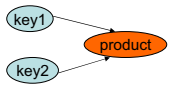
Biological networks

Recap

- Gene regulatory networks
  - Transcription Factors: special proteins that function as “keys” to the “switches” that determine whether a protein is to be produced
  - Gene regulatory networks try to show this “key-product” relationship and understand the regulatory mechanisms that govern the cell.

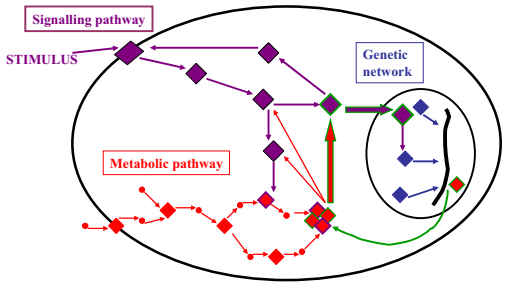


  - We went over a simple algorithm for detecting significant patterns in these networks

Other networks?

- Apart from regulation there are other events in a cell that require interaction of biological molecules
- Other types of molecular interactions that can be observed in a cell
  - enzyme – ligand
    - **enzyme:** a protein that catalyzes, or speeds up, a chemical reaction
    - **ligand:** extracellular substance that binds to receptors
    - metabolic pathways
  - protein – protein
    - cell signaling pathways
    - proteins interact physically and form large complexes for cell processes

Pathways are inter-linked



Interactions → Pathways → Network

- A collection of interactions defines a network
- Pathways are a subset of networks
  - All pathways are networks of interactions, however not all networks are pathways!
  - Difference in the level of annotation or understanding
- We can define a pathway as a biological network that relates to a **known** physiological process or complete function

Pathways

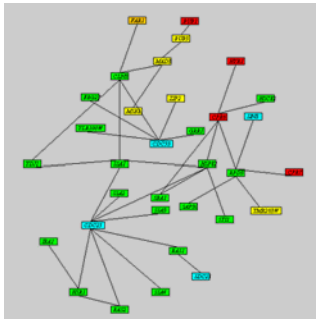
- However, there is no *precise* biological definition of a pathway
- Current automated partitioning of networks into pathways is somewhat arbitrary
  - We choose the start/finish points based on “important” or functionally known proteins
  - Gives us the ability to conceptualize the mapping of genotype → phenotype
    - understand functions of genes/proteins in a systematic way: systems biology

Pathway Databases

- KEGG
- BioCyc
- Reactome
- GenMAPP
- BioCarta
- TransPATH
- …175 more at Pathway Resource List http://www.cbio.mskcc.org/prl/index.php

The "interactome"

- The complete wiring of a proteome.
- Each vertex represents a protein.
- Each edge represents an "interaction" between two proteins.



An edge between two proteins if...

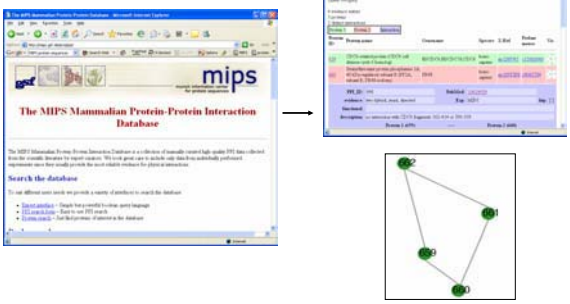
- The proteins interact physically and form large complexes
- The proteins are enzymes that catalyze two successive chemical reactions in a pathway
- One of the proteins regulates the expression of the other

Sources for interaction data

- Literature: research labs have been conducting small-scale experiments for many years!
- Interaction dabases:
  - MIPS (Munich Information center for Protein Sequences)
  - BIND (Biomolecular Network Interaction Database)
  - GRID (General Repository for Interaction Datasets)
  - DIP (Database of Interacting Proteins)
- Experiments:
  - Y2H (yeast two-hybrid method)
  - APMS (affinity purification coupled with mass spectrometry)

MIPS

- http://mips.gsf.de/



GRID

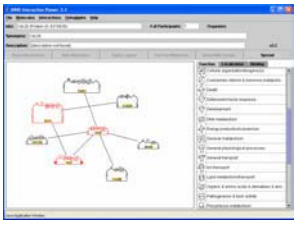
Mount Sinai Hospital, Toronto, Canada

- search GRID

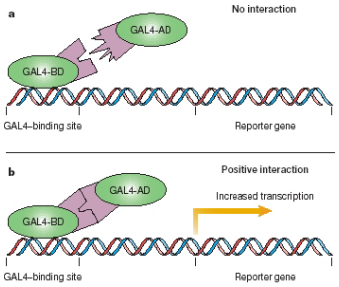
BIND

University of Toronto, Canada

- search BIND



Yeast two-hybrid method



S.-K. Ng and S.-H. Tan, "Discovering Protein-Protein Interactions", J. *of Bioin. and Comp. Bio.*, 1(4):711-741, 2004.

- These methods provide the ability to perform genome/proteome-scale experiments.
  - For yeast: 50,000 unique interactions involving 75% of known open reading frames (ORFs) of yeast genome
  - However, for *C. elegans* they provide relatively small coverage of the genome with ~5600 interactions.
- Problems with high-throughput experiments:
  - Low quality, false positives, false negatives
  - Fraction of biologically relevant interactions: 30%-50% (Deane *et al.* 2002)

Solution:

- User other indirect data sources to create a probabilistic protein network.
- Other sources include:
  - Genome data:
    - Existence of genes in multiple organisms
    - Locations of the genes
  - Bio-image data
  - Gene Ontology annotations
  - Microarray experiments
  - Sub-cellular localization data

Quality of interaction sources

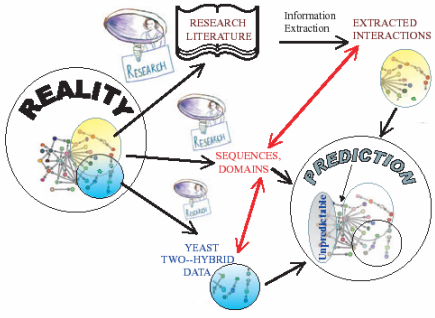
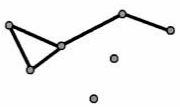


Image from Iossifov *et al.* (Bioinformatics 20(8), 2004)

Integration of sources into a single PPI network

- Binary approach
  - If there is at least one supporting evidence that there is interaction, assume the proteins do interact.
  - The “interaction” exists or do not exist. We don’t care about the quality of the source



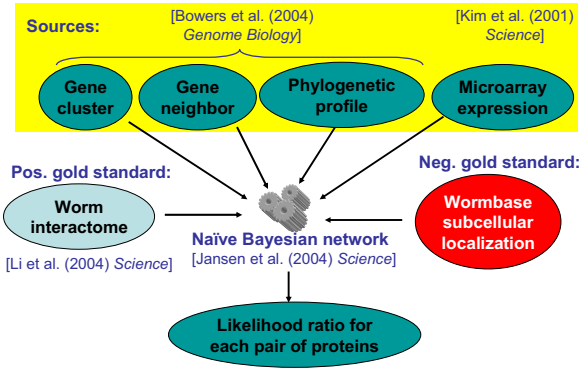
Probabilistic network approach

- Each “interaction” link between two proteins has a posterior probability of existence, based on the quality of supporting evidence.

Bayesian Network approach

- Jansen *et al.* (2003) *Science*. Lee *et al.* (2004) *Science*.
- Combine individual probabilities of likelihood computed for each data source into a single likelihood (or probability)
- Naive Bayes:
  - Assume independence of data sources
  - Combine likelihoods using simple multiplication

Combining data sources



Bayesian Approach

- A scalar score for a pair of genes is computed separately for each information source.
- Using gold positives (known interacting pairs) and gold negatives (known non-interacting pairs) interaction likelihoods for each information source is computed.
- The product of likelihoods can be used to combine multiple information sources
  - Assumption: A score from a source is independent from a score from another source.

Computing the likelihoods

- Partition the pair scores of an information source into bins and provide likelihoods for score-ranges
- E.g. Using the microarray information source and using Pearson correlation for scoring protein pairs you may get scores between -1 and 1. You want to know what is the likelihood of interaction for a protein pair that gets a Pearson correlation of 0.6.

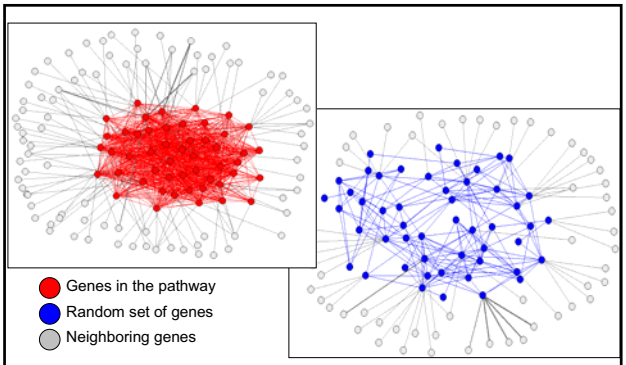
Partitioning the scores

pearson corr.	likelihood
(0.8,1.0]	
(0.6,0.8]	
(0.4,0.6]	
(0.2,0.4]	
(0.0,0.2]	
(-0.2,0.0]	
(-0.4,-0.2]	
(-0.6,-0.4]	
(-0.8,-0.6]	
[-1.0,-0.8]	

Computing the likelihood

- $$L = \frac{P(\text{Interaction} \mid \text{Score}) / P(\text{Interaction})}{P(\sim\text{Interaction} \mid \text{Score}) / P(\sim\text{Interaction})}$$

- Example



The 50 genes in the pathway are strongly clustered (with ~1900 functional linkages), however a random set of 50 genes contains fewer linkages (~300 functional linkages).