M-Finder: Functional association mining from protein interaction networks weighted by semantic similarity

Young-Rae Cho, Ph.D.

Bioinformatics Program
Department of Computer Science
Baylor University

Backgrounds

- Omics-Data Management, Integration & Analysis
  - Genome, Proteome, Transcriptome, Metabolome, . . .
  - Interactome
    - Protein-protein interactions (PPIs) in the genomic scale
    - Represented as an undirected, unweighted network
    - Functional association patterns between proteins

- Computational Systems Biology
  - Computational approaches on the system-level

- Challenges
  - Unreliable data → Preprocessing
    - False positives & False negatives
  - Large scale → Scalability / Efficiency
  - Complex connectivity → Robustness / Accuracy

Approach

- Functional Associations
  - Connection patterns between functionally associated proteins
  - Sub-networks of the genome-wide PPI networks
  - Functional modules

- Functional Association Mining
  - Discovery of functional associations given a query protein

- Strategies
  (1) Weighting PPI networks
    - Integration of Gene Ontology and its annotation data
    - Computing semantic similarity for each PPI
  (2) Simulating functional propagation
    - Information propagation model
    - Implementing a dynamic propagation algorithm

Semantic Similarity

- Definition
  - A function that returns a numeric value reflecting closeness in meaning between ontological concepts (i.e., GO terms)
- Application
  - Measuring functional similarity between genes or proteins
- Categorization of Existing Semantic Similarity
  - **Edge-based** methods
    - Path length between GO terms, or normalized depth to SCA with depth to GO terms
  - **Node-based** methods
    - Ratio of common ancestor GO terms
  - **Annotation-based** methods
    - Information content (IC) of SCA, or normalized IC of SCA with ICs of GO terms
  - **Hybrid** methods
    - G-SESAME, simGIC, IntelliGO, . . .

Improvements of Semantic Similarity (1)

- Previous Experiments
  - Observed that IC of SCA (Resnik's method) has a good performance
  - Focuses on commonality, not difference of two GO terms

- *simICNP* (similarity by **I**nformation **C**ontent **N**ormalized with **P**ath length)
  - Normalization of the Resnik's method with the shortest path length between GO terms

$$sim_{ICNP}(C_1, C_2) = \frac{-\log P(C_0)}{1 + len(C_1, C_2)}$$

- *simICND* (similarity by **I**nformation **C**ontent **N**ormalized with **D**istance)
  - Normalization of the Resnik's method with the distance between the information contents of GO terms

$$sim_{ICND}(C_1, C_2) = \frac{-\log P(C_0)}{1 - \log P(C_1) - \log P(C_2) + 2 \cdot \log P(C_0)}$$

Improvements of Semantic Similarity (2)

- *simICS* (similarity by **I**nformation **C**ontent of annotation **S**et)
  - $\alpha |\Gamma_{C0}| + \beta(|\Gamma_{C0}| - |\Gamma_{C1} \cap \Gamma_{C2}|) - \gamma |\Gamma_{C1} \cap \Gamma_{C2}|$

    where $|\Gamma_{Ci}|$ denotes the size of the set of proteins annotating to the GO term C_i
  - $2|\Gamma_{C0}| + |\Gamma_{C0}| - |\Gamma_{C1} \cap \Gamma_{C2}| - |\Gamma_{C1} \cap \Gamma_{C2}| = 3|\Gamma_{C0}| - |\Gamma_{C1}| - |\Gamma_{C2}|$

$$sim_{ICS}(C_1, C_2) = -\log \frac{3|\Gamma_{C_0}| - |\Gamma_{C_1}| - |\Gamma_{C_2}|}{|\Gamma_{C_r}|}$$

- Functional Similarity
  - Best-match averaging method

$$sim_{BMA}(g_1, g_2) = \frac{\sum_{C_1 \in S_1} \max_{C_2 \in S_2} sim(C_1, C_2) + \sum_{C_2 \in S_2} \max_{C_1 \in S_1} sim(C_1, C_2)}{|S_1| + |S_2|}$$

Information Propagation Model

- Path Strength
  - $$S(p) = \lambda \cdot w(v_0, v_1) \prod_{i=1}^{n-1} \frac{w(v_i, v_{i+1})}{d_i}$$
    - $p = \langle v_0, v_1, \ldots, v_n \rangle$
    - λ : a scale parameter
    - $w(v_i, v_{i+1})$: a weight between v_i and v_{i+1} where $0 < w(v_i, v_{i+1}) \leq 1$
    - d_i : a shape parameter (degree of v_i)
  - Factors
    - edge weights
    - node degree
    - path length

- Functional Association between two proteins
  - Sum of strength of all possible paths (including cycles) between two nodes in the PPI network

Dynamic Functional Propagation Algorithm (1)

- Path Strength Computation
  - Dynamic programming

$$S(\langle v_s, v_1, \cdots, v_n \rangle) = S(\langle v_s, v_1, \cdots, v_{n-1} \rangle) \cdot \frac{w(v_{n-1}, v_n)}{d_{n-1}}$$

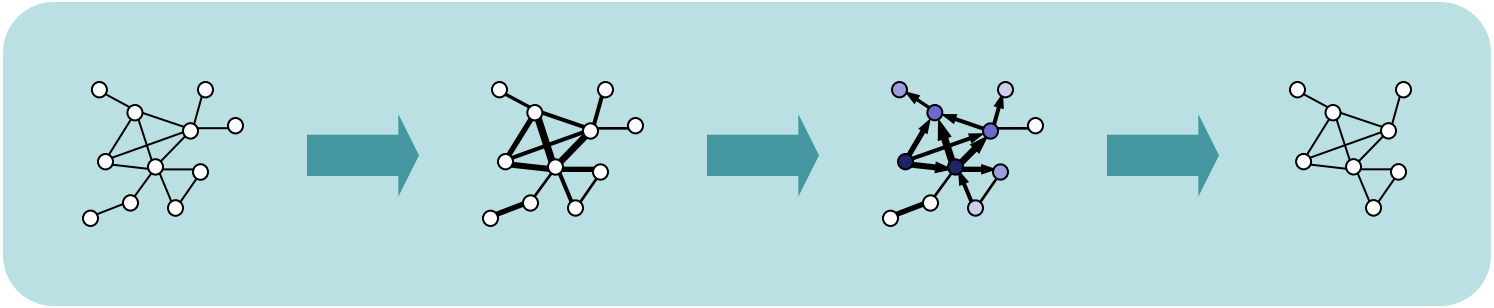
- Functional Flow Computation
  - Sum of strength of the paths in a fixed length
  - Dynamic programming

$$F_s(v_i) = \sum_{v_j \in N(v_i)} F_s(v_j) \cdot \frac{w(v_j, v_i)}{d_j}$$

- Functional Association Computation
  - Sum of strength of the paths in any length
  - Cumulative score of functional flow requires iterative computation of functional flow

Dynamic Functional Propagation Algorithm (2)

- Algorithm
  - (1) Assign an initial quantity to the query protein.
  - (2) Compute $F_s(v_i)$ for all proteins.
    (If $F_s(v_i)$ for any protein is less than a threshold, then switch it to 0.)
  - (3) Compute the cumulative $F_s(v_i)$ for all proteins.
  - (4) Repeat step (2) and (3) until no change in the cumulative $F_s(v_i)$.
  - (5) Return a sub-network consisting of the proteins whose cumulative $F_s(v_i)$ is greater than a threshold.



Evaluation of Semantic Similarity

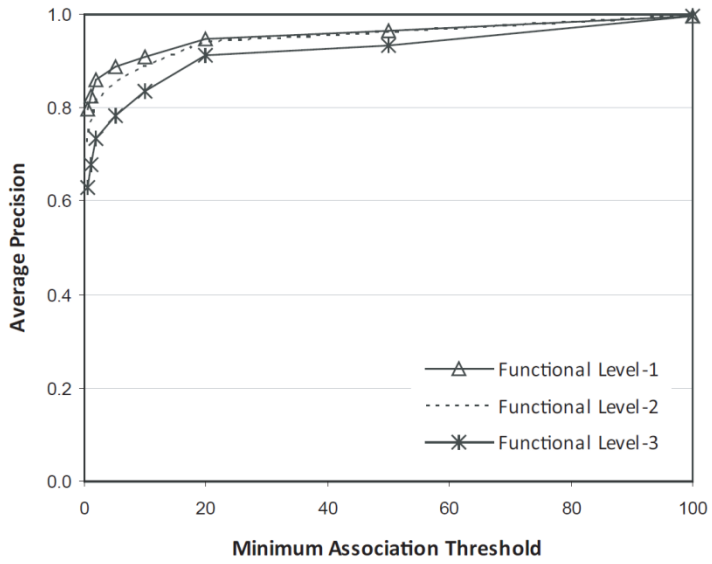
- PPI Data
  - Yeast PPI data from BioGRID
- Functional Consistency
  - Manually curated small-scale functional annotation data from FunCat
  - Compute the Jaccard coefficient of two sets of functions
- Correlation
  - Compute the Pearson coefficient between semantic similarity and functional consistency
- Comparison

Category	Semantic Similarity	Correlation
Edge-based	path length	0.699
	normalized depth	0.614
Node-based	NTO	0.571
	simUI	0.679
Annotation-based	Resnik	0.646
	Lin	0.663
	Jiang	0.716
Hybrid	G-SESAME	0.667
	simGIC	0.688
New Approaches	simICNP	**0.726**
	simICND	**0.733**
	simICS	0.678

Accuracy of Functional Associations

- Precision
  - Manually curated small-scale functional annotation data from FunCat
  - X : the set of proteins generated given a query protein
  - P_i : the set of proteins annotating to a functional category which has the query protein
  - Precision = (X ∩ P_i) / X

- Results

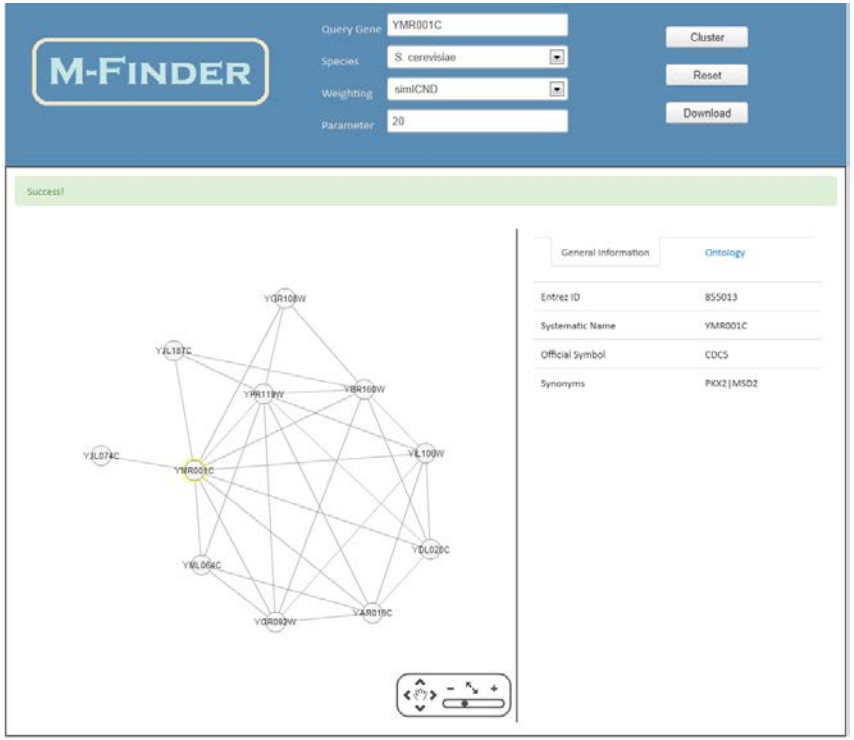


- Comparison

Method	Avg. Size	Avg. Precision
M-Finder	39.56	0.78
MCODE	40.71	0.43
DPClus	5.28	0.68
IPCA	47.24	0.52

M-Finder

- M-Finder
  - Interactive web application
  - To visualize the functional association pattern of a query protein
  - To find additional functional information of proteins and interaction information
  - http://bionet.ecs.baylor.edu/mfinder/



Conclusion

- Significance
  - Integration of interactome data with GO annotation data
  - Systematic approach to analyze interactome data
    - Information propagation model to formulate functional impact
    - Dynamic propagation algorithm for efficient computation
  - Application tool

- Future Works
  - Management of complete weighted PPI data across model species
  - Additional functional information, e.g., homologs, protein domain, . . .

- Acknowledgement
  - Supported by the grant from Young Investigator Development Program (YIDP) Award by Baylor University