

# Assessing and Predicting Protein Interactions Using Both Local and Global Network Topological Metrics

Guimei Liu	National University of Singapore
Jinyan Li	Nanyang Technological University
Limsoon Wong	National University of Singapore

1

## Outline

- Background
- Related work
- Our method
  - Local network topological metric
  - Global network topological metric
- Experiments
- Summary

2

## Background

- Protein-protein interactions play a critical role in most cellular processes and form the basis of biological mechanisms.
- High-throughput experimental techniques enable the study of protein-protein interactions at the proteome scale.
- However, high-throughput protein interaction data are often associated with high false positive and false negative rates
  - limitations of the associated experimental techniques
  - dynamic nature of protein interaction maps
  - ...

3

## Computational methods

- A weight is assigned to each interaction such that the higher the weight is, the more likely the interaction is true
- Various Information have been used
  - 3D protein structures
  - co-evolution
  - co-localization
  - gene fusion
  - literature
  - **network topology**
  - protein domains/motifs
  - ...

4

## Methods based on network topology

- Represents PPI networks as undirected graphs, where vertices are proteins, and edges represent interactions between proteins.
- IG1 [Saito et al. 2002]
  - The first one on evaluating the reliability of PPIs using solely PPI network topology
  - Mainly for PPI data generated by yeast-two-hybrid experiments
  - Given a protein pair (u,v), IG1 is calculated based on the number of proteins that interact with and only with either u or v
- IG2 [Saito et al. 2002]
  - Uses 5 local network motifs
  - Performs better than IG1
- IRAP [Chen et al. 2005]
  - the collective reliability of the strongest alternative path between two proteins
  - Expensive to compute
- CD-distance [Brun et al. 2003] and FSWeight [Chua et al. 2006]
  - Based on the number of common neighbors of two proteins
  - Easy to compute
  - They are initially proposed for function prediction
  - Outperforms the previous three methods on large PPI networks [Chen et al. 2006]

5

## CD-distance

- Given a pair of proteins (u, v) in a PPI network  $G=(V, E)$ 
  - $N_u$ : the set of neighbors of u in G
  - $N_v$ : the set of neighbors of v in G
- $$CD(u,v) = \frac{2|N_u \cap N_v|}{|N_u| + |N_v|}$$
- Consider relative intersection size of the two neighbor sets, not absolute intersection size
  - Case 1:  $|N_u|=1, |N_v|=1, |N_u \cap N_v|=1, CD(u,v)=1$
  - Case 2:  $|N_u|=10, |N_v|=10, |N_u \cap N_v|=10, CD(u,v)=1$

6

## FSWeight

- Try to overcome the weakness of CD-distance

$$\bullet \text{FS}(u,v) = \frac{2 | N_u \cap N_v |}{| N_u | + | N_u \cap N_v | + \lambda_u} \times \frac{2 | N_u \cap N_v |}{| N_v | + | N_u \cap N_v | + \lambda_v}$$

Where  $\lambda_u$  and  $\lambda_v$  are used to penalize those proteins with very few neighbors

$$\lambda_u = \max \left\{ 0, \frac{\sum_{x \in G} | N_x |}{| V |} - | N_u | \right\}, \quad \lambda_v = \max \left\{ 0, \frac{\sum_{x \in G} | N_x |}{| V |} - | N_v | \right\}$$

- Suppose the average degree is 4, then
  - Case 1:  $|N_u|=1, |N_v|=1, |N_u \cap N_v|=1, \text{FS}(u,v)=4/25=0.16$
  - Case 2:  $|N_u|=10, |N_v|=10, |N_u \cap N_v|=10, \text{FS}(u,v)=1$

7

## Our method

- CD-distance and FSWeight are local metrics
- We use both local and global network topological metric
  - Local metric
    - a variant of CD-distance
    - Computed iteratively
  - Global metric
    - Computed using interacting protein group pairs
- These two metrics are combined to get the final score of an interaction

8

## Local topological metric

- A variant of CD-distance which penalizes proteins with few neighbors

$$w_L(u,v) = \frac{2 |N_u \cap N_v|}{|N_u| + \lambda_u + |N_v| + \lambda_v}$$

$$\lambda_u = \max \left\{ 0, \frac{\sum_{x \in G} |N_x|}{|V|} - |N_u| \right\}, \quad \lambda_v = \max \left\{ 0, \frac{\sum_{x \in G} |N_x|}{|V|} - |N_v| \right\}$$

(same as in FSWeight)

- Iterate local topological metric
  - Motivation: the weight of an interaction reflects its reliability, so can we get better results if we use this weight to re-calculate the score of other interactions?

9

## Iterate local metric

- $w_L^0(u,v) = 1$  if  $(u,v) \in G$ , otherwise  $w^0(u,v) = 0$

$$\bullet w_L^1(u,v) = \frac{|N_u \cap N_v| + |N_u \cap N_v|}{|N_u| + \lambda_u + |N_v| + \lambda_v}$$

$$\bullet w_L^k(u,v) = \frac{\sum_{x \in N_u \cap N_v} w_L^{k-1}(u,x) + \sum_{x \in N_u \cap N_v} w_L^{k-1}(v,x)}{\sum_{x \in N_u} w_L^{k-1}(u,x) + \lambda_u + \sum_{x \in N_v} w_L^{k-1}(v,x) + \lambda_v}$$

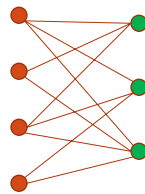
$$\bullet \lambda_u^k = \max \left\{ 0, \frac{\sum_{x \in V} \sum_{y \in N_x} w_L^{k-1}(x,y)}{|V|} - \sum_{x \in N_u} w_L^{k-1}(u,x) \right\}$$

$$\bullet \lambda_v^k = \max \left\{ 0, \frac{\sum_{x \in V} \sum_{y \in N_x} w_L^{k-1}(x,y)}{|V|} - \sum_{x \in N_v} w_L^{k-1}(v,x) \right\}$$

10

## Global topological metric

- Observation:
  - if one group of proteins interact with another group of proteins, then it is likely that the interaction between these two protein groups is mediated by an underlying complementary binding domain/motif pair.
  - In a protein pair participates in an interacting group pair, then the interaction between them is likely to be true



11

## Calculating global topological metric

- Three steps:
  - Step 1: generate protein groups that have common interacting partners
  - Step 2: calculate the interacting score of the generated protein groups
  - Step 3: calculate the global topological score of a protein pair

12

## Step 1: generate protein groups

- A protein group is considered if it
    - Contains at least  $s$  proteins
    - Its members have at least  $t$  common interacting partners
  - The adjacency matrix of an undirected graph can be regarded as a transaction database
    - Each adjacency list is a transaction
    - Each protein is an item
- ⇒ finding protein groups can be mapped to finding frequent itemsets that contain at least  $s$  items and appear in at least  $t$  transactions.
- ⇒ We use a frequent itemset mining algorithm to find qualified protein groups

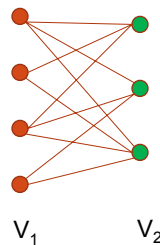
13

## Step 2: calculate interaction score of protein groups

- Let  $V_1$  and  $V_2$  be two protein groups generated in Step 1, the interacting confidence score of  $V_1$  and  $V_2$  is defined as

$$\text{conf}(V_1, V_2) = \frac{\# \text{interactions between } V_1 \text{ and } V_2}{\# \text{possible protein pairs between } V_1 \text{ and } V_2}$$

- Example
  - #interactions: 10
  - #possible protein pairs: 12
  - $\text{conf}(V_1, V_2) = 10/12 = 0.833$



14

### Step 3: calculating global metric score

- The global interacting score of a protein pair is computed based on the interacting confidence score of the interacting group pairs it participates in and the degree of its participation

- Given a protein pair (u,v)

$$w_G(u,v) = \max\{conf(V_1, V_2) \times \frac{2|N_u \cap V_2|}{|N_u| + |V_2|} \times \frac{2|N_v \cap V_1|}{|N_v| + |V_1|} \mid u \in V_1, v \in V_2\}$$

- $\frac{2|N_u \cap V_2|}{|N_u| + |V_2|}$  is u's participation degree in interacting protein group pair (V<sub>1</sub>, V<sub>2</sub>)
- $\frac{2|N_v \cap V_1|}{|N_v| + |V_1|}$  is v's participation degree in (V<sub>1</sub>, V<sub>2</sub>)

15

### Combine local metric and global metric

- The final score of a protein pair (u,v) is defined as the sum of its local metric score and global metric score

$$LGTweight(u, v) = w_L^k(u, v) + w_G(u, v)$$

16

## Experiments

- PPI dataset: DIP yeast (dated 07-Oct-2007)
  - 4932 proteins and 17491 interactions
  - Core dataset: 6459 interactions
- Evaluation methods:
  - Functional homogeneity
    - Use Gene Ontology (GO) annotations
  - Localization coherence
    - use Gene Ontology (GO) annotations
  - 5-fold cross validation
    - DIP core dataset

17

## GO annotations

- Select only informative GO terms.
  - A GO term is informative if no less than 30 proteins are annotated with that term, and none of its descendant terms has at least 30 proteins
- 50 molecular function terms and 110 biological process terms
  - 3251 proteins and 11229 interactions have functional annotations.
- 42 cellular component terms
  - 1615 proteins and 4246 interactions have cellular component annotations

18

## Functional Homogeneity

- Given a set of protein pairs, its functional homogeneity is defined as

$$\frac{\text{\#protein pairs sharing same function annotation}}{\text{\#protein pairs that have function annotations}}$$

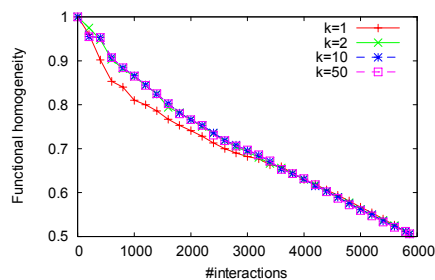
- Similarly, localization coherence is defined as

$$\frac{\text{\#protein pairs sharing same localization annotation}}{\text{\#protein pairs that have localization annotations}}$$

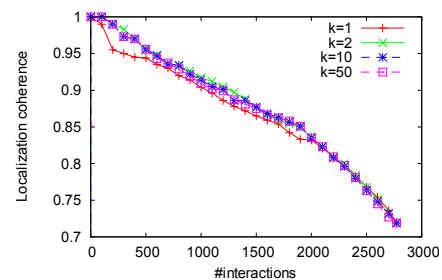
19

## Experiment 1: the effect of k to local metric

- Assessing the reliability of PPIs in DIP dataset



Functional homogeneity

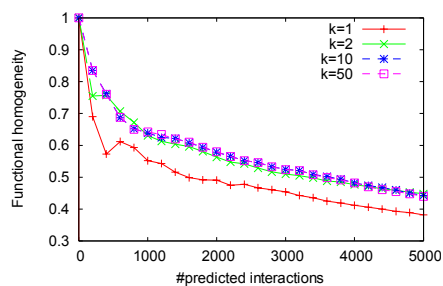


Localization coherence

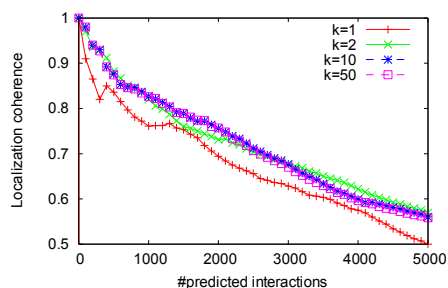
20

## The effect of k to local metric

- Predicting new PPIs



Functional homogeneity

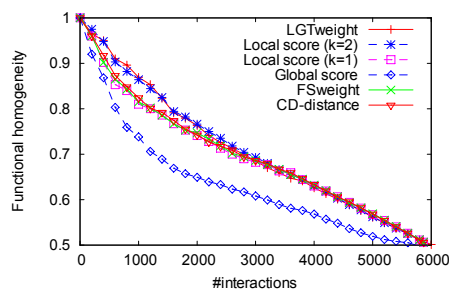


Localization coherence

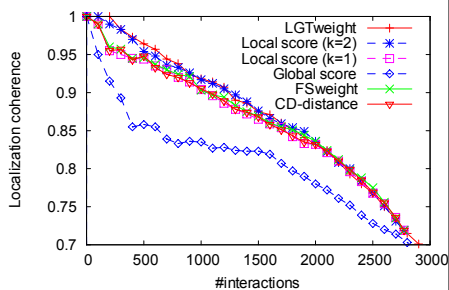
21

## Experiment 2: comparing different scoring methods

- Assessing the reliability of PPIs in DIP dataset
- For global topological metric, we set  $s=5$ ,  $t=1$



Functional homogeneity

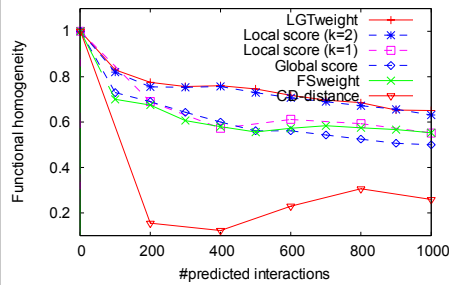


Localization coherence

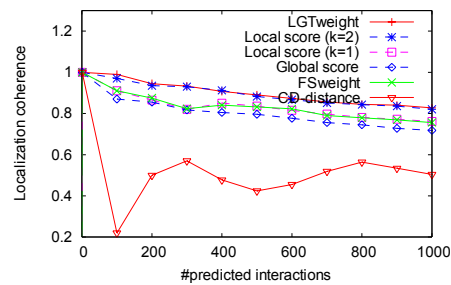
22

## Comparing different scoring methods

- Predicting new PPIs



Functional homogeneity



Localization coherence

23

## Experiment 3: 5-fold cross-validation

- Use the DIP core dataset as the golden standard
- Divide the proteins in the DIP full yeast dataset into 5 disjoint groups.
- For each group of proteins
  - Training data: remove the interactions between proteins in the group, and use the remaining interactions as training data
  - Testing data: all the protein pairs within this group
  - Correct answer PPIs: the pairs of proteins in the group that are in the core dataset

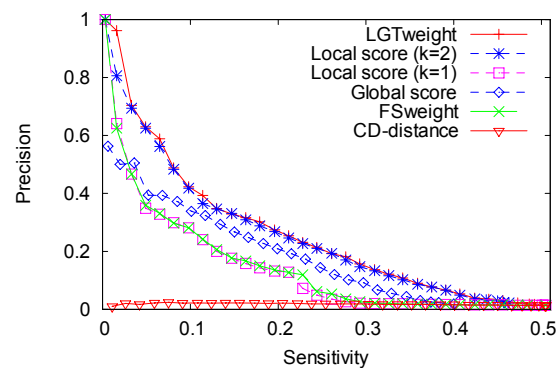
24

## 5-fold cross-validation

- Average number of proteins in 5 groups: 986
- Average number of interactions in 5 training datasets: 16723
- Average number of interactions in 5 testing datasets: 486591
- Average number of correct answer interactions: 307
- Measures:
  - sensitivity =  $TP / (TP + FN)$
  - specificity =  $TN / (TN + FP)$ 
    - #negatives  $\gg$  #positives, specificity is always very high
    - $>97.8\%$  for all scoring methods
  - precision =  $TP / (TP + FP)$

25

## 5-fold cross-validation



26

## Summary and Conclusion

- Assessing the reliability of PPIs
  - CD-distance, FSWeight, and Local metric show similar performance
  - Iterating local metric can improve the performance slightly
- Predicting new interactions
  - CD-distance is not good at predicting new interactions
  - Iterating local metric can improve the performance significantly
- CD-distance and FSWeight can also be iterated, and they show similar improvement as local metric
- The global metric does not improve the performance much, but if an interaction has both high local metric score and high global metric score, then the interaction

27

## Q&A

- Thank you for your attention

28

## Rank difference and score difference

- Given an interaction  $(u,v)$ , the rank difference of  $(u,v)$  at  $k$ -th iteration is

$$\text{rank\_diff}^k(u,v) = |\text{rank}^k(u,v) - \text{rank}^{k-1}(u,v)|$$

- Given a set of interactions  $E$ , the average ranking difference of all the interactions in  $E$  at  $k$ -th iteration is defined as

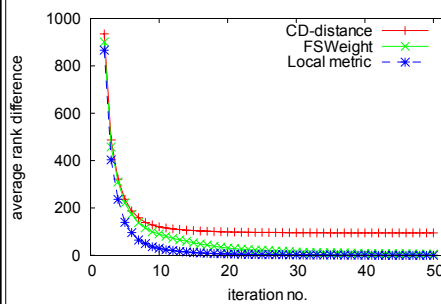
$$\text{avg\_rank\_diff}^k(E) = \frac{\sum_{(u,v) \in E} |\text{rank}^k(u,v) - \text{rank}^{k-1}(u,v)|}{|E|}$$

- Similarly, we can define average score difference of all the interactions in  $E$  at  $k$ -th iteration

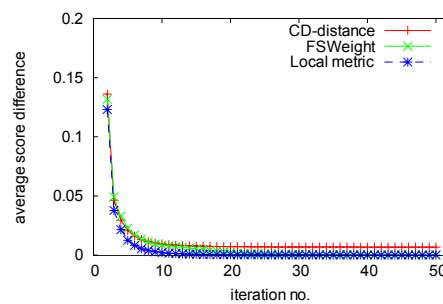
$$\text{avg\_score\_diff}^k(E) = \frac{\sum_{(u,v) \in E} |\text{score}^k(u,v) - \text{score}^{k-1}(u,v)|}{|E|}$$

29

## Results on the DIP dataset



Rank difference



Score difference

30