Sim_{VICD}: An integrative measure of graph- and vector-based semantic similarity using information content distance

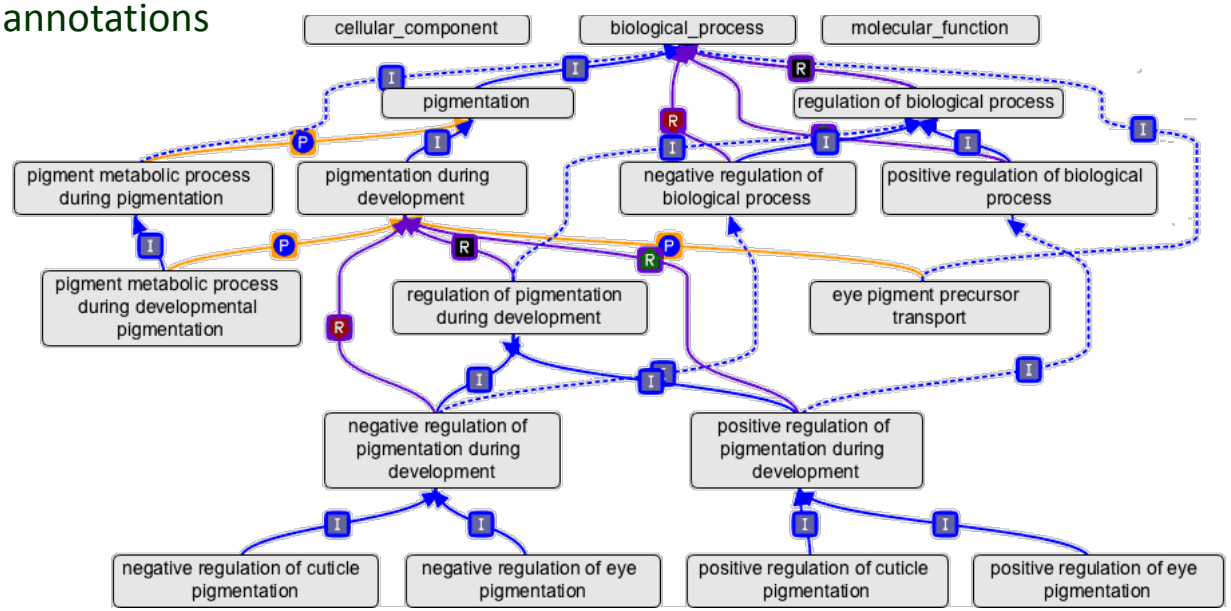
Qiaoli Hu, Yong Shi, and **Young-Rae Cho**

Baylor University

Gene Ontology (GO)

- **Significant Features as a Benchmark**
  - Comprehensiveness of GO terms on three ontologies
  - Extensive coverage of annotating genes and gene products in a genomic scale

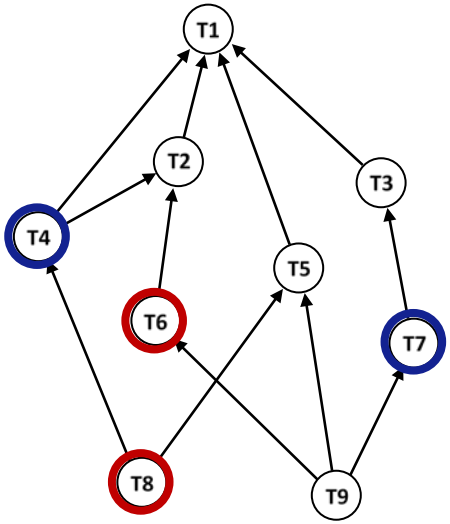
- **Properties for Effective Analysis**
  - DAG structure of GO terms
  - Transitivity property of annotations

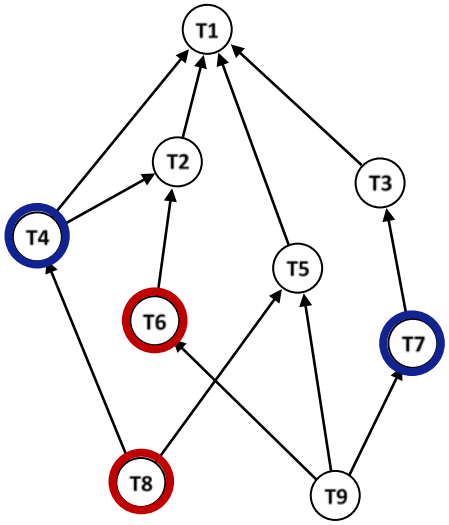


1. *http://geneontology.org/page/ontology-structure*

Semantic Similarity

- **Definition**
  - Function to return a numeric value reflecting closeness in meaning between terms
- **Application**
  - Quantifying functional similarity between two genes or two proteins
- **Types**
  - Pairwise approaches
  - Group-wise approaches





Survey of Semantic Similarity Measures (1)

- **Node-based Methods**
  - Common-term-based
    - Measuring the ratio of common ancestor terms between two terms
    - pairwise method

$$sim_{node}(t_i, t_j) = \frac{|N(t_i) \cap N(t_j)|}{|N(t_i) \cup N(t_j)|}$$

    - $N(t_i)$ is a set of all ancestor terms of t_i including t_i itself
    - group-wise method

$$sim_{ui}(S_i, S_j) = \frac{|N(S_i) \cap N(S_j)|}{|N(S_i) \cup N(S_j)|}$$

    - S_i is a set of terms annotating a gene g_i
    - $N(S_i)$ is a set of all ancestor terms of the terms in S_i including the terms in S_i themselves

Survey of Semantic Similarity Measures (2)

- **Edge-based Methods**
  - Distance-based
    - Measuring distance (path length) between two terms

$$sim_{dist}(t_i, t_j) = \frac{1}{len(t_0,t_i)+len(t_0,t_j)+1}$$

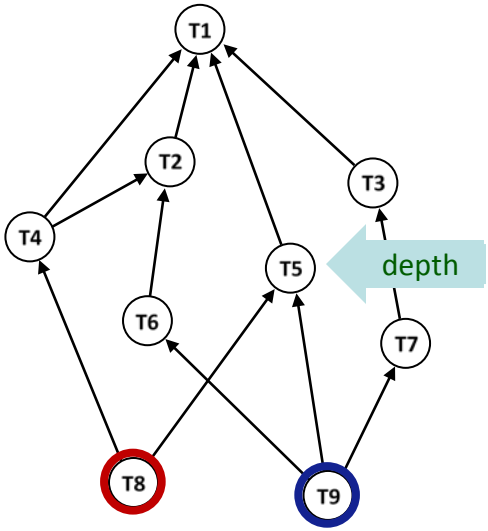
    - t_0 is the closest common ancestor term of t_i and t_j

  - Depth-based
    - Measuring specificity of the closest common term of two terms
    - Specificity of a term is defined as its depth

$$sim_{depth}(t_i, t_j) = \frac{2\times len(t_r,t_0)}{2\times len(t_r,t_0)+len(t_0,t_i)+len(t_0,t_j)}$$

    - t_r is the root term

  - Most of edge-based methods are pairwise



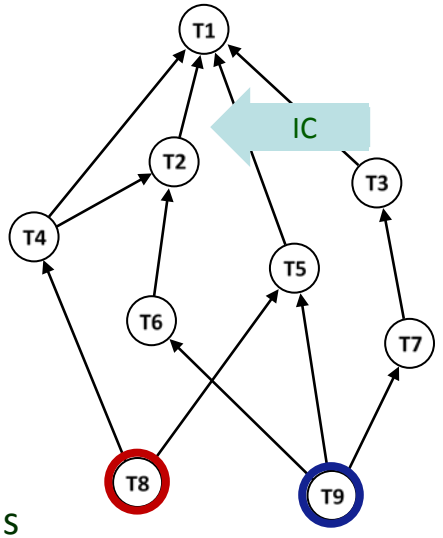
Survey of Semantic Similarity Measures (3)

- **Annotation-based Methods**
  - IC (Information Content)-based
    - Measuring specificity of the most informative common term of two terms
    - Specificity of a term is defined as its information content

$$sim_{ic}(t_i, t_j) = \max_{t_0 \in N(t_i) \cap N(t_j)}(-\log P(t_0))$$

$$sim_{ic-norm}(t_i, t_j) = \max_{t_0 \in N(t_i) \cap N(t_j)} \frac{2 \times \log P(t_0)}{\log P(t_i) + \log P(t_j)}$$

    - $P(t_i)$ is the ratio of annotating genes on the term t_i



  - IC-distance-based
    - Measuring information content distance between two terms

$$sim_{ic-dist}(t_i, t_j) = \max_{t_0 \in N(t_i) \cap N(t_j)} \frac{1}{2 \times \log P(t_0) - \log P(t_i) - \log P(t_j) + 1}$$

  - Most of annotation-based methods are pairwise

Survey of Semantic Similarity Measures (4)

- **Integrative Methods**
  - IC-based + IC-distance-based

$$sim_{icnd}(t_i, t_j) = \frac{-\log P(t_0)}{2\times\log(t_0)-\log P(t_i)-\log P(t_j)+1}$$

    - t_0 is the most informative common ancestor term of t_i and t_j
    - pairwise method

  - Common-terms-based (Node-based) + IC-based

$$sim_{gic}(S_i, S_j) = \frac{\sum_{t_m \in N(S_i) \cap N(S_j)} \log P(t_m)}{\sum_{t_n \in N(S_i) \cup N(S_j)} \log P(t_n)}$$

    - group-wise method

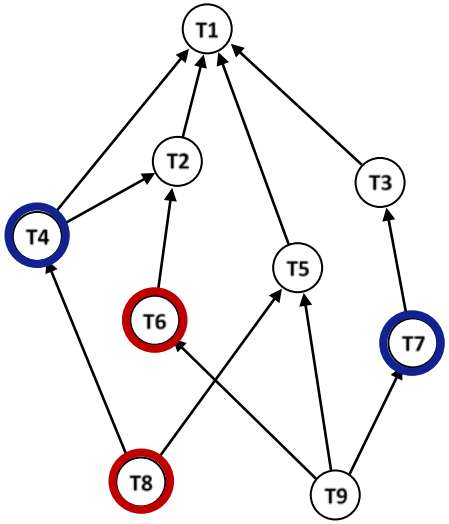
  - Common-term-based (Node-based) + Depth-based (Edge-based)

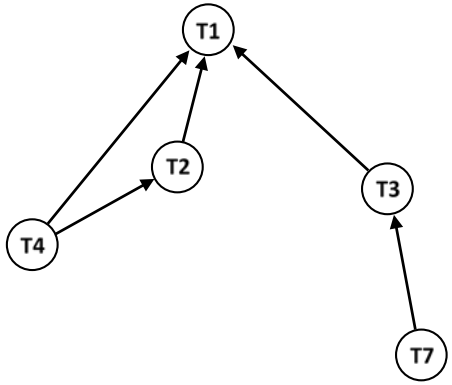
$$sim_{wtd}(S_i, S_j) = \frac{\sum_{t_m \in N(S_i) \cap N(S_j)} len(t_r, t_m)}{\sum_{t_n \in N(S_i) \cup N(S_j)} len(t_r, t_n)}$$

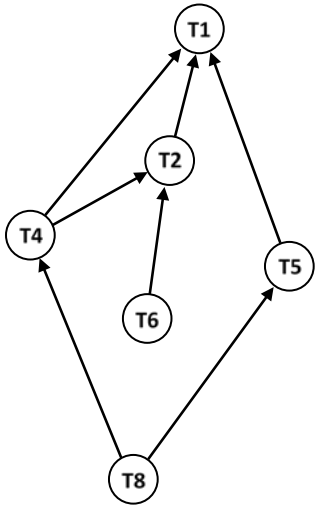
    - group-wise method

Proposed Semantic Similarity Measure (1)

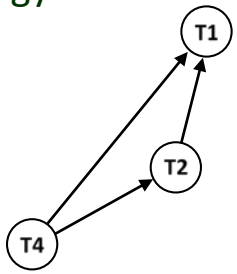
- **Graph-based Method**
  - Induced sub-ontology
    - Graph which consists of all terms in $N(S_i)$ and their directed edges





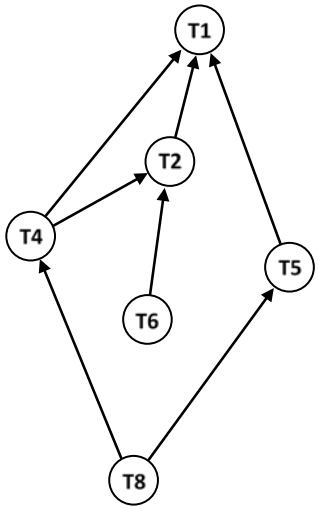


  - Common induced sub-ontology



Proposed Semantic Similarity Measure (2)

- **Vector-based Method**
  - Shortest-path vector
    - Dimensions: all directed shortest-paths in the directed graph
    - Values: IC (information content) distance



p(T2 → T1) ⟶ $\log P(T1) - \log P(T2)$
p(T4 → T1)
p(T5 → T1)
p(T6 → T1)
p(T8 → T1)
p(T4 → T2)
p(T6 → T2)
p(T8 → T2)
p(T8 → T4)
p(T8 → T5)

  - Magnitude of a shortest-path vector
    - Standard Euclidean norm

$$\mathcal{M}(G) = \sqrt{\sum_{\forall t_i, t_j \text{ s.t. } p(t_i \to t_j) \text{ in } G} (\log P(t_j) - \log P(t_i))^2}$$

Proposed Semantic Similarity Measure (3)

- **simVICD (similarity of vectors by information content distance)**
  - Magnitude of the common induced sub-ontology
    - Normalized by using geometric mean

$$sim_{vicd}(S_i, S_j) = \frac{\mathcal{M}(G_i \cap G_j)}{\sqrt{\mathcal{M}(G_i) \cdot \mathcal{M}(G_j)}}$$

    - G_i is the induced sub-ontology of the terms in S_i
    - $G_i \cap G_j$ is the common induced sub-ontology of G_i and G_j

- **Summary of the Proposed Method**
  - Combination of graph-based (node-based and edge-based), vector-based, and IC-based approaches
  - group-wise method

Experiment-1, Validation of PPIs

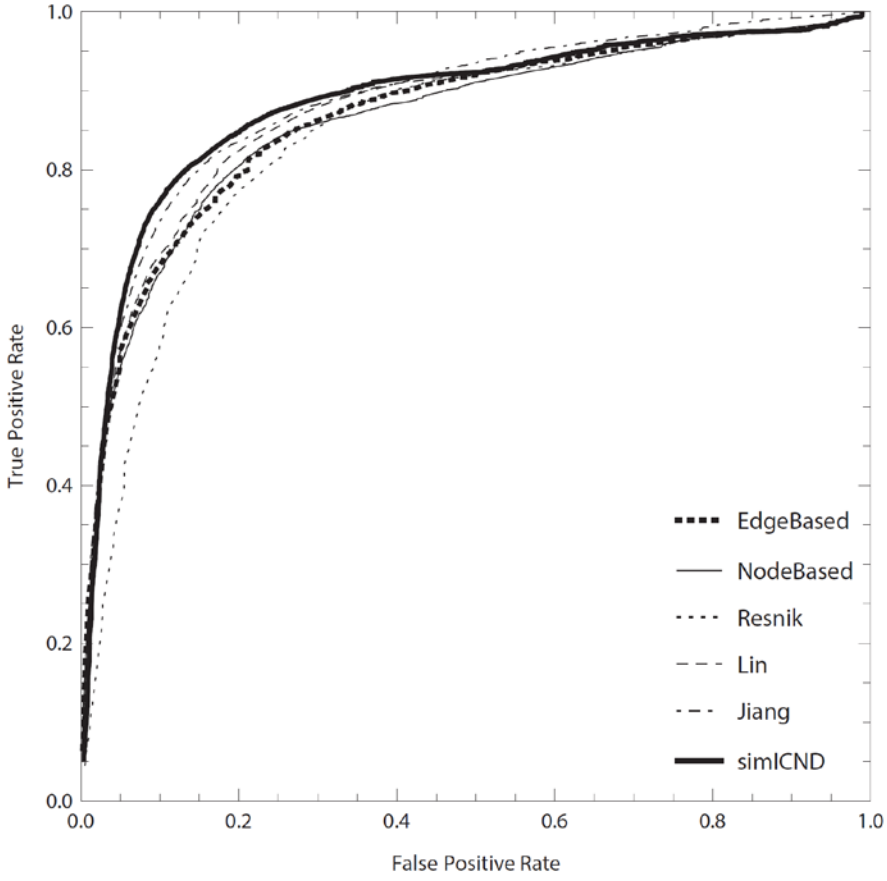
- **Validation Method**
  - Measuring semantic similarity of interacting pairs
  - Comparison between similarity scores of interacting pairs and protein complexes

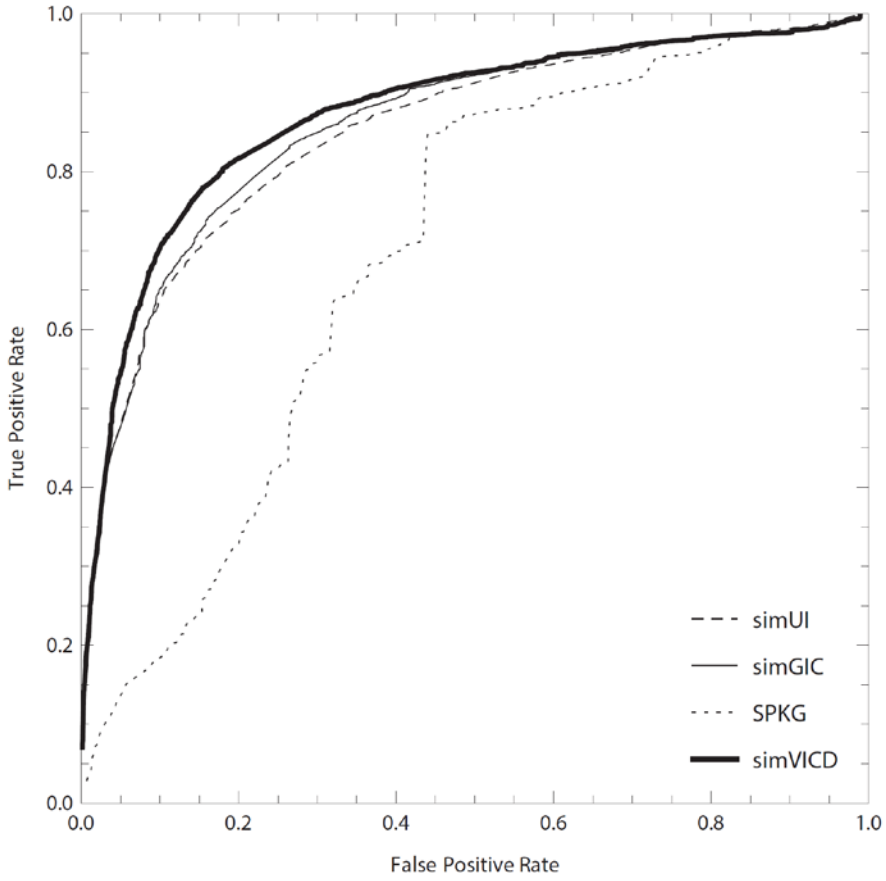
- **Data Source**
  - Yeast PPI data in a genomic scale from BioGRID
  - BP and MF ontologies and yeast annotation data from GO
  - Yeast protein complex data from Pu *et al*. 2009 (as reference data)

- **Experimental Setting**
  - Interacting pair which occur in the same protein complex → True data
  - Interacting pair which do not occur in the same protein complex → False data
  - Interacting pair which have higher semantic similarity than a threshold → Positive result
  - Interacting pair which have lower semantic similarity than a threshold → Negative result

Experiment-1, ROC Analysis (1)

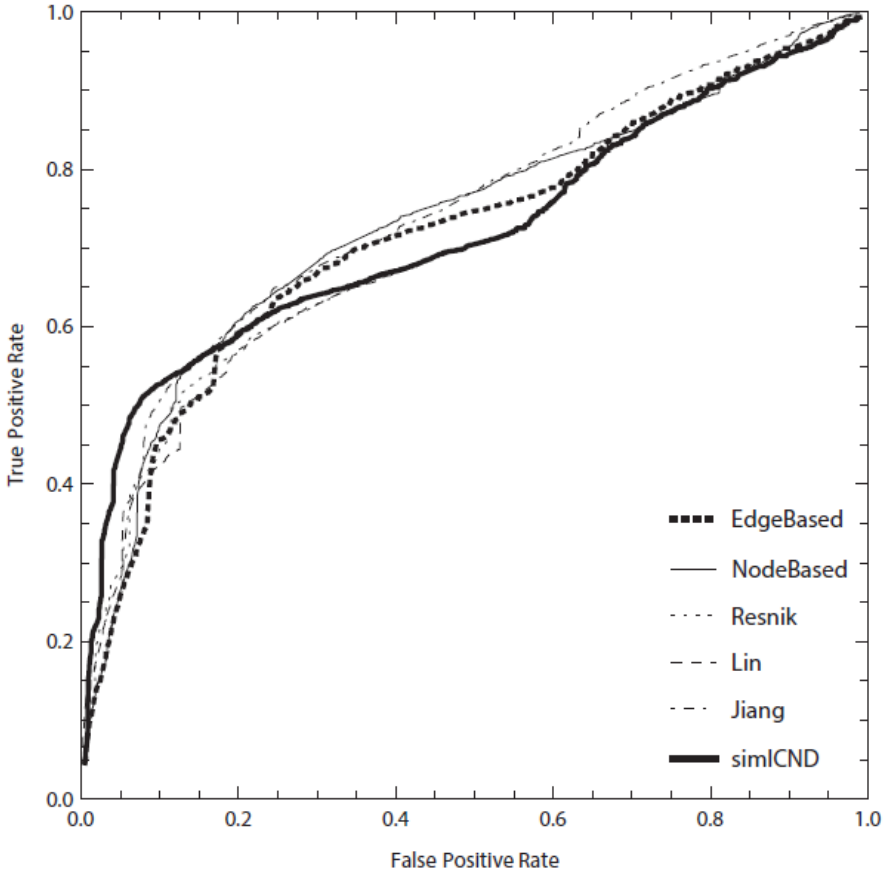
- **ROC Curves**
  - Comparison of semantic similarity measures in BP ontology

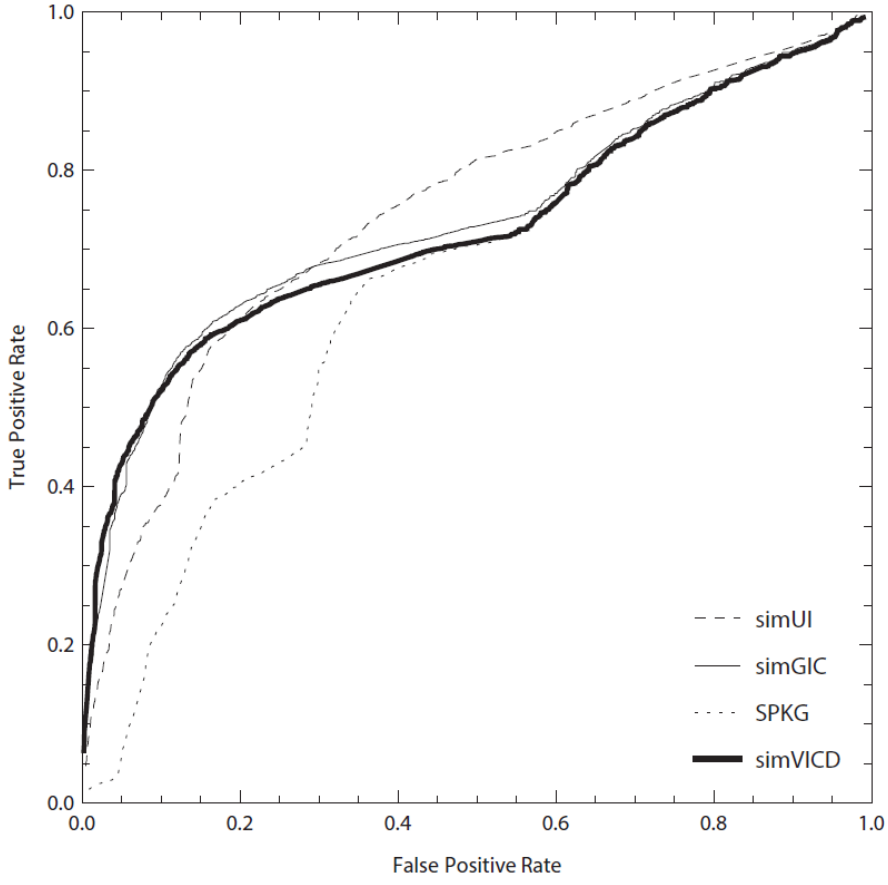




Experiment-1, ROC Analysis (2)

- **ROC Curves**
  - Comparison of semantic similarity measures in MF ontology





Experiment-1, Correlation Analysis

- **Pearson Correlation**
  - Reference scores by assigning 1 to true PPIs and 0 to false PPIs
  - Correlation between semantic similarity scores and reference scores across all PPIs

Ontology	Edge-based	Node-based	Resnik's	Lin's	Jiang's	simICND
BP	0.43	0.50	0.43	0.49	**0.58**	**0.58**
MF	0.31	0.34	0.32	0.34	0.36	**0.42**

Ontology	simUI	simGIC	spgk	simVICD
BP	0.52	0.55	0.16	**0.57**
MF	0.34	0.43	0.04	**0.46**

Experiment-2, Essential Protein Prediction

BAYLOR UNIVERSITY

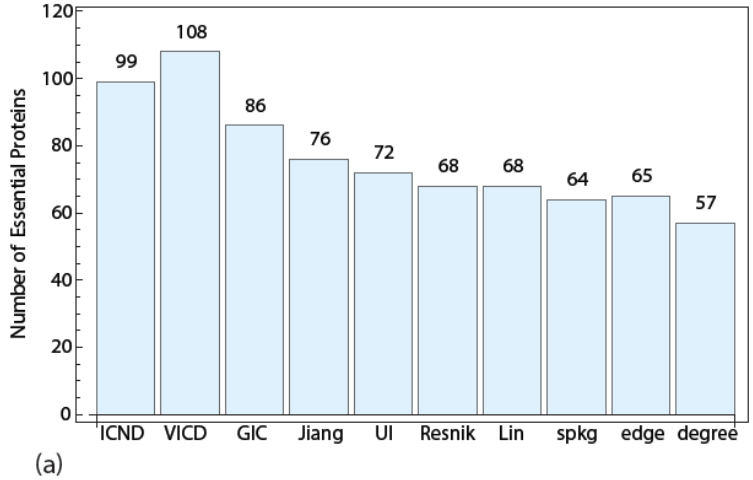
- **Prediction Method**
  - Predicting functionally essential proteins from PPI networks by graph centrality measures
  - Applying weighted degree by semantic similarity as centrality

- **Data Source**
  - Yeast PPI data in a genomic scale from BioGRID
  - BP and MF ontologies and yeast annotation data from GO
  - Yeast essential proteins from MIPS and DEG (as reference data)

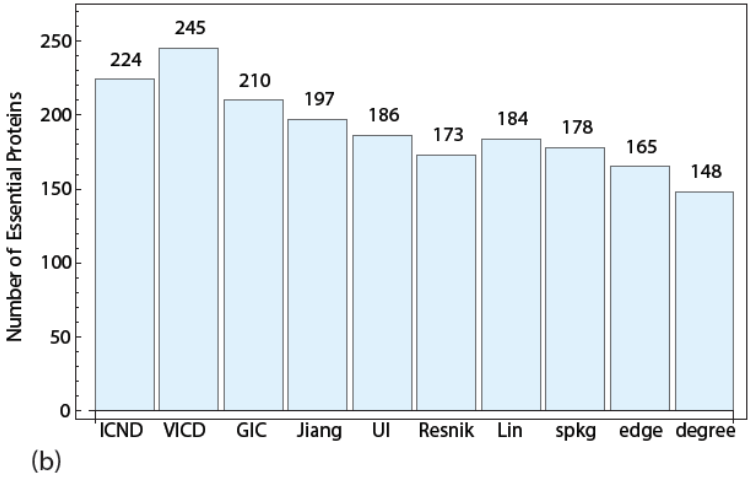
Experiment-2, Prediction Accuracy Comparison (1)

❑ Accuracy of Essential Protein Selection

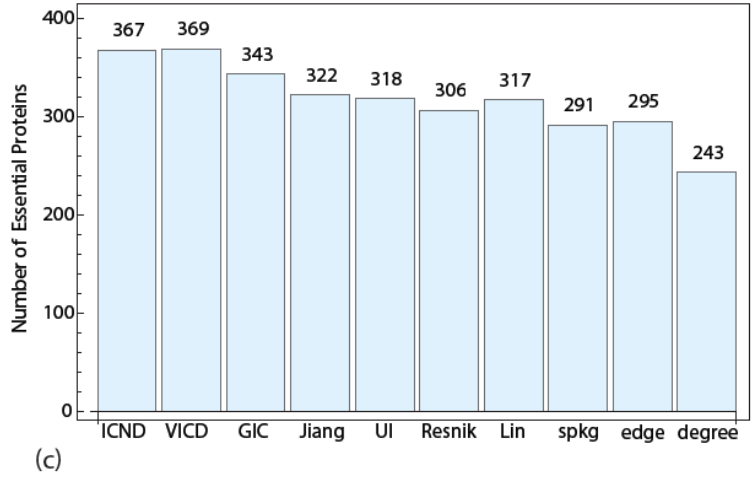
- Number of essential proteins selected from the ordered list of proteins by weighted degree



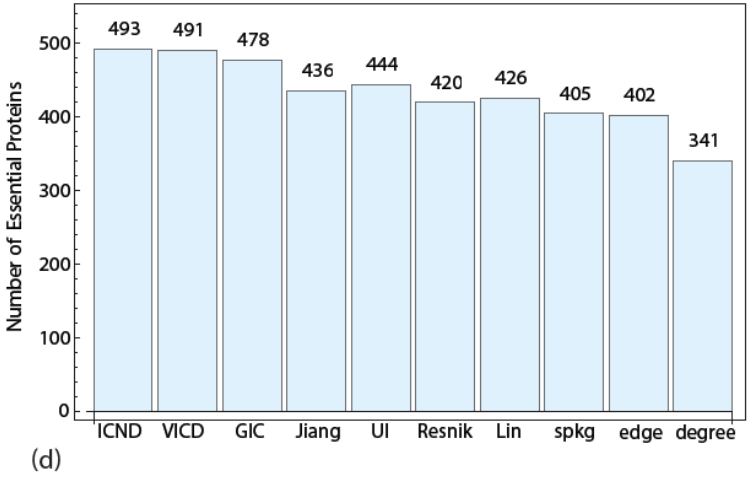
(a)



(b)



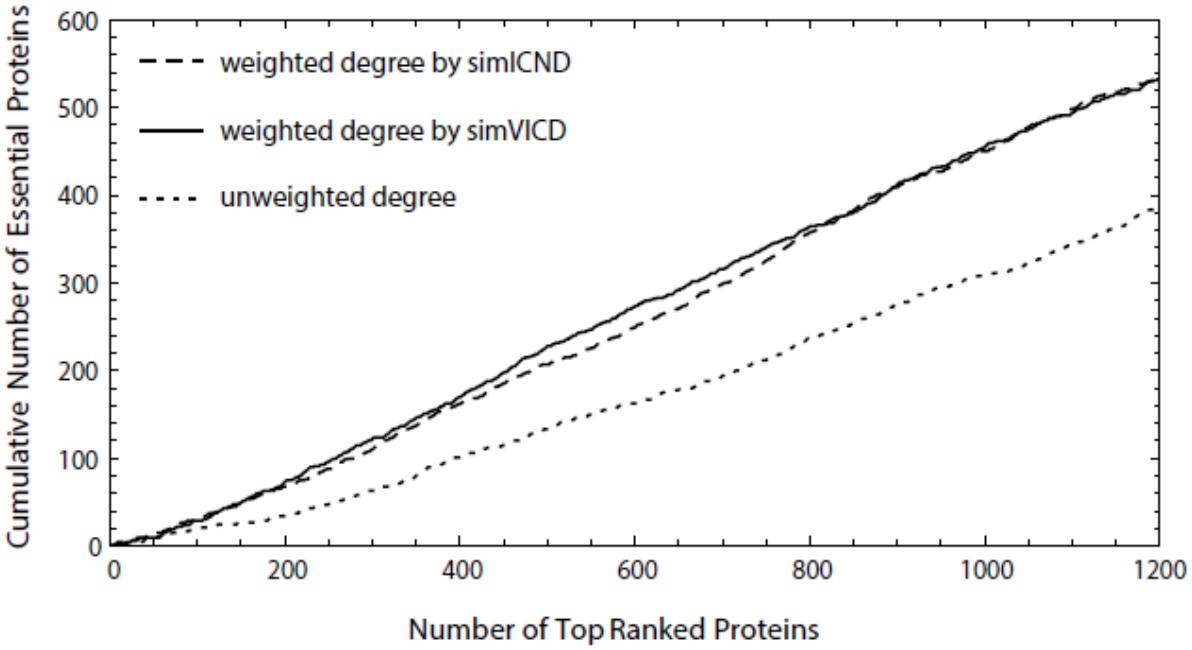
(c)



(d)

Experiment-2, Prediction Accuracy Comparison (2)

- **Accuracy of Essential Protein Selection**
  - Cumulative count of essential proteins selected from the ordered list of proteins by weighted degree

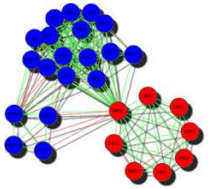


Conclusion

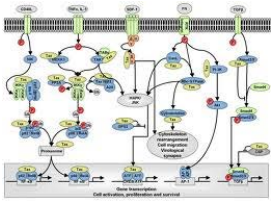
- **Summary**
  - Aim
    - Measuring functional similarity of interacting pairs by semantic similarity
  - Experimental results
    - Among pairwise methods,
      (1) simICND and (2) IC distance-based method were the most competitive
    - Among group-wise methods,
      (1) simVICD and (2) simGIC were the most competitive
  - Issues
    - Aggregation method for pairwise approaches
    - Informative annotation selection, and IEA exclusion
    - Application to the species of sparse annotation

Future Research Direction

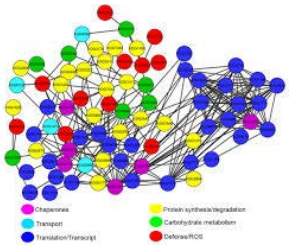
Weighting PPI Networks



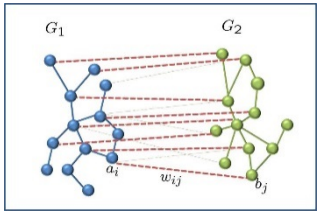
Protein complexes or Functional modules



Signaling pathways



Essential proteins



Network alignment

Acknowledgement

- This research was supported in part by funds from the Baylor University Research Committee and the Vice Provost for Research.

- Reference: http://bionet.ecs.baylor.edu/