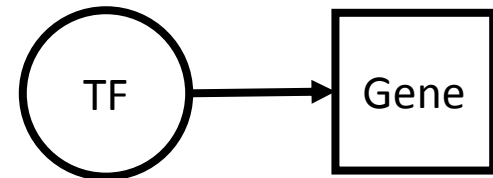
[Appling Graz](#)



- Metabolic

- Protein-DNA (transcription factor-gene)

[Yeger-Lotem2009](#)





- Genes and proteins are different node types


Weighting interactions

- Probability-like confidence of the interaction

Proteins

	MP2K1_HUMAN	Homo sapiens	<i>Temporarily not available for viewing in Netility.</i>
	MK01_HUMAN	Homo sapiens	<i>Temporarily not available for viewing in Netility.</i>

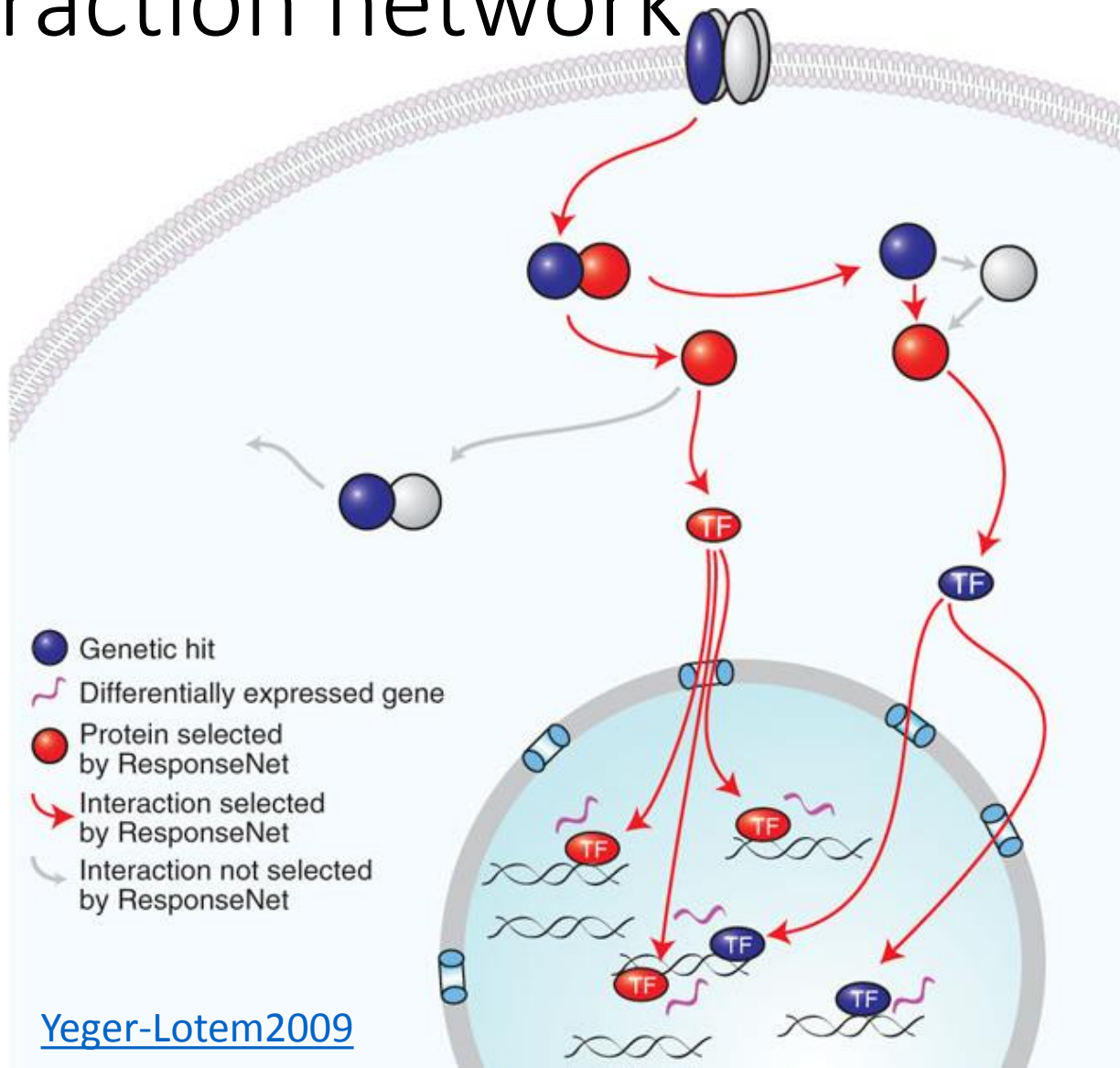
Evidence

Source DB ↕	Source ID ↕	Interaction Type ↕	PSI MI Code ↕	PubMed ID ↕	Detection Type ↕	PSI MI Code ↕
biogrid	857930	direct interaction	MI:0407	12788955	enzymatic study	MI:0415
ophid	17231	aggregation	MI:0191	11352917	confirmational text mining	MI:0024
ophid	17231	aggregation	MI:0191	15657099	deglycosylase assay	MI:1006
ophid	17234	aggregation	MI:0191	11352917	confirmational text mining	MI:0024
ophid	17234	aggregation	MI:0191	15657099	deglycosylase assay	MI:1006
biogrid	259225	direct interaction	MI:0407	12697810	t7 phage display	MI:0108
intact	EBI-8279991 	phosphorylation reaction	MI:0217	23241949	biosensor	MI:0968

- Example evidence: edge score of 1.0
- 16 distinct publications supporting the edge

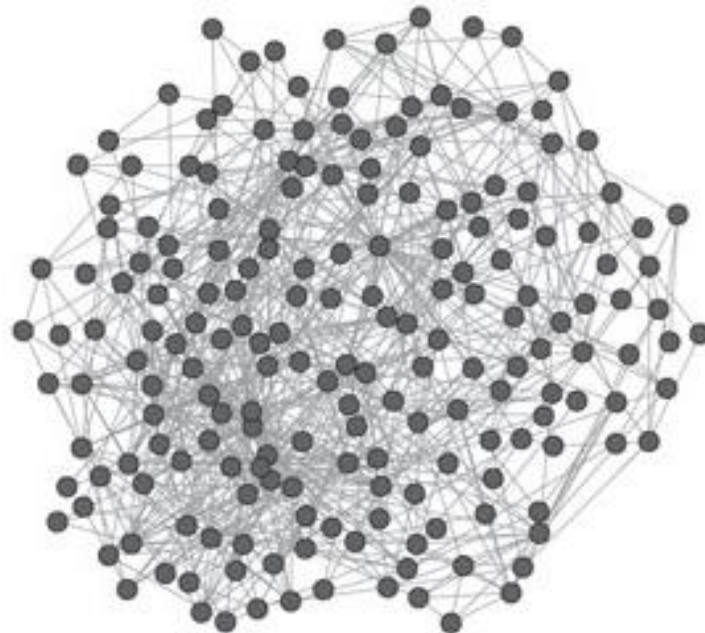
[iRefWeb](#)

Identify connections within an interaction network



Hairball networks

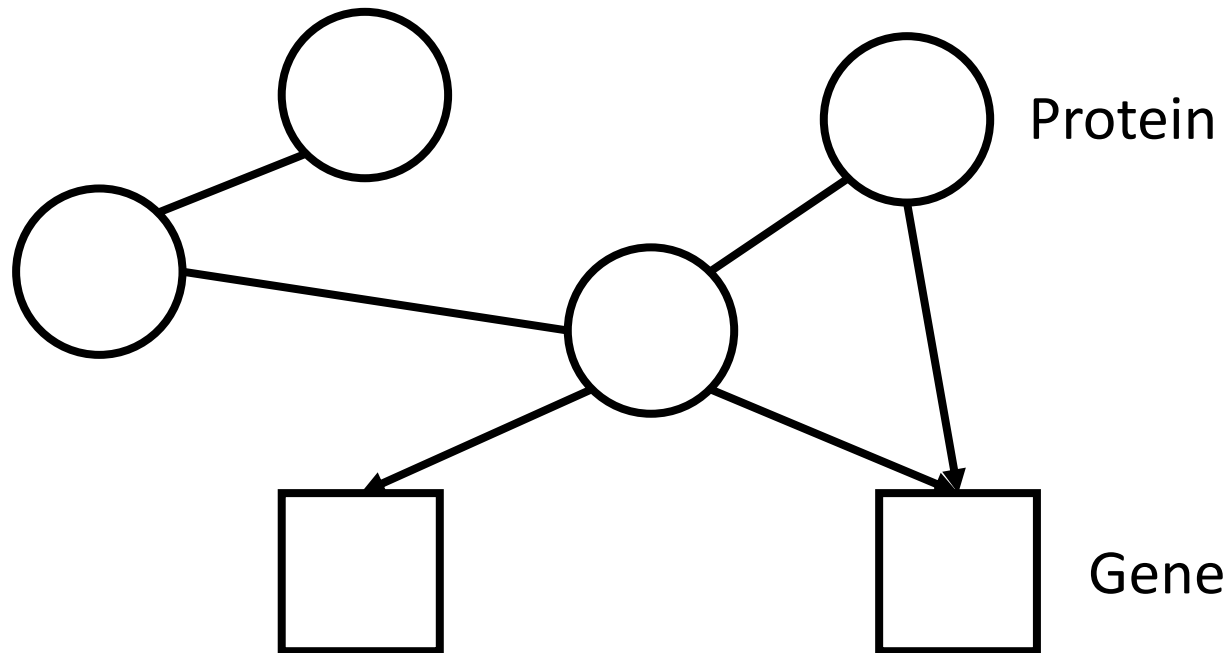
- Networks are highly connected
- Can't use naïve strategy to connect screen hits and differentially expressed genes



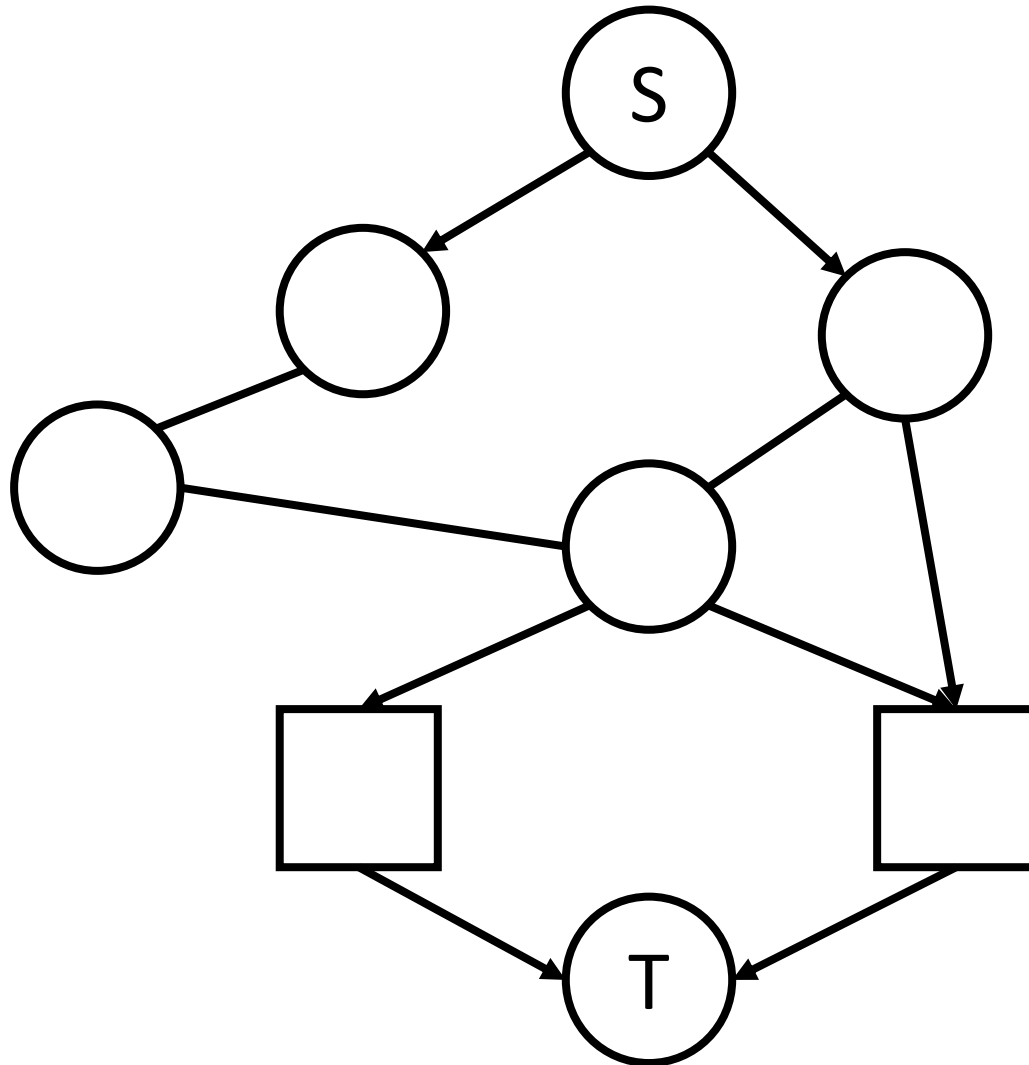
Framing an optimization problem

- ResponseNet network optimization goals
 - Connect screen hits and differentially expressed genes
 - Recover sparse connections
 - Identify intermediate proteins missed by the screens
 - Prefer high-confidence interactions

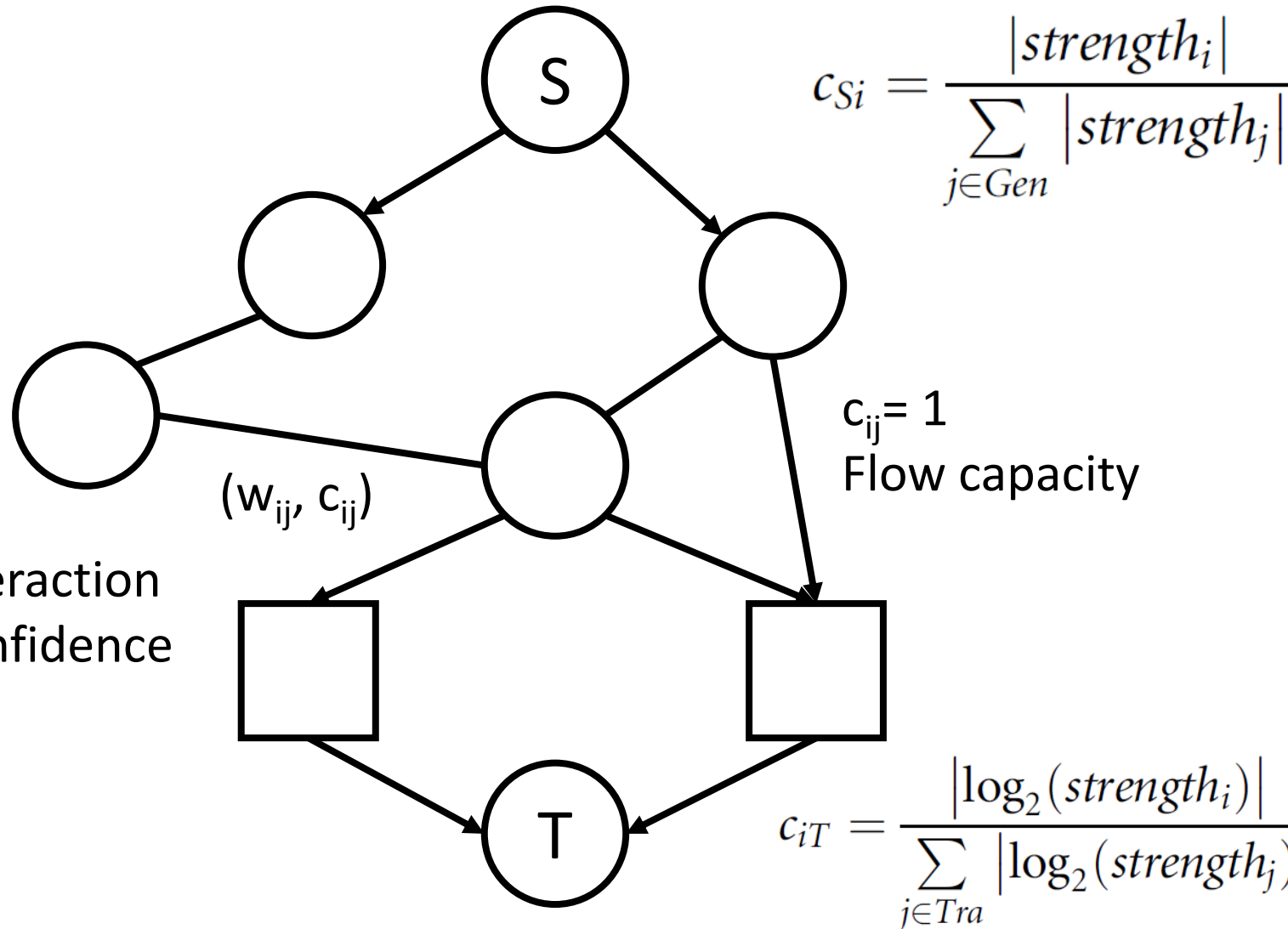
Construct the interaction network



Transform to a flow problem

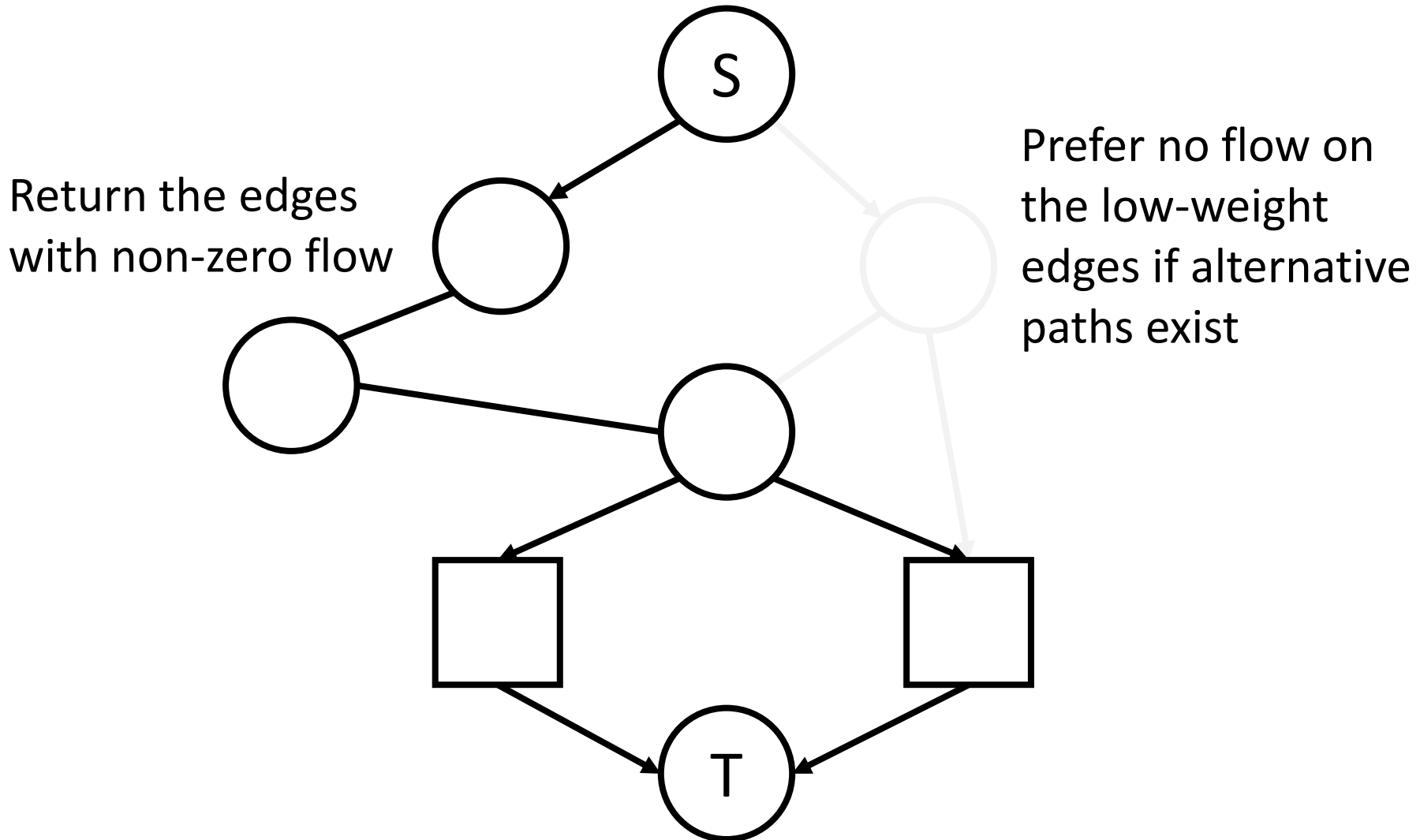


Weights and capacities on edges



w_{ij} from interaction network confidence

Find the minimum cost flow



Formal minimum cost flow

$$\min_f \left(\sum_{i \in V', j \in V'} -\log(w_{ij}) * f_{ij} - (\gamma * \sum_{i \in Gen} f_{Si}) \right)$$

Positive flow on an edge incurs a cost

Cost is greater for low-weight edges

Flow on an edge

Parameter controlling the amount of flow from the source

Formal minimum cost flow

$$\min_f \left(\sum_{i \in V', j \in V'} -\log(w_{ij}) * f_{ij} \right) - (\gamma * \sum_{i \in Gen} f_{Si})$$

Subject to:

$$\sum_{j \in V'} f_{ij} - \sum_{j \in V'} f_{ji} = 0 \quad \forall i \in V' - \{S, T\}$$

Flow coming in to a node
equals flow leaving the node

Formal minimum cost flow

$$\min_f \left(\sum_{i \in V', j \in V'} -\log(w_{ij}) * f_{ij} - (\gamma * \sum_{i \in Gen} f_{Si}) \right)$$

Subject to:

$$\sum_{j \in V'} f_{ij} - \sum_{j \in V'} f_{ji} = 0 \quad \forall i \in V' - \{S, T\}$$

$$\sum_{i \in Gen} f_{Si} - \sum_{i \in Tra} f_{iT} = 0$$

Flow leaving the
source equals flow
entering the target

Formal minimum cost flow

$$\min_f \left(\sum_{i \in V', j \in V'} -\log(w_{ij}) * f_{ij} - (\gamma * \sum_{i \in Gen} f_{Si}) \right)$$

Subject to:

$$\sum_{j \in V'} f_{ij} - \sum_{j \in V'} f_{ji} = 0 \quad \forall i \in V' - \{S, T\}$$

$$\sum_{i \in Gen} f_{Si} - \sum_{i \in Tra} f_{iT} = 0$$

Flow is non-negative
and does not exceed
edge capacity

$$0 \leq f_{ij} \leq c_{ij} \quad \forall (i, j) \in E'$$

Formal minimum cost flow

$$\min_f \left(\sum_{i \in V', j \in V'} -\log(w_{ij}) * f_{ij} \right) - \left(\gamma * \sum_{i \in Gen} f_{Si} \right)$$

Subject to:

$$\sum_{j \in V'} f_{ij} - \sum_{j \in V'} f_{ji} = 0 \quad \forall i \in V' - \{S, T\}$$

$$\sum_{i \in Gen} f_{Si} - \sum_{i \in Tra} f_{iT} = 0$$

$$0 \leq f_{ij} \leq c_{ij} \quad \forall (i, j) \in E'$$

Linear programming

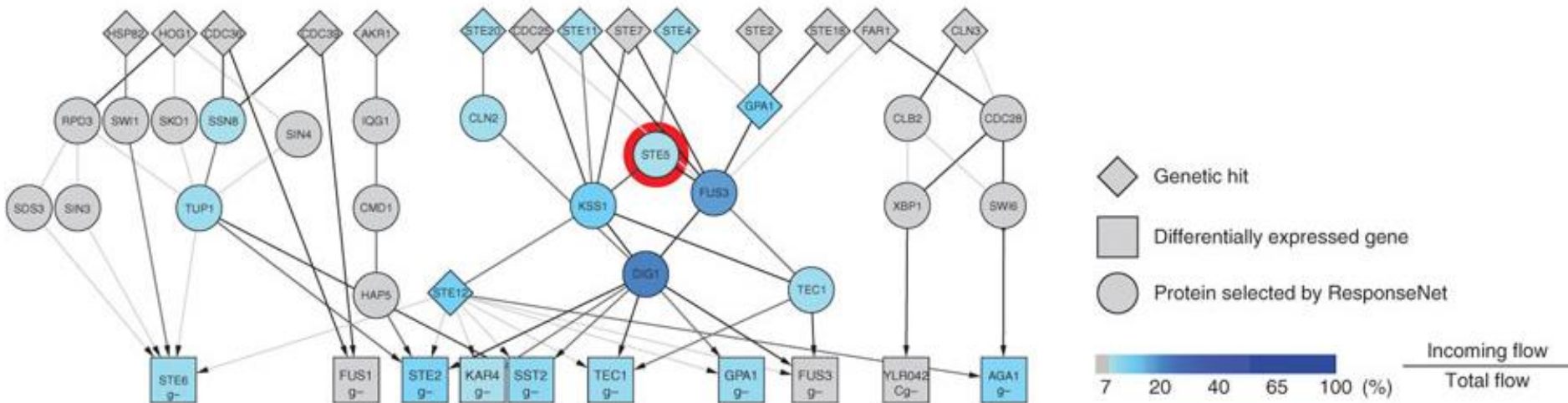
- Optimization problem is a linear program
- Canonical form

$$\begin{array}{ll} \text{maximize} & \mathbf{c}^T \mathbf{x} \\ \text{subject to} & \mathbf{Ax} \leq \mathbf{b} \\ \text{and} & \mathbf{x} \geq \mathbf{0} \end{array}$$

[Wikipedia](#)

- Polynomial time complexity
- Many off-the-shelf solvers
- [Practical Optimization: A Gentle Introduction](#)
 - [Introduction to linear programming](#)
 - [Simplex method](#)
 - [Network flow](#)

ResponseNet pathways



- Identifies pathway members that are neither hits nor differentially expressed
- Ste5 recovered when *STE5* deletion is the perturbation

ResponseNet summary

- Advantages

- Computationally efficient
- Integrates multiple types of data
- Incorporates interaction confidence
- Identifies biologically plausible networks

- Disadvantages

- Direction of flow is not biologically meaningful
- Path length not considered
- Requires sources and targets

Alternative pathway identification algorithms

- k-shortest paths
 - [Ruths2007](#)
 - [Shih2012](#)
- Random walks / network diffusion / circuits
 - [Tu2006](#)
 - eQTL electrical diagrams ([eQED](#))
 - [HotNet](#)
- Integer programs
 - Signaling-regulatory Pathway INference ([SPINE](#))
 - [Chasman2014](#)

Alternative pathway identification algorithms continued

- Path-based objectives
 - Physical Network Models ([PNM](#))
 - Maximum Edge Orientation ([MEO](#))
 - Signaling and Dynamic Regulatory Events Miner ([SDREM](#))
- Steiner tree
 - Prize-collecting Steiner forest ([PCSF](#))
 - Belief propagation approximation ([msgsteiner](#))
 - [Omics Integrator](#) implementation (unpublished)
- Hybrid approaches
 - [PathLinker](#): random walk + shortest paths
 - [ANAT](#): shortest paths + Steiner tree

Recent developments in pathway discovery

- Multi-task learning: jointly model several related biological conditions
 - ResponseNet extension: [SAMNet](#)
 - Steiner forest extension: [Multi-PCSF](#)
 - SDREM extension: [MT-SDREM](#)
- Temporal data
 - ResponseNet extension: [TimeXNet](#)
 - [Steiner forest extension](#)
 - [Temporal Pathway Synthesizer](#) (unpublished)

Condition-specific genes/proteins used as input

- Genetic screen hits (as causes or effects)
- Differentially expressed genes
- Transcription factors inferred from gene expression
- Proteomic changes (protein abundance or phosphorylation)
- Genetic variants or DNA mutations
- Enzymes regulating metabolites
- Receptors or sensory proteins
- Protein interaction partners
- Pathway databases or other prior knowledge

If you're still interested

- Computational Network Biology
 - Fall 2016 special topics course
 - BMI 826/CS 838
 - Professor Sushmita Roy