

PPI Network Alignment

02-715 Advanced Topics in Computational
Genomics

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
- Algorithm for ranking hyperlinked webpages in the network of webpages
 - Node is each webpage
 - Directed edge from a linking page to the hyperlinked page
- Pages with higher PageRank are returned as search hits

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, and 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

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

- \mathbf{p} : the vector of length N
 - \mathbf{e} : the vector of N ones
 - $\mathbf{D}_c = \text{diag}(\mathbf{c})$: diagonal elements are c_i
 - \mathbf{L} : $N \times N$ matrix of 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

- \mathbf{p}/N is the stationary distribution of a Markov chain over the N webpages
- In order to find \mathbf{p} , we use power method
 - Initialize $\mathbf{p} = \mathbf{p}_0$
 - Iterate to find fixed point \mathbf{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

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

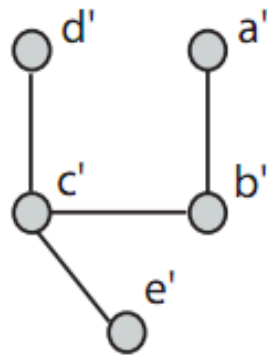
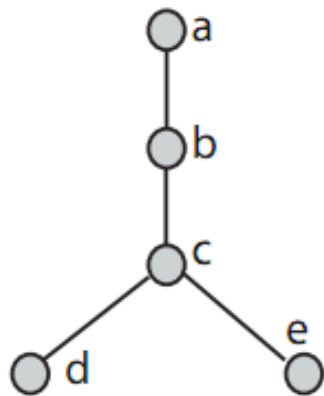
– $N(i)$: the set of neighbors of node u within the graph of u

- 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

- 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

$$R = \alpha AR + (1 - \alpha)E, \quad 0 \leq \alpha \leq 1, \text{ or}$$

$$R = (\alpha A + (1 - \alpha)E1^T)R.$$

- $\alpha = 0$: only sequence similarity information is used but no network information is used.
- $\alpha = 1$: 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

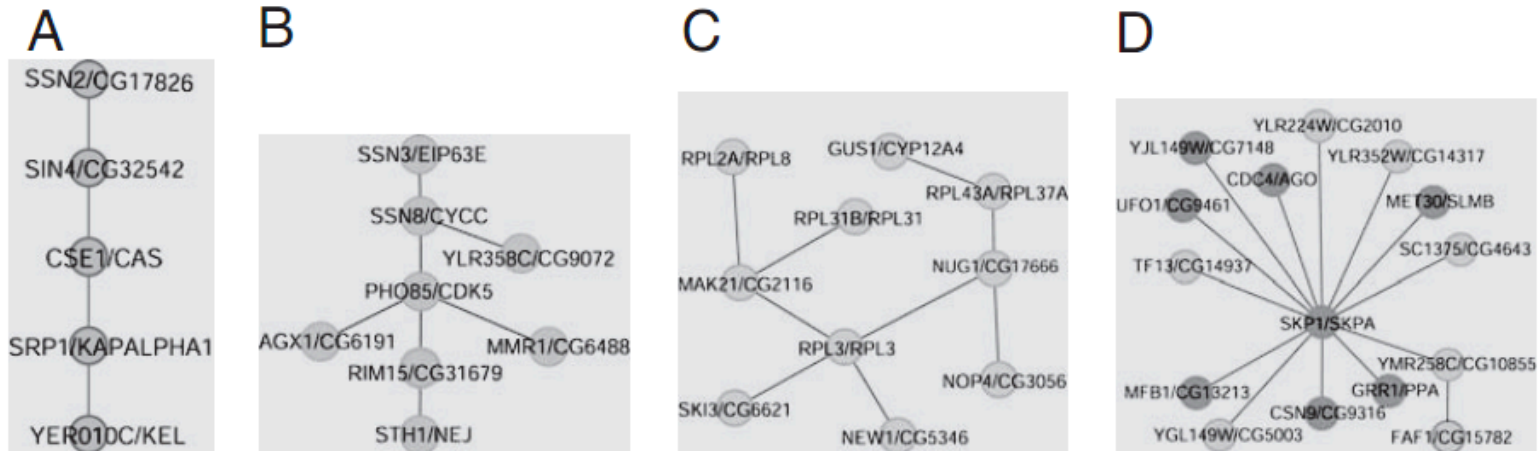
- Many-to-many mapping
 - Greedy approach
 - Form a k -partite graph with k graphs
 - Iterate until k -partite graph has no edges
 - Finding seed pair:
 - select the edge (i,j) with the highest score R_{ij} (i,j are from two different graphs G_1 and G_2)
 - Extend the seed:
 - In (G_3, \dots, G_k) , 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



Summary

- IsoRank uses both sequence similarity and network information to align multiple PPI networks from different species
- IsoRank adopts PageRank algorithm to solve the problem of global PPI network alignment