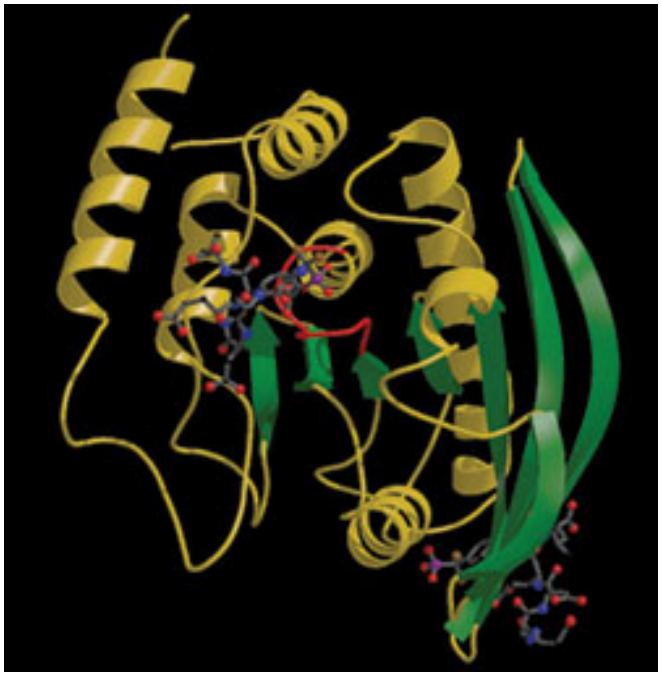
Protein Protein Interactions

02-710 Computational Genomics

Protein Interactions



Assigning Function to Proteins

- While ~25000 genes have been identified in the human genome, for most, we still do not know exactly what they do
- Determining the function of the protein can be done in several ways.
  - Sequence similarity to other (known) proteins
  - Using domain information
  - Using three dimensional structure
  - Based on high throughput experiments (when does it functions and who it interacts with)

Protein Interaction

- In order to fulfill their function, proteins interact with other proteins in a number of ways including:
  - Pathways, for example A -> B -> C
  - Post translational modifications
    - E.g., protein phosphorylation to regulate enzymes
  - Forming protein complexes

Protein interaction

- Traditionally protein interactions were studied in small scale experiments
- Many new proteins from complete genome sequences
- New methods for genome wide interaction data

PPI Lab Experiments

- *Small-scale* PPI experiments
  - One protein or several proteins at a time
  - Small amount of available data
  - Expensive and slow lab process
- *High-throughput* PPI experiments
  - Hundreds / thousands of proteins at a time
  - Highly noisy and incomplete data
  - Surprisingly little overlap among different sets

Methods

- Yeast two-hybrid screens
- Protein complex purification techniques using mass spectrometry

Direct

- Correlated messenger RNA expression profiles
- Genetic interaction data
- '*in silico*' interactions

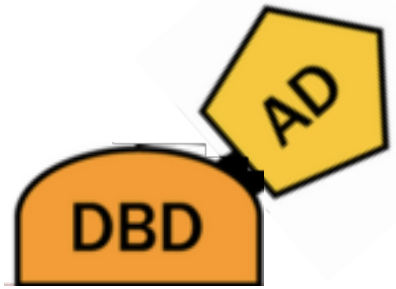
Indirect

- Analysis of PPI networks
  - Classification
  - Network alignment

Yeast two-hybrid assay

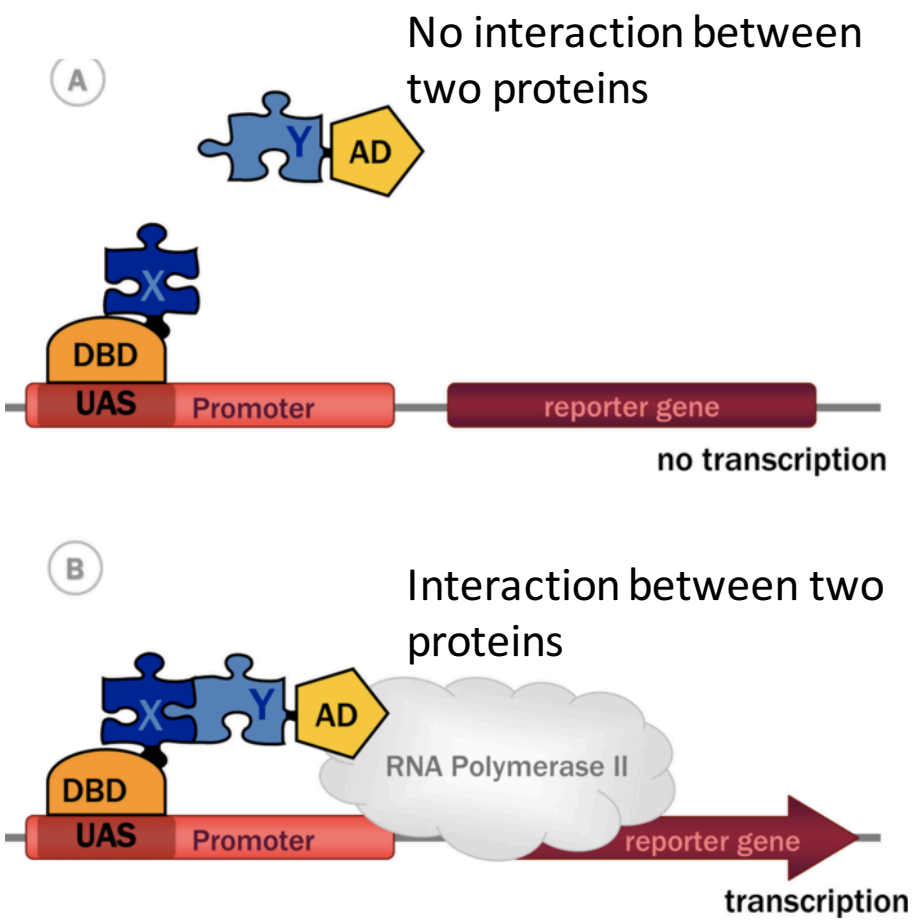
- Yeast transcription factor has a binding domain (BD) and activation domain (AD)
  - BD binds to upstream of the target gene on DNA
  - AD is required to activate transcription

  - BD and AD function independently



Yeast two-hybrid assay

- Bait (X) and prey (Y): two proteins to be tested for interaction
  - Bait is attached to the BD
  - Prey is attached to the AD
- If bait and prey interact,
  - a proper transcription factor is formed
  - the reporter gene is transcribed
- If bait and prey does not interact,
  - a proper transcription factor is not formed
  - the reporter is not transcribed

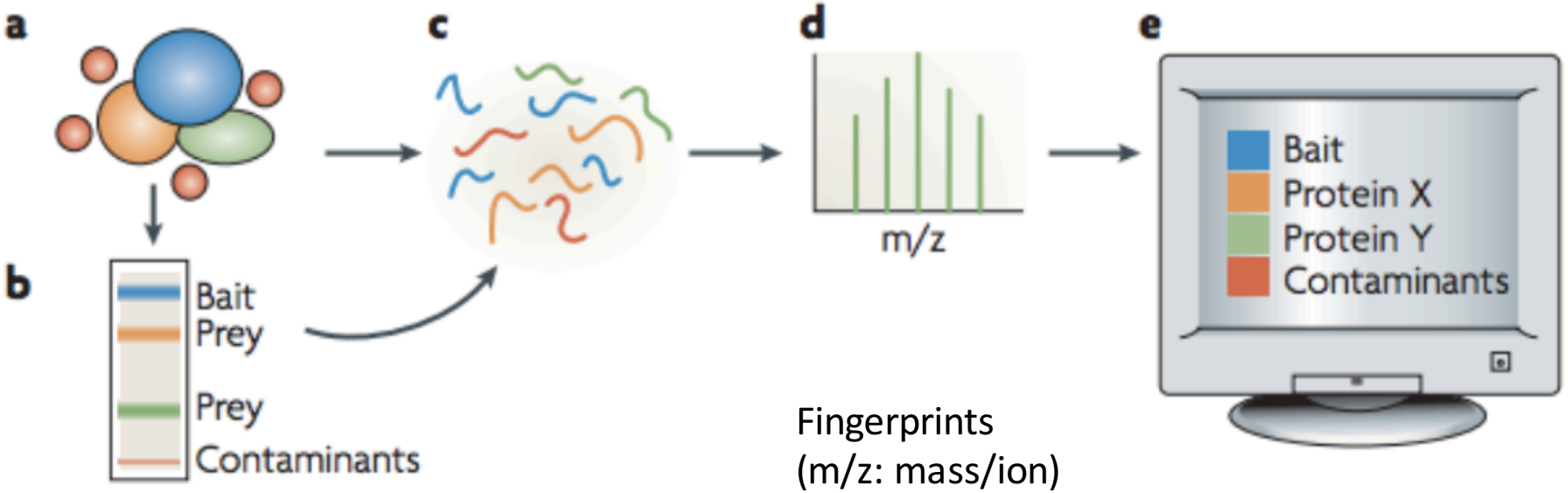


Mass spectrometry (MS) of purified complexes

- Affinity purification and mass spectrometry
  - Multiprotein complexes are isolated directly from cell lysates through one or more affinity purification steps
  - Complex components are then identified by MS

- Unlike two-hybrid assay,
  - MS can can be performed under near physiological conditions in the relevant organism and cell type
  - MS does not perturb post translational modification, thus the effects of post translational modification can be detected

Mass spectrometry (MS) of purified complexes



Digest with enzymes

Fingerprints (m/z: mass/ion) from mass spectrometer

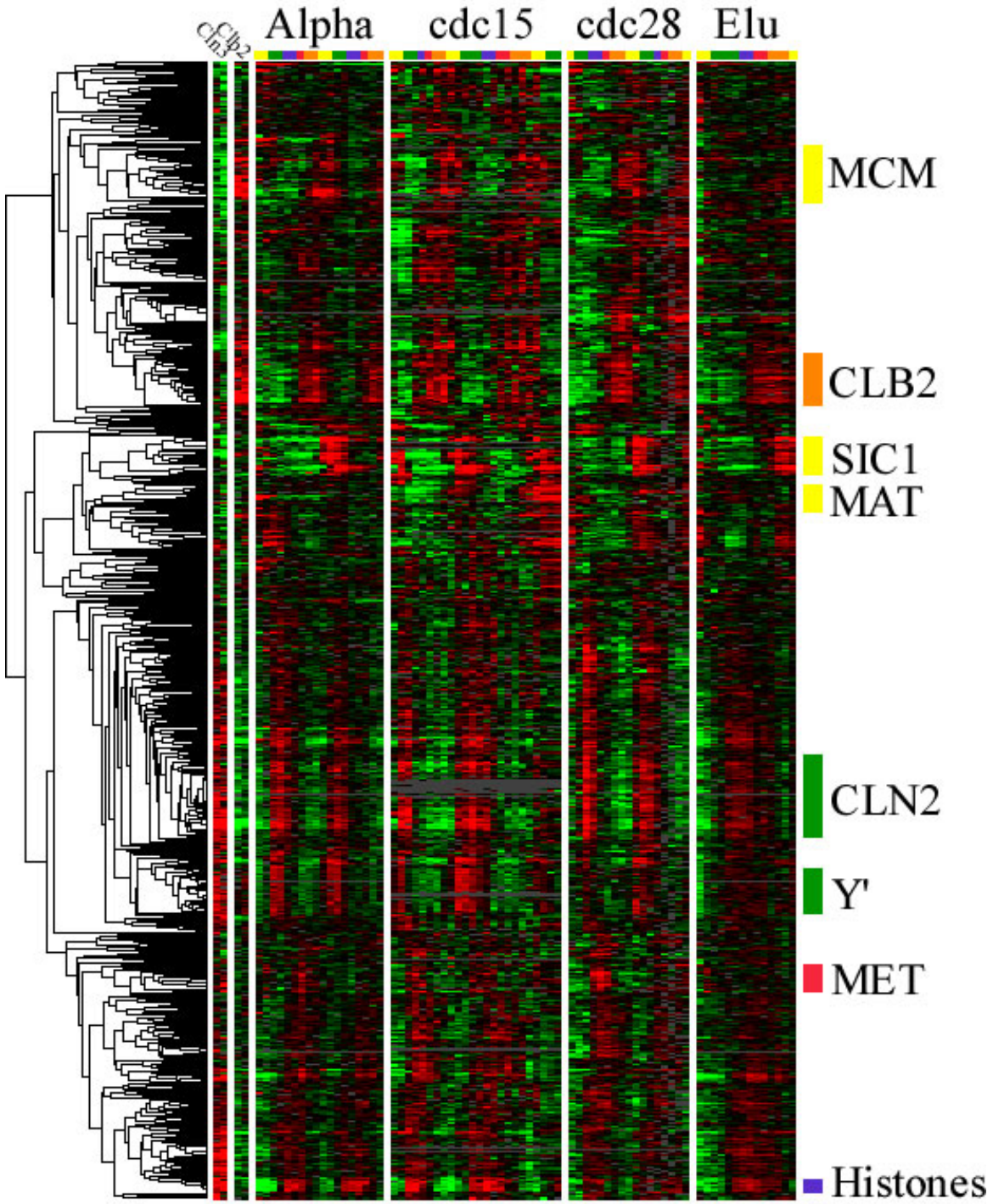
Mass spectrometry: Fast, high-throughput methods for protein sequencing

Identifying PPI from Mass Spectrometry Data

- MS data alone only provide the protein composition not the protein-protein interaction

- Limitation: what happens when one bait protein participates in multiple complexes?

mRNA Expression



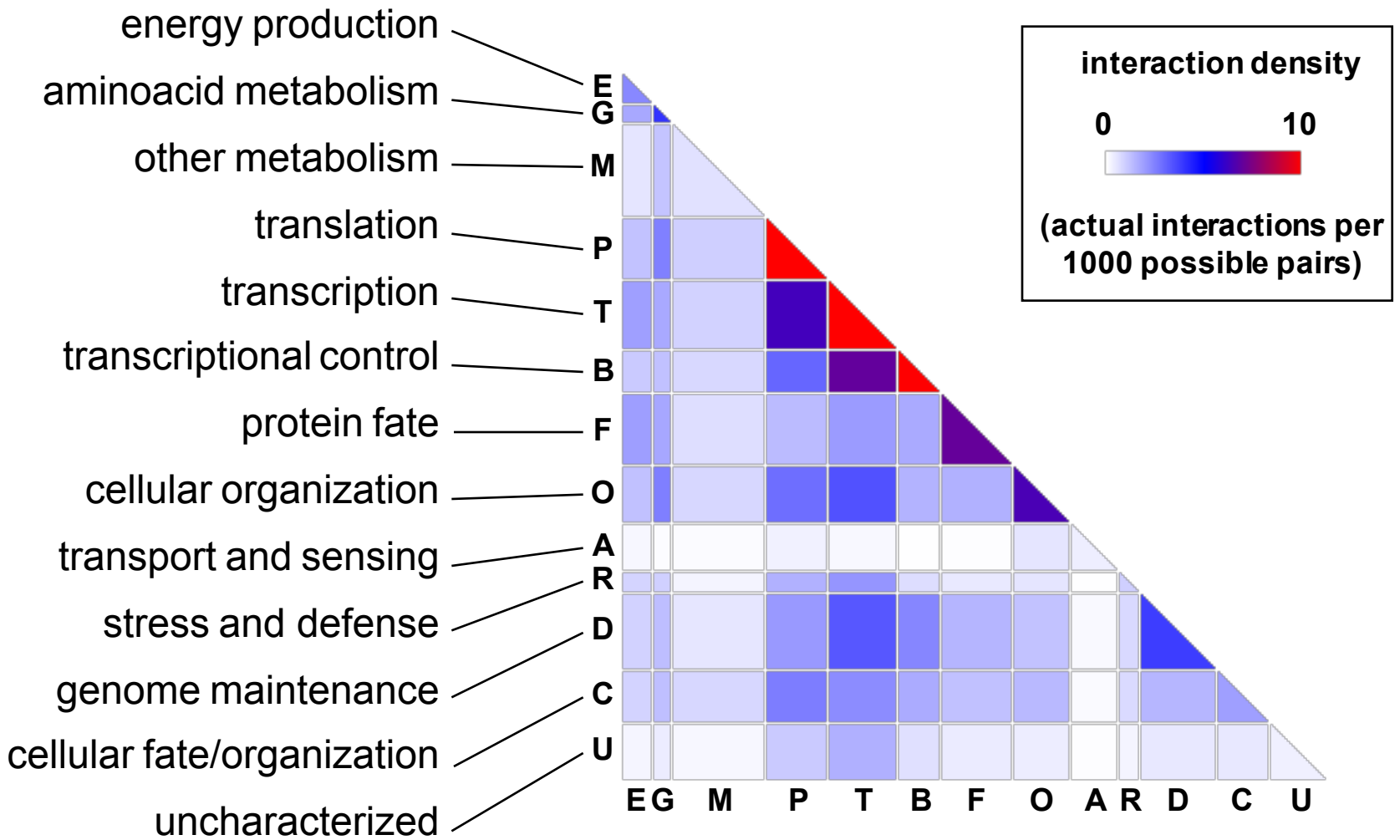
Genetic interactions (synthetic lethality).

- Two nonessential genes that cause lethality when mutated at the same time form a synthetic **lethal interaction**.
- Such genes are often functionally associated and their encoded proteins may also interact physically.

In silico predictions through genome analysis.

- Whole genomes can be screened for three types of interaction evidence:
  - In prokaryotic genomes, interacting proteins are often encoded by conserved operons
  - Interacting proteins have a tendency to be either present or absent together from fully sequenced genomes, that is, to have a similar 'phylogenetic profile';
  - Proteins are sometimes found fused into one polypeptide chain. This is an indication for a physical interaction.

Distribution of interacting proteins (e.g., TAP complexes)



Benchmarking

- Comparing the data with a reference set of trusted interactions allows the estimation of **lower limits** for accuracy and coverage.
- The highest accuracy is achieved for interactions supported by more than one method

Biases in coverage

- Most protein interaction data (including the curetted complexes) are biased towards proteins of high abundance.
- The two “genetic” approaches (two-hybrid and synthetic lethality) appear relatively unbiased.
- Data sets are biased towards particular cellular localizations. For example mitochondrial proteins in the case of the *in silico* predictions. (such protein are of bacterial descent)

Methods

- Yeast two-hybrid screens
- Protein complex purification techniques using mass spectrometry

Direct

- Correlated messenger RNA expression profiles
- Genetic interaction data
- '*in silico*' interactions

Indirect

- Analysis of PPI networks
  - Classification
  - Network alignment

Protein interaction as a classification problem

- Given these direct and indirect datasets, we can design a classifier which will take as an input high throughput data for a pair of proteins

Challenges

- Features are heterogenuous
- Most features are noisy
- Most features have missing values
- Highly skewed class distribution
  - Much more non-interacting pairs than interacting pairs
  - No negative (not interacting) set available
- Only a small positive (interacting) set available

Species	Database (Small-scale PPI)	Genome Size	Predicted # of Interactions	Estimated Ave. Num. Partners Per Protein
Yeast	DIP (3867 interactions ; 1773 proteins)	~6300	~30,000	~10
Human	HPRD (14608 interactions; 5712 proteins)	~25,000	~90,000	~6

PPI Network Alignment

- Comparative analysis of PPI networks across different species by aligning the PPI networks
  - Find functional orthologs of proteins in PPI network of different species
  - Discover conserved subnetwork motifs in the PPI network
- Global vs. local alignment
  - Most of the previous work was focused on local alignment
  - Global alignment can better capture the global picture of how conserved subnetwork motifs are organized – but this is more challenging

PPI Network Alignment

- Challenges
  - How can we align *multiple* PPI networks?: pair-wise alignment is an easier problem
  - How can we use both sequence conservation information and local network topology during the alignment?
    - Conserved subnetworks across species have proteins with conserved sequences as well as conserved interactions with other proteins
    - Most of the previous work was focused on finding orthologs based on the sequence similarities

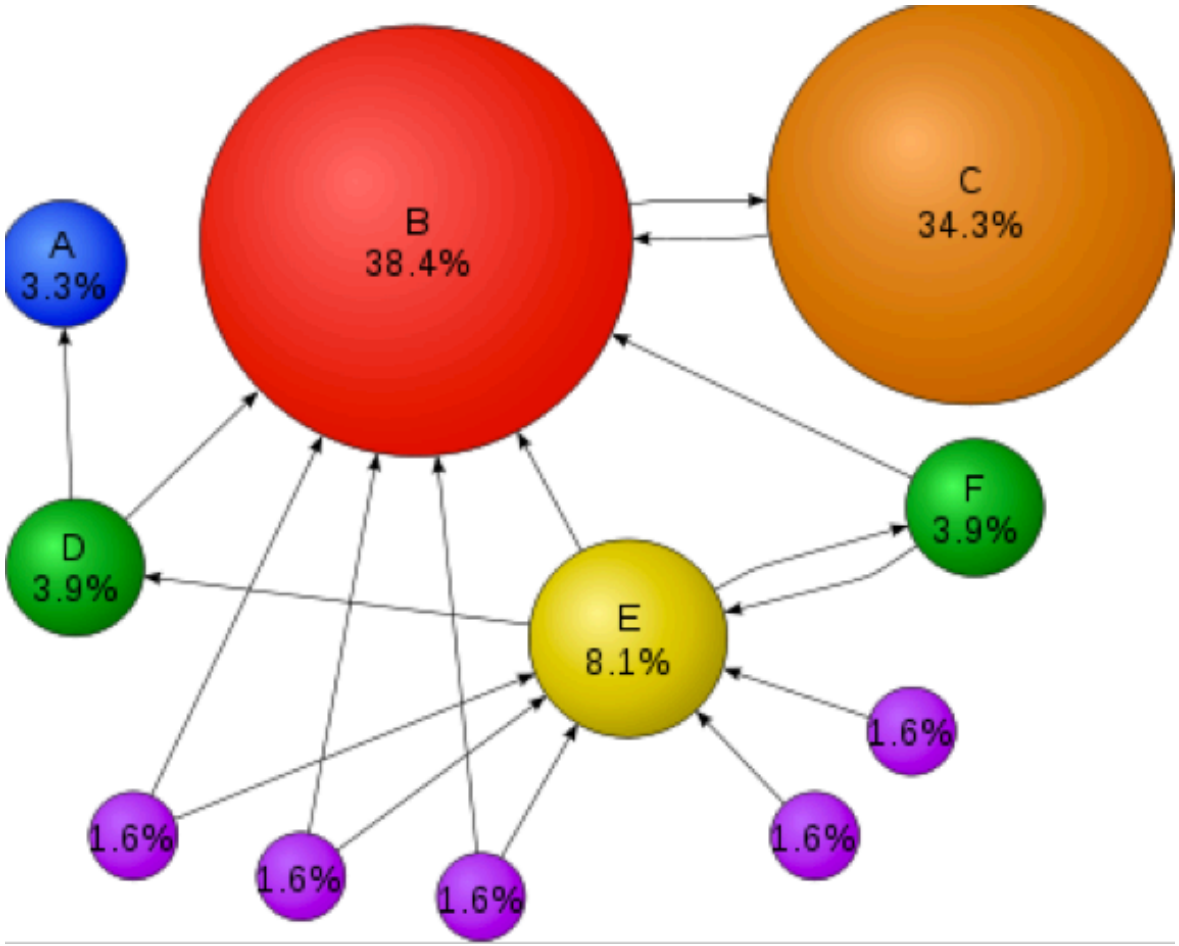
IsoRank and IsoRank-Nibble

- Multiple PPI network alignment for multiple species
- Global alignment
- Alignment based on both sequence and local connectivity conservations
- Based on Google PageRank

PageRank Overview

- Developed by Larry Page and used in Google search engine
- Pages with higher PageRank are returned as search hits
- Algorithm for ranking hyperlinked webpages in the network of webpages
  - Node is each webpage
  - Directed edge from a linking page to the hyperlinked page

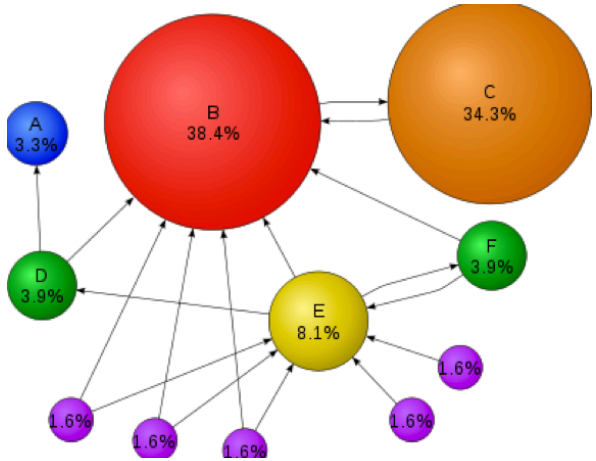
PageRank Illustration



PageRank Overview

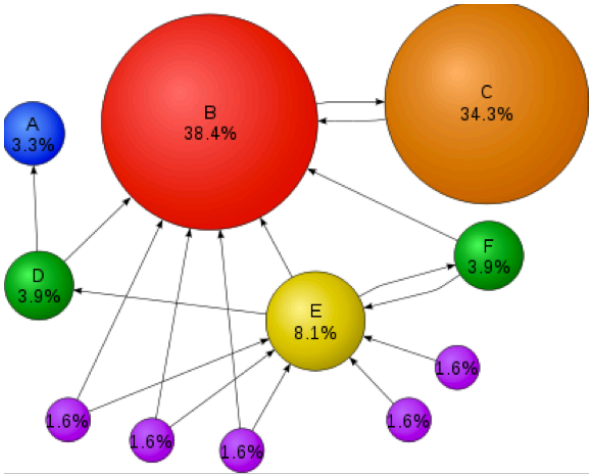
- PageRank models the user behavior

- PageRank for each page is the probability that a websurfer who starts at a random page and takes a random walk on this network of webpages end up at that page
  - With probability d (damping factor), the websurfer jumps to a different randomly selected webpage and starts a random walk
  - Without the damping factor, only the webpages with no outgoing edges will get non-zero PageRanks



PageRank

- The webpages with a greater number of pages linked to it are ranked higher
- If a webpage has multiple hyperlinks, the vote of each outgoing edge is divided by the number of hyperlinks
- The vote of each hyperlink depends on the PageRank of the linking webpage
  - Recursive definition of PageRanks



PageRank

- PageRank p_i of page *i* is given as

$$p_i = (1-d) + d\sum_{j=1}^{N}\left(\frac{L_{ij}}{c_j}\right)p_j$$

  - *d*: damping factor, it ensures each page gets at least (1-*d*) PageRank
  - *N*: the number of webpages
  - L_{ij}=1 if page *j* points to page *i*, and 0 otherwise
  - $c_j = \sum_{i=1}^{N} L_{ij}$

PageRank

- Using matrix notation

$$\mathbf{p} = (1-d)\mathbf{e} + d \cdot \mathbf{L}\mathbf{D}_c^{-1}\mathbf{p}$$

  - p: the vector of length *N*
  - e: the vector of *N* ones
  - $\mathbf{D}_c = \mathrm{diag}(\mathbf{c})$: diagonal elements are c_i
  - **L**: *N*x*N* matrix of $\mathbf{L}_{ij}$'s

- Introduce normalization $\mathbf{e}^T\mathbf{p} = N$ so that average PageRank is 1

$$\begin{aligned} \mathbf{p} &= \left[(1-d)\mathbf{e}\mathbf{e}^T/N + d\mathbf{L}\mathbf{D}_c^{-1}\right]\mathbf{p} \\ &= \mathbf{A}\mathbf{p} \end{aligned}$$

PageRank

- p/N is the stationary distribution of a Markov chain over the N webpages

- In order to find **p**, we use power method
  - Initialize $\mathbf{p} = \mathbf{p}_0$
  - Iterate to find fixed point p

$$\mathbf{p}_k \leftarrow \mathbf{A}\mathbf{p}_{k-1}; \quad \mathbf{p}_k \leftarrow N\frac{\mathbf{p}_k}{\mathbf{e}^T\mathbf{p}_k}$$

IsoRank

- Stage 1: Given two networks G_1 and G_2, compute the similarity scores R_{ij} for a pair of protein for node i in vertex set V_1 in G_1 and protein for node j in vertex set V_2 in G_2
  - Use PageRank algorithm

- Stage 2: Given the matrix R of R_{ij}, find the global alignment using a greedy algorithm

From PageRank to IsoRank

- PageRank ranks webpages, whereas IsoRank ranks the pairs of proteins from the two networks to be aligned.

- PageRank uses the hyperlink information from neighboring nodes to recursively compute the ranks, whereas IsoRank uses the sequence similarity and network connectivity with other neighboring nodes to define the ranks.

IsoRank

- Similarly to PageRank, pairwise similarity score R_{ij} is recursively defined as

$$R_{ij} = \sum_{u \in N(i)} \sum_{v \in N(j)} \frac{1}{|N(u)||N(v)} R_{uv} \quad i \in V_1, j \in V_2$$

  - $N(i)$: the set of neighbors of node u within the graph of u

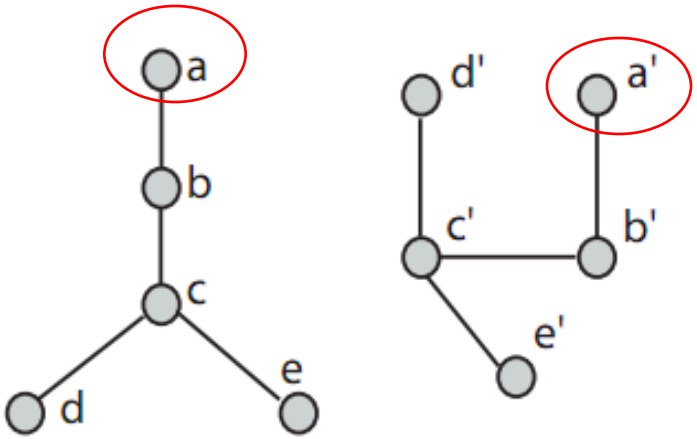
- Using matrix notation

$$R = AR, \quad \text{where}$$

$$A[i, j][u, v] = \begin{cases} \frac{1}{|N(u)||N(v)|} & \text{if } (i, u) \in E_1, (j, v) \in E_2 \\ 0 & \text{otherwise} \end{cases}$$

  - A is a large but sparse matrix

IsoRank Example



R

	a'	b'	c'	d'	e'
a	0.0312		0.0937		
b		0.1250		0.0625	0.0625
c	0.0937		0.2812		
d		0.0625		0.0312	0.0312
e		0.0625		0.0312	0.0312

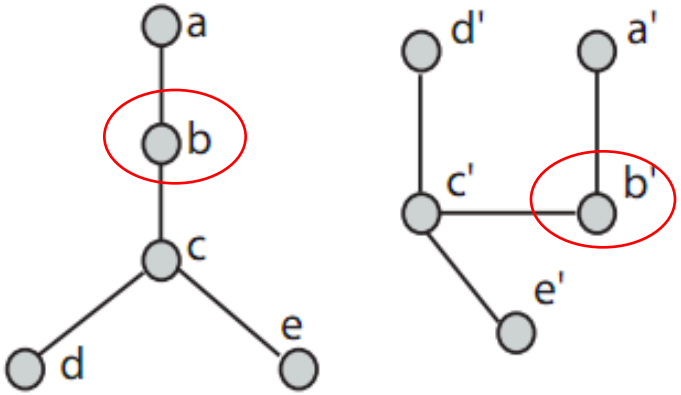
$$R_{aa'} = \frac{1}{4} R_{bb'}$$

$$R_{bb'} = \frac{1}{3} R_{ac'} + \frac{1}{3} R_{a'c} + R_{aa'} + \frac{1}{9} R_{cc'}$$

$$R_{dd'} = \frac{1}{9} R_{cc'}$$

$$R_{cc'} = \frac{1}{4} R_{bb'} + \frac{1}{2} R_{be'} + \frac{1}{2} R_{bd'} + \frac{1}{2} R_{eb'} + \frac{1}{2} R_{db'} + R_{ee'} + R_{ed'} + R_{de'} + R_{dd'}$$

IsoRank Example



R

	a'	b'	c'	d'	e'
a	0.0312		0.0937		
b		0.1250		0.0625	0.0625
c	0.0937		0.2812		
d		0.0625		0.0312	0.0312
e		0.0625		0.0312	0.0312

$$R_{aa'} = \frac{1}{4} R_{bb'}$$

$$R_{bb'} = \frac{1}{3} R_{ac'} + \frac{1}{3} R_{a'c} + R_{aa'} + \frac{1}{9} R_{cc'}$$

$$R_{dd'} = \frac{1}{9} R_{cc'}$$

$$R_{cc'} = \frac{1}{4} R_{bb'} + \frac{1}{2} R_{be'} + \frac{1}{2} R_{bd'} + \frac{1}{2} R_{eb'} + \frac{1}{2} R_{db'} + R_{ee'} + R_{ed'} + R_{de'} + R_{dd'}$$

IsoRank

- When the network edges are weighted

$$R_{ij} = \sum_{u \in N(i)} \sum_{v \in N(j)} \frac{w(i,u)w(j,v)}{\sum_{r \in N(u)} w(r,u) \sum_{q \in N(v)} w(q,v)} R_{uv}$$

$$i \in V_1, j \in V_2$$

- Power method can be used to compute R_{ij}'s

IsoRank

- Incorporating sequence similarity information E

$$R = \alpha A R + (1 - \alpha) E, \qquad 0 \leq \alpha \leq 1, \text{ or}$$
$$R = (\alpha A + (1 - \alpha) E 1^{T}) R.$$

  - α = 0: only sequence similarity information is used but no network information is used.
  - α = 1: only network information is used

IsoRank: Stage 2

- Extracting node-mapping information for global alignment given pairwise similarity scores R_{ij}
  - One-to-one mapping
    - Any node is mapped to at most one node in the network from other species
    - Efficient computation
    - Ignores gene duplication
  - Many-to-many mapping
    - Finds clusters of orthologous genes across networks from different species
  - Mapping criterion: identify pairs of nodes that have high R_{ij} scores, while ensuring the mapping obeys transitive closures – if the mapping contains (a,b) and (b,c), it should contain (a,c)

IsoRank: Stage 2

- One-to-one mapping
  - Greedy approach
  - Select the highest scoring pair

IsoRank: Stage 2

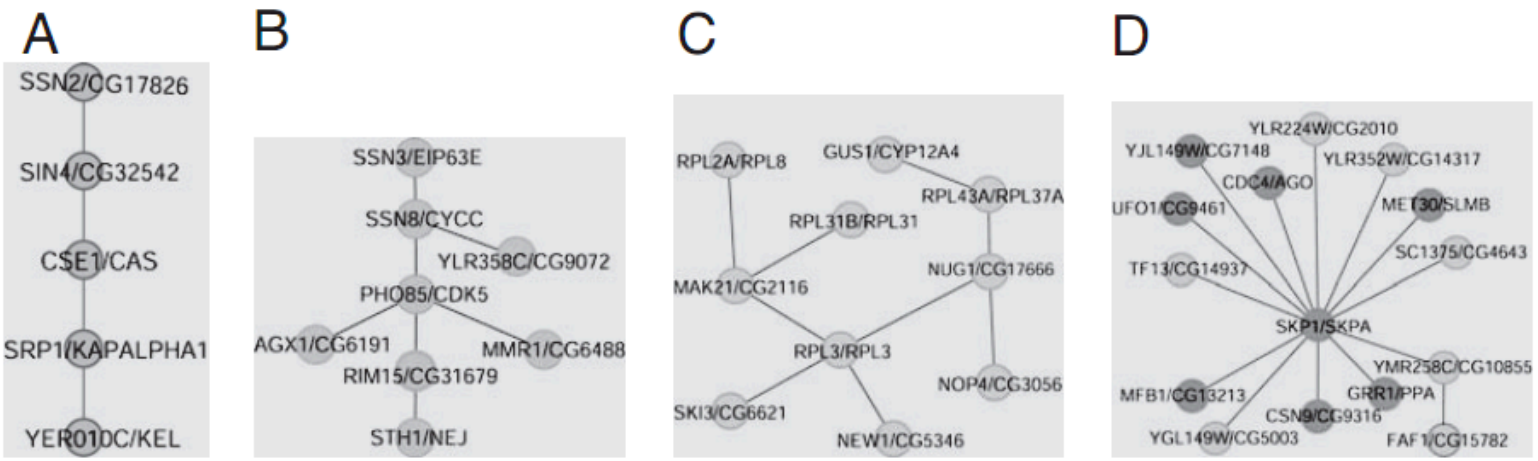
- Many-to-many mapping
  - Greedy approach
  - Form a *k*-partite graph with *k* graphs
  - Iterate until *k*-partite graph has no edges
    - Finding seed pair:
      - select the edge (*i,j*) with the highest score R_{ij} (*i,j* are from two different graphs G_1 and G_2)
    - Extend the seed:
      - In (G_3, …, G_k), find a node *l*, such that 1) R_{lj} and R_{li} are the highest scores between *l* and any node in G_1 and G_2, and 2) R_{li} and R_{lj} exceed a certain threshold
    - Remove from *k*-partite graph the match set

Results

- Alignment PPI networks from five species
  - S. cerevisiae, D. Melanogaster, C. elegans, M. musculus, H. sapiens
  - The common subgraph supported by the global alignment contains
    - 1,663 edges supported by at least two PPI networks
    - 157 edges supported by at least three networks
  - The alignment by sequence-only (no network) method contains
    - 509 edges with support in two or more species
    - 40 edges supported by at least three networks

Results

- Subgraphs selected from yeast-fly PPI network alignment



What you should know

- Different techniques for detecting protein—protein interactions

- Computational methods for analysis of protein-protein interaction data
  - Classification
  - Network alignment