SUNY Korea CSE549
Spring 2017
Instructor: Sael lee

Biological Networks

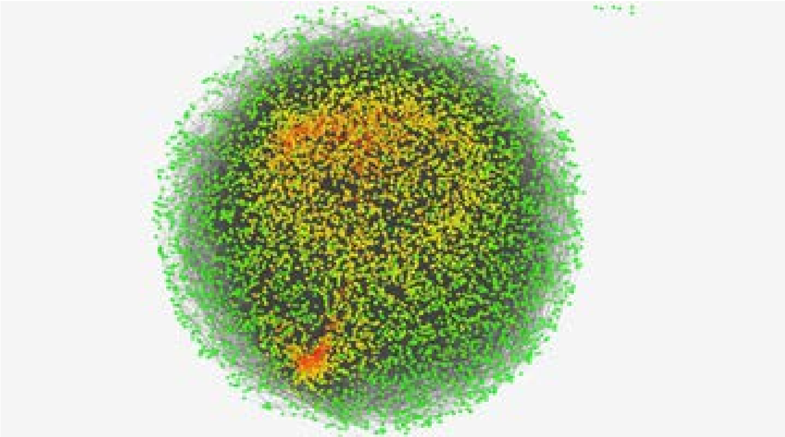
Ref: M. Zitnik and J. Leskovec’s CS2224W slides on bio-network.

Types of Biological Networks

- ❑ There are several types of bio-networks.
- ❑ Classification:
  - ❑ Gene co-expression networks
  - ❑ Protein-protein interaction networks
  - ❑ Signal transduction and gene regulatory networks (pathways)
  - ❑ Metabolic networks (pathways)
  - ❑ Other types of networks
    - ❑ Phylogenetic trees
    - ❑ Mixture of networks

Gene Co-expression Network

- Description:
  - **Gene co-expression** is process where set of genes are expressed in coordination to produce proteins.
  - **Gene co-expression networks** contains information on the **correlation of the gene expression** in different biological or environmental conditions.

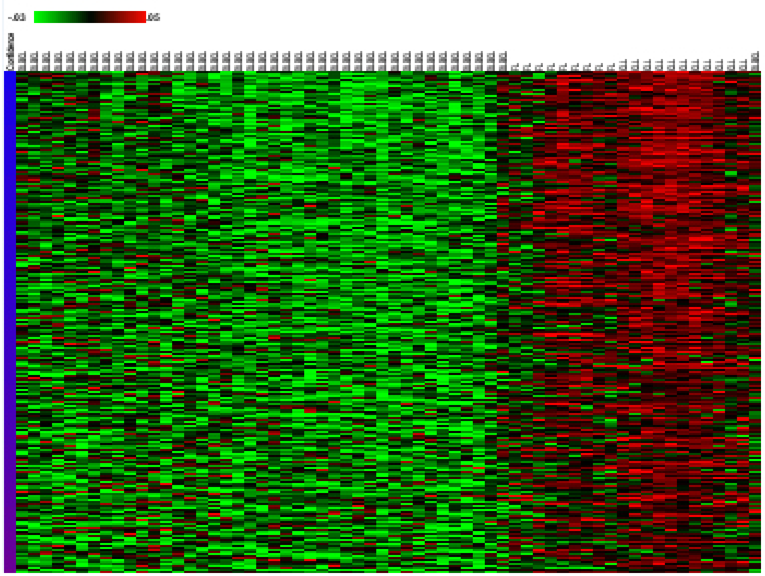


“A gene co-expression network constructed from a microarray dataset containing gene expression profiles of 7221 genes for 18 gastric cancer patients - S. Mohammad H. Oloomi ”

Gene Co-expression Network cont.

- ❑ Construction:
  - ❑ form edges between pairs of genes that show similar expressio n patterns across biological conditions,
  - ❑ where the activation levels of two co-expressed genes rise and fall together across conditions.

- ❑ Major DBs:
  - ❑ The Cancer Genome Atlas
  - ❑ NCBI Gene Expression Omnibus
  - ❑ GeneMANIA
  - ❑ EBI Array Express
  - ❑ GTEx Data Portal
  - ❑ MGI-Mouse Gene
  - ❑ Expression Database
  - ❑ **STRING (PPI)**
  - ❑ Bgee.

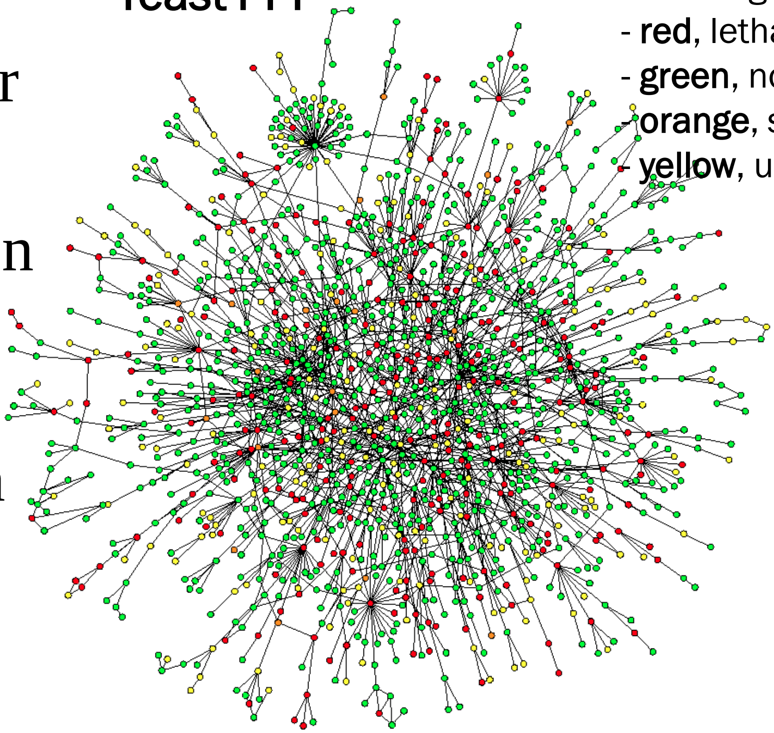


Protein-Protein Interaction Networks (PPI)

- ❑ Description: Networks where nodes represent proteins and edges represent interactions between the two protein.

- ❑ Types of Interactions:
  - ❑ to build a protein complex or to activate/deactivate.
  - ❑ However, types of interaction in PPI, i.e. “activation”, “binding to”, or “phosphorylation”, are often unknown

Yeast PPI



Color signifies the phenotypic effect of removing a protein
- **red**, lethal
- **green**, non-lethal
- **orange**, slow growth
- **yellow**, unknown

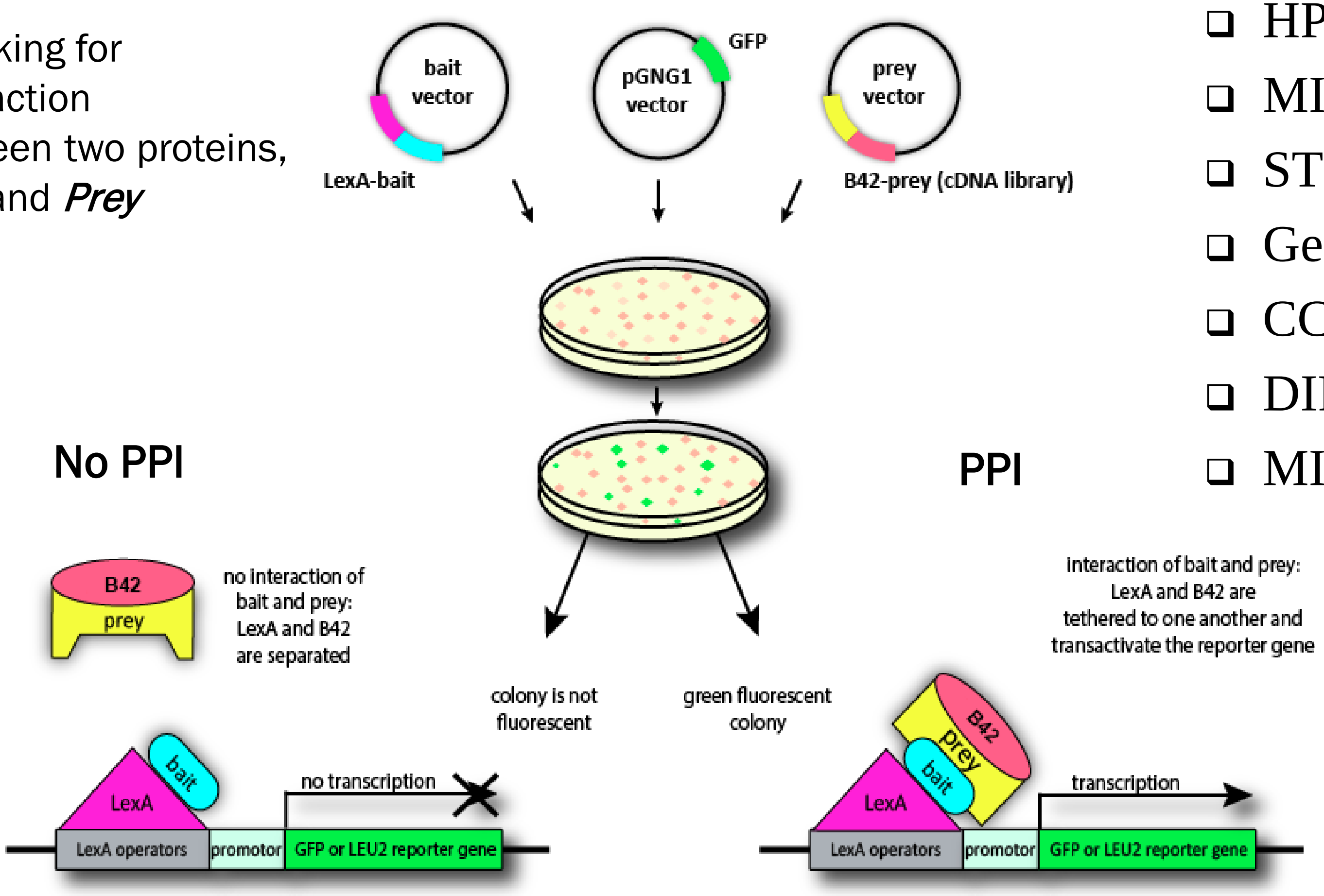
Jeong et al., Nature 2001

Rolland et al., Cell 2014

Protein-Protein Interaction cont.

- ❑ Construction
  - ❑ Yeast-two-hybrid screening

Checking for interaction between two proteins, *Bait* and *Prey*


bait vector
LexA-bait
pGNG1 vector
GFP
prey vector
B42-prey (cDNA library)
No PPI
PPI
B42
prey
no interaction of bait and prey: LexA and B42 are separated
colony is not fluorescent
green fluorescent colony
interaction of bait and prey: LexA and B42 are tethered to one another and transactivate the reporter gene
LexA
bait
no transcription
LexA operators
promotor
GFP or LEU2 reporter gene
B42
prey
bait
LexA
transcription
LexA operators
promotor
GFP or LEU2 reporter gene

- ❑ Major DBs:
  - ❑ BioGRID
  - ❑ HPRD
  - ❑ MIntAct
  - ❑ STRING,
  - ❑ Gene-MANIA,
  - ❑ CCSB Interactome,
  - ❑ DIP,
  - ❑ MINT.

Signal Transduction and Regulatory Networks

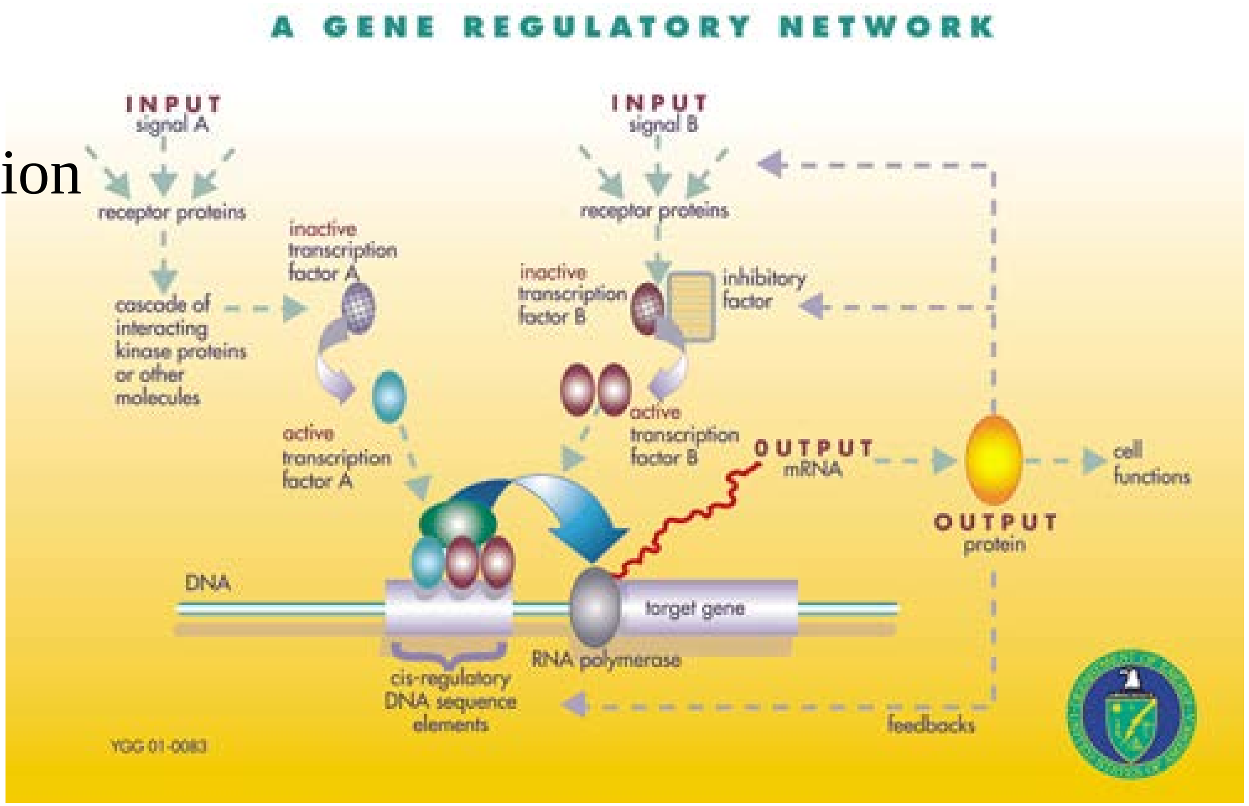
- **Signal transduction**
  - Communication process within a cell to coordinate its responses to an environmental change.
  - Response is a reaction of the cell, e.g., the activation of a gene or the production of energy.
- **Signal transduction network** of a cell
  - Complete network of all signal transduction pathways.
  - Signal transduction pathways: directed network of chemical reactions in a cell from a stimulus to the response

Signal Transduction and Regulatory Networks cont.

- **Gene regulation** is a type of response of a cell to an internal stimulus where expression of a gene is regulated by protein called a transcription factor.
- **Gene regulatory network** is a directed network where nodes represent genes and directed edges represent regulatory interactions
  - Ex> binding of a transcription factor (i.e., source of an edge) to a gene (i.e., target of an edge).
  - Compared to a gene co-expression network, a gene regulatory network attempts to represent the causal <u>(directed)</u> relationships between genes.

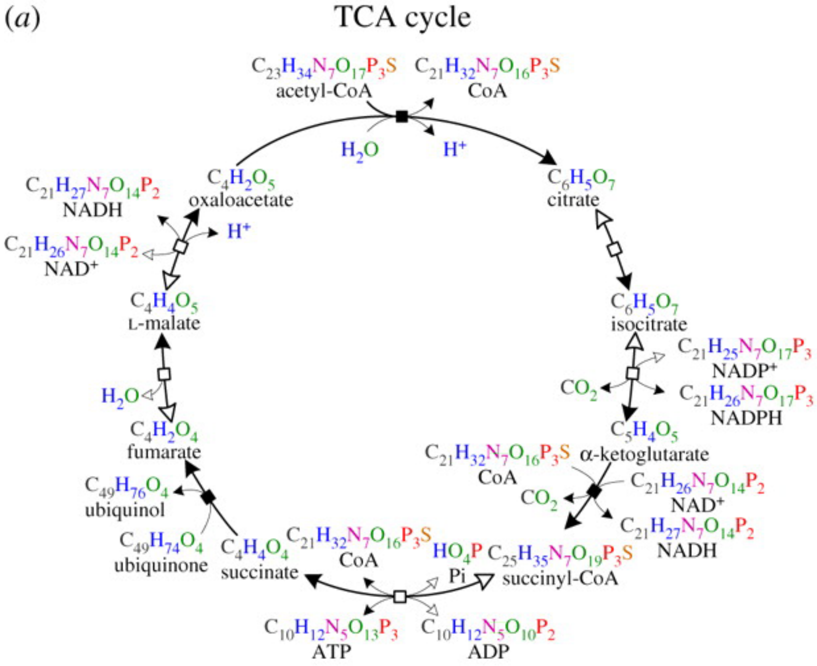
Signal Transduction and Regulatory Networks cont.

- Major DB:
  - Netpath,
  - Pathway Commons,
  - WikiPathways,
  - NCINature
  - Pathway Interaction Database,
  - RegulonDB,
  - TRANSFAC.


A GENE REGULATORY NETWORK
INPUT
signal A
INPUT
signal B
receptor proteins
receptor proteins
inactive transcription factor A
cascade of interacting kinase proteins or other molecules
inactive transcription factor B
inhibitory factor
active transcription factor A
active transcription factor B
OUTPUT
mRNA
OUTPUT
protein
cell functions
DNA
target gene
RNA polymerase
cis-regulatory DNA sequence elements
feedbacks
YGG 01-0083

Metabolic Networks

- **Metabolic reaction** is a chemical process that transforms chemical substances or metabolites (i.e., reactants) into other substances (i.e., products) usually catalyzed by enzymes.
- **Metabolic networks** are directed networks where each
  - Node represents a metabolite (a molecule) and
  - Edge represents a metabolic reaction.


(a)
TCA cycle
$C_{23}H_{34}N_7O_{17}P_3S$
acetyl-CoA
$C_{21}H_{32}N_7O_{16}P_3S$
CoA
H_2O
H^+
$C_4H_2O_5$
oxaloacetate
$C_6H_5O_7$
citrate
$C_{21}H_{27}N_7O_{14}P_2$
NADH
H^+
$C_{21}H_{26}N_7O_{14}P_2$
NAD⁺
$C_4H_4O_5$
L-malate
$C_6H_5O_7$
isocitrate
$C_{21}H_{25}N_7O_{17}P_3$
NADP⁺
CO_2
$C_{21}H_{26}N_7O_{17}P_3$
NADPH
H_2O
$C_4H_2O_4$
fumarate
$C_5H_4O_5$
α-ketoglutarate
$C_{21}H_{32}N_7O_{16}P_3S$
CoA
$C_{21}H_{26}N_7O_{14}P_2$
NAD⁺
CO_2
$C_{49}H_{76}O_4$
ubiquinol
$C_{21}H_{27}N_7O_{14}P_2$
NADH
$C_{49}H_{74}O_4$
ubiquinone
$C_4H_4O_4$
succinate
$C_{21}H_{32}N_7O_{16}P_3S$
CoA
HO_4P
Pi
$C_{25}H_{35}N_7O_{19}P_3S$
succinyl-CoA
$C_{10}H_{12}N_5O_{13}P_3$
ATP
$C_{10}H_{12}N_5O_{10}P_2$
ADP

Metabolic Networks cont

- **Metabolic pathway** is a connected sub-network of the metabolic network either representing specific processes or defined by functional boundaries.
  - Ex> network between an initial and a final chemical substance.
  - **Hyper-graph**: The nodes represent the substances and the directed hyper-edges represent the reactions from reactants to products and is labeled with the enzymes that catalyze the reaction.
  - **Directed bipartite graph**: $G = (V_s; V_r;E)$ with inV_s representing substances, nodes V_r representing metabolic reactions and directed edges E representing the transformation of substance.

Metabolic Networks cont.

- Major DBs
  - BRENDA
  - KEGG PATHWAY Database
  - MANET
  - Reactome
  - Small Molecule Pathway Database
  - MetaNetX.

Other types of networks

- Gene-phenotype network
  - Phenotypes: diseases
- Phylogenetic trees
- Gene Ontology

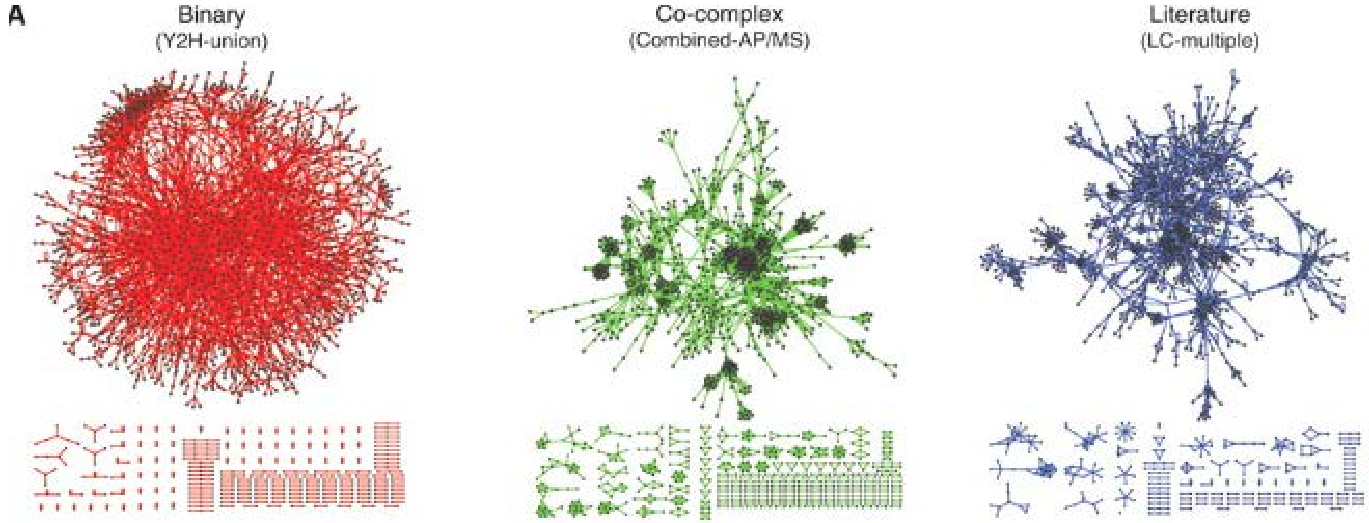
Applications of PPI

- Finding disease modules in networks
  - Method 1: Community detection
- Predicting biological attributes, such as protein functions
  - Method 2: Guilt-by-association principle
  - Method 3: Gene recommender systems

Yu et al., Science 2008

PPI Analysis

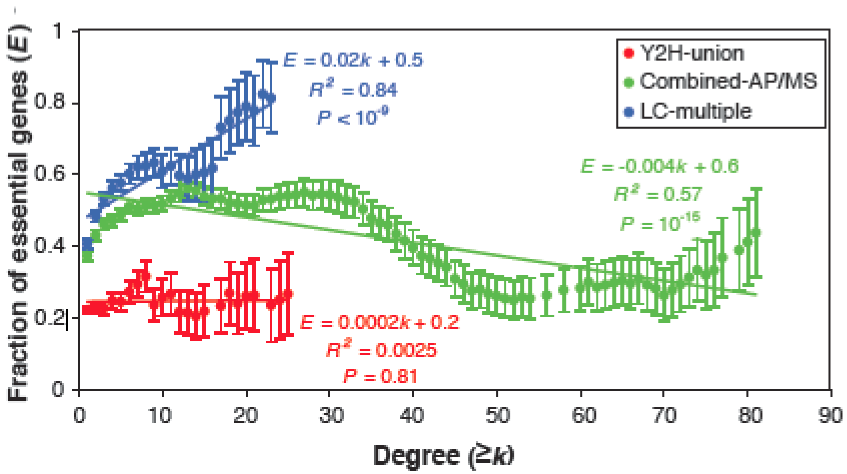
- Yeast Interactome Network (PPI) Data:
  - Three yeast protein-protein interaction (PPI) networks
  - List of **essential** yeast proteins, these proteins form a minimal protein set required for a living cell
  - Mapping of proteins to **phenotypes** associated with **deletion of each protein**

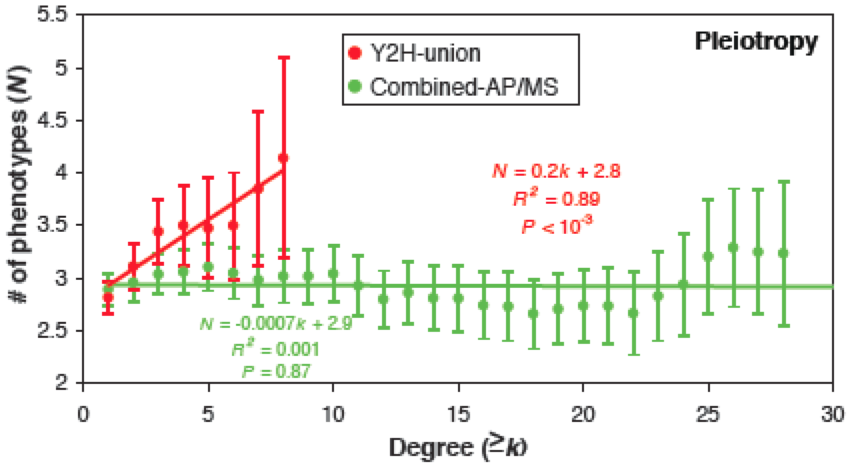


Yu et al., Science 2008

Hub Proteins

- **Hub proteins:** 20% nodes in the network with the highest degree
- Observations:
  - **Hub proteins** associate with **essential proteins**, confirmed in many but not all networks
  - **Hub proteins** associate with **larger numbers of phenotypes** than non-hub proteins





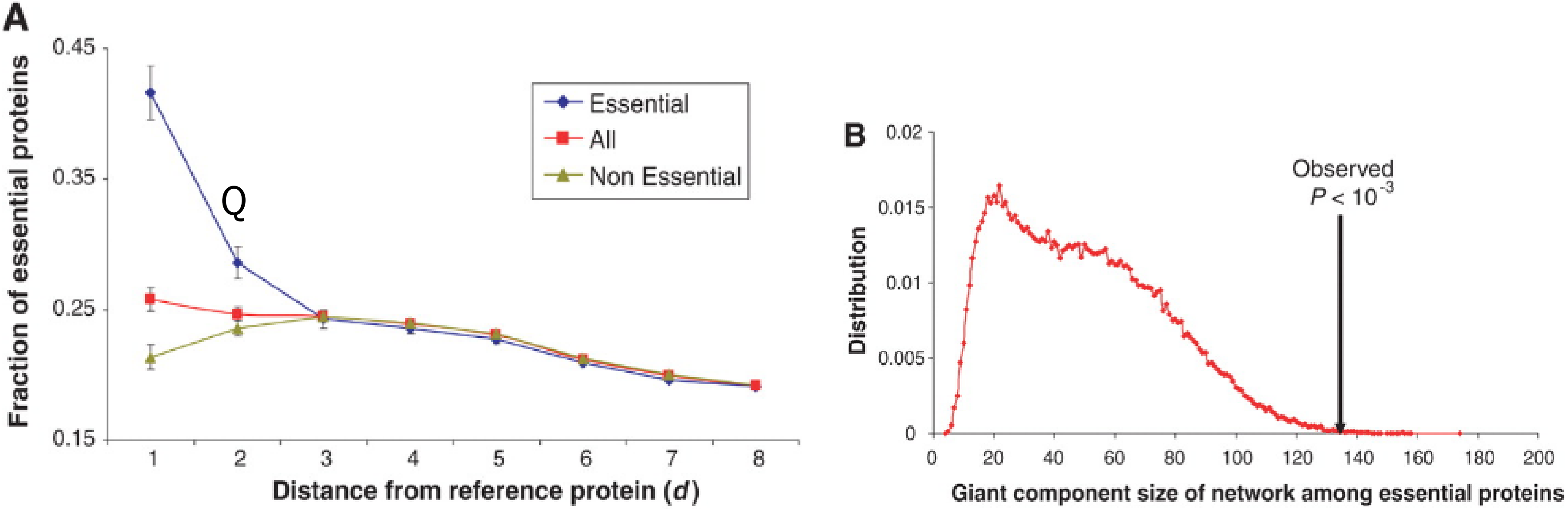
Yu et al., Science 2008

Essential Proteins in PPI

- For a protein p_1, take the **fraction of essential proteins** among all proteins whose distance to protein p_1 is equal to d:

$$Q(p_1, d) = \sum_{p \in S_d(p_1)} \frac{I(p \text{ is essential})}{\left|S_{d(p_1)}\right|}$$

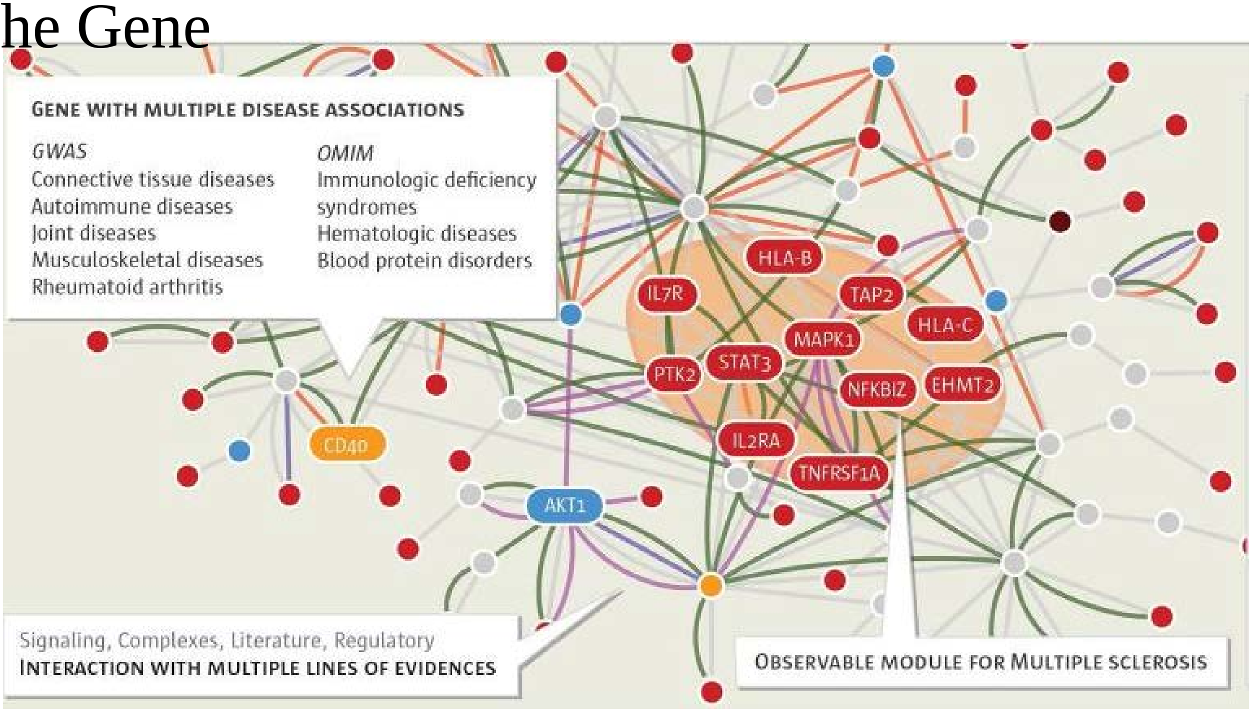
I(x) = 1 if x true
I(x) = 0 otherwise



Menche et al., Science 2015

Disease Protein/Gene

- Given **disease proteins**, compute **shortest path distance** d_s of each disease protein to the closest disease protein $P(d_s)$ is **shifted towards smaller** d_s compared to the random expectation $P^{\text{rand}}(d_s)$
  - ⇒ Disease proteins **agglomerate** in one network neighborhood of increasing the Gene


GENE WITH MULTIPLE DISEASE ASSOCIATIONS
GWAS
Connective tissue diseases
Autoimmune diseases
Joint diseases
Musculoskeletal diseases
Rheumatoid arthritis
OMIM
Immunologic deficiency syndromes
Hematologic diseases
Blood protein disorders
HLA-B
IL7R
TAP2
HLA-C
MAPK1
PTK2
STAT3
NFKBIZ
EHMT2
IL2RA
TNFRSF1A
CD40
AKT1
Signaling, Complexes, Literature, Regulatory
INTERACTION WITH MULTIPLE LINES OF EVIDENCES
OBSERVABLE MODULE FOR MULTIPLE SCLEROSIS

Disease Protein/Gene

- **Disease module** assumption**:** Disease proteins **tend to cluster** in one network neighborhood
- **Local interaction** assumption**:** Disease proteins **tend to interact** with each other
- Mutations in interacting proteins tend to lead to diseases with **similar phenotypes** (i.e., symptoms)

- Disease Module finding/prediction is important!

Mostafavi, Morris, Proteomics 2012

Functional Interaction Networks

- PPI or co-expression network

- Types of protein/gene function prediction
  - **"What does my gene do?"**
    - **Goal:** Determine a gene's function based on who it interacts with – "**guilt-by-association**"
  - **"Give me more genes that function like these"**
    - E.g., Find more multiple sclerosis genes, find new ciliary genes, find more members of a protein complex
  - **"Should there be a connection between A & B"**
    - Drug protein interaction prediction

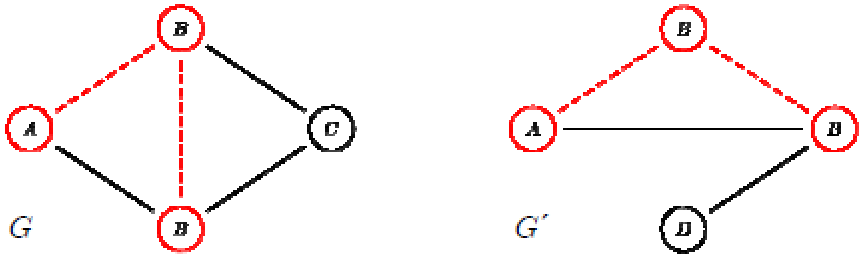
Graph Comparison

Definition 1 **(Graph Comparison Problem)**

Given two graphs G and G′ from the space of graphs G. The problem of graph comparison is to find a mapping

$$s : G \times G' \rightarrow R$$

such that s(G,G′) quantifies the similarity (or dissimilarity) of G and G′.



Isomorphism

Graph isomorphism

Find a mapping f of the vertices of G_1 to the vertices of G_2 such that G_1 and G_2 are identical; i.e. (x,y) is an edge of G_1 iff (f(x),f(y)) is an edge of G_2. Then f is an **isomorphism**, and G_1 and G_2 are called **Isomorphic**

- No polynomial-time algorithm is known for graph isomorphism
- Neither is it known to be NP-complete

Isomorphism

Subgraph isomorphism

G_1 and G_2 are **isomorphic** if there exists a subgraph isomorphism
from G_1 to G_2 and from G_2 to G_1

- Subgraph isomorphism is NP-complete

We want polynomial-time similarity measure for graphs

Measuring graph Similarity 1: Edit Distances

- **Principle**
  - Count operations that are necessary to transform G1 into G2
  - Assign costs to different types of operations (edge/node insertion/deletion, modification of labels)

- **Advantages**
  - Captures partial similarities between graphs
  - Allows for noise in the nodes, edges and their labels
  - Flexible way of assigning costs to different operations
- **Disadvantages**
  - Contains subgraph isomorphism check (NP-complete) as one intermediate step
  - Choosing cost function for different operations is difficult

Measuring graph Similarity 2: Topological Descriptors

- **Principle**
  - Map each graph to a feature vector (ex> finger printing methods)
  - Use distances and metrics on vectors for learning on graphs

- **Advantages**
  - Reuses known and efficient tools for feature vectors

- **Disadvantages**
  - Most feature vector transformation leads to loss of topological information
  - Or includes subgraph isomorphism as one step

Topological Descriptors cont.

feature vectors (chemical **fingerprints)**

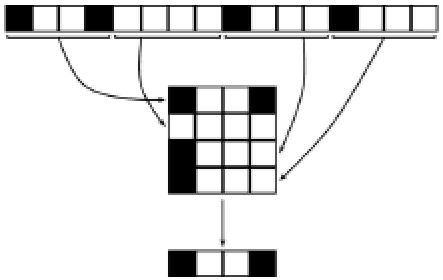
$$\phi(A) = (\phi_s(A))_{s \text{ substructure}}$$

where

$$\phi_s(A) = \begin{cases} 1 & \text{if } s \text{ occurs in } A \\ 0 & \text{otherwise} \end{cases}$$

0	1	0	0	0	1	1	0	1	0	0	0	…

Modulo Compression (lossy)



Elias-Gamma Monotone Encoding (lossless)

[Baldi et al., 2007]

- index $j \to \lfloor log(j) \rfloor$ 0 bits + binary encoding of j
- $j_i < j_{i+1}$: $\lfloor log(j_{i+1}) \rfloor \to \lfloor log(j_i) - log(j_{i+1}) \rfloor$
- average compressed size $= 1,800$ bits

Measuring graph Similarity 3: Graph Kernels

- Kernels on pairs of graphs

- **Principle**
  - Let ϕ(x) be a vector representation of the graph x
  - The kernel between two graphs is defined by:

$$K(x, x') = \phi(x)^T \phi(x')$$

  - To solve convex optimization with kernels, kernels needs to be
    - Symmetric, that is, $k(x, x') = k(x', x)$, and
    - Positive semi-definite (p.s.d.)
  - Comparing nodes in a graph involves constructing a kernel between nodes
  - Comparing graphs involves constructing a kernel between graphs.

Graph Kernels cont.

- **Advantages**
  - Similarity of two graphs are inferred through kernel function

- **Disadvantages**
  - Defining a kernel that captures the semantics inherent in the graph structure and is reasonably efficient to evaluate is the key challenge.

Brief history of graph kernels

- The idea of **constructing kernels *on* graphs** (i.e., between the nodes of a single graph) was first proposed by Kondor and Lafferty (2002), and extended by Smola and Kondor (2003).

- Idea of **kernels *between* graphs** were proposed by G¨artner et al. (2003) and later extended by Borgwardt et al. (2005).

- Idea of **marginalized kernels** (Tsuda et al., 2002) was extended to graphs by Kashima et al. (2003, 2004), then further refined by Mah´e et al. (2004).