Ranking for Phenotype Relevance Based on the Diffusion State Distance

Yuelin Liu*, Sam Slate*, Jisoo Park, Lenore Cowen

Department of Computer Science, Tufts University, Medford, MA

Tufts UNIVERSITY

Department of Computer Science

MOTIVATION

The **Diffusion State Distance metric (DSD)**, introduced by Cao et al. in 2013[1], has been shown to improve existing methods for function prediction and community detection in protein-protein interaction networks.

Diffusion-based methods have also been used extensively in **gene prioritization**, where the goal is, given a set of genes known to be associated a disease or phenotype, to use network information to rank a set of candidate genes by their likelihood to share relevance with the disease or phenotype.

Therefore, we asked whether DSD-based methods were superior to standard ranking methods such as **Random Walk with Restart (RWR)**[2] for this problem, as well.

DATASETS

Species: S. cerevisiae (Baker's yeast)

PPI Network: STRINGdb[3] (v10.5)

Phenotype-gene Associations:

1) MIPS FunCat[4]
2) Gene Ontology[5]

METHODS

Given a set of seed genes known to be associated with a phenotype, a gene from the candidate set can be ranked by:

1. Average RWR from Seed Genes

RWR computes the probability of reaching another gene from a target gene via a random walk with a certain restart probability. Candidate genes with high average RWR to seed genes were ranked higher.

2. Average DSD from Seed Genes

Genes with low average DSD to all seed genes were ranked higher.

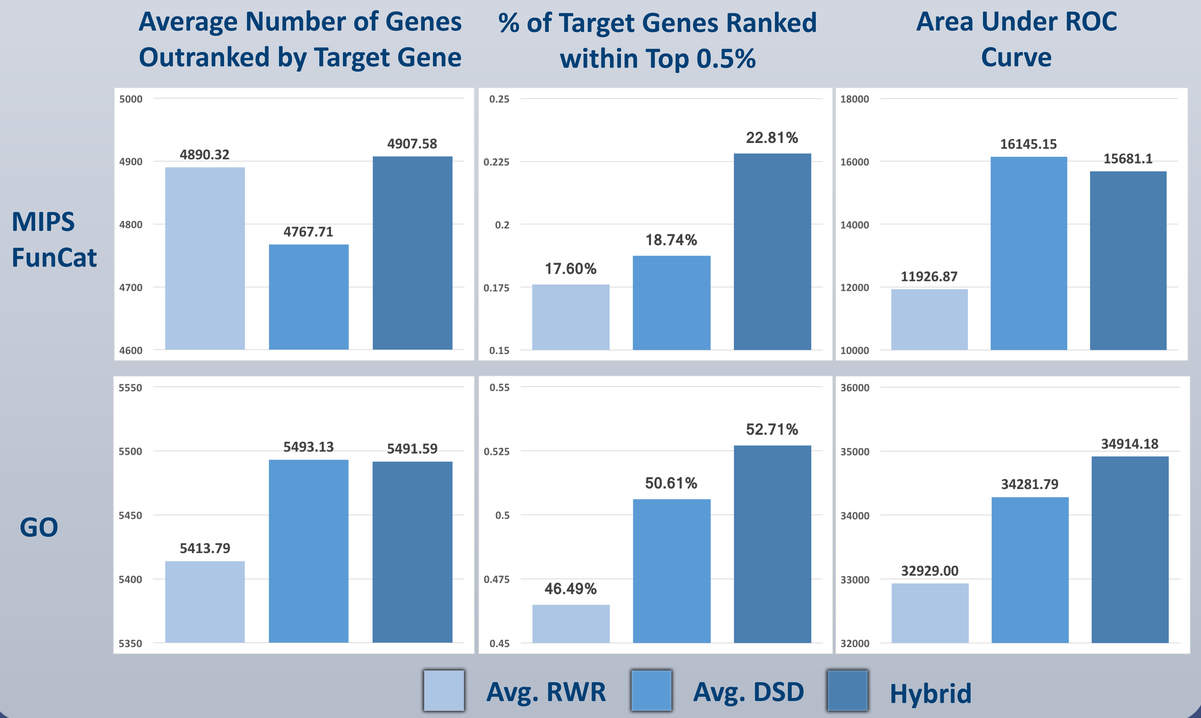
3. Hybrid Score

The hybrid score of a candidate gene was the average of its ranks obtained using the previous 2 methods.

We performed **leave-one-out cross validation** on each of the ranking methods: we considered the rank of a target gene with all other genes of its phenotype as the seed set.

RESULTS

Results reported were computed with RWR with **restart probability = 0.9** and **DSD step = 1**. Higher values indicate better performance.



CONCLUSION

Although DSD-based methods did not consistently outperform RWR-based methods, hybrid algorithms that ranked based on the average of RWR and DSD rankings slightly outperformed either method alone.

FUTURE WORK

Moving forward, we are interested in further investigating the following:

1. Testing our hypothesis on human PPI networks and gene-disease phenotype association data
2. Exploring new methods to compute the distance from one node to a set of nodes in a network
3. Exploring ways of incorporating DSD-based clustering in solving gene prioritization problem

REFERENCES

[1] M. Cao et al., *ISMB 2014 Proceedings*, 30 (2014): i219-i227

[2] Köhler et al. *The American Journal of Human Genetics* 82.4 (2008): 949-958.

[3] D. Szklarczyket al., *Nucleic Acids Research*, 39 (2011): D561–D568

[4] A. Ruepp et al., *Nucleic Acids Research*, 32 (2004): 5539-5545 (Accessed: 8/3/17)

[5] The Gene Ontology Consortium. *Nucleic Acids Research*, 43 (2015): D1049–D1056. (Accessed: 7/14/17)