Graph-based Information Diffusion Method for Prioritizing Functionally Related Genes in Protein-Protein Interaction Networks

Minh Pham and Olivier Lichtarge

Department of Molecular and Human Genetics

Baylor College of Medicine

Houston, Texas, USA

Motivation

Predictions (●)

Gene A

Gene B

Gene C

…

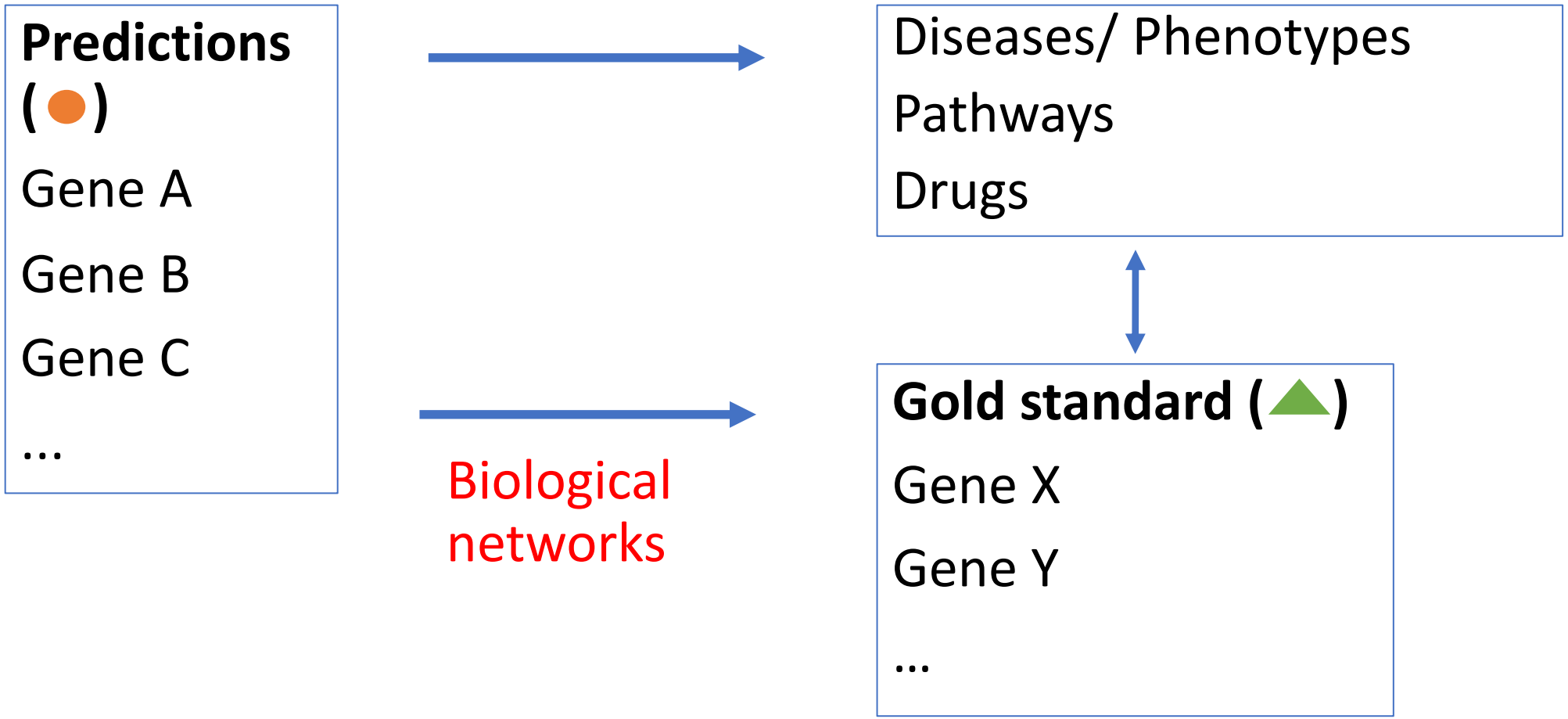
How do we validate whether they are related?

Diseases/ Phenotypes

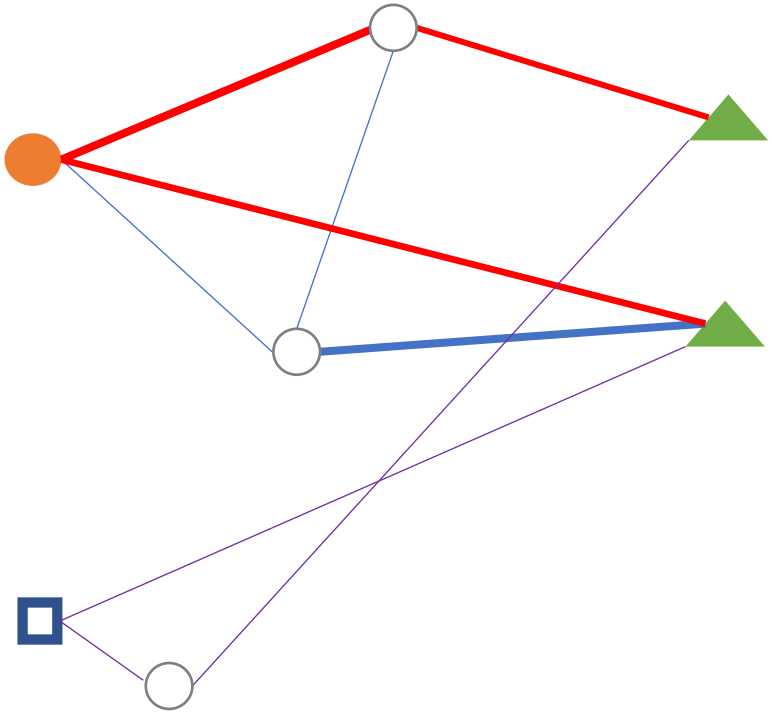
Pathways

Drugs

Motivation


Predictions (●)
Gene A
Gene B
Gene C
…
Biological networks
Diseases/ Phenotypes
Pathways
Drugs
Gold standard (▲)
Gene X
Gene Y
…

Network analysis to validate functionally related genes using biological network info



- Shortest path length (SPL) methods [1] are routinely used [2]

● and □ are equally connected to ▲

Average non-weighted SPL:

● → ▲ : (2+1)/2 = 1.5

□ → ▲ : (2+1)/2 = 1.5

▲ Gold standard

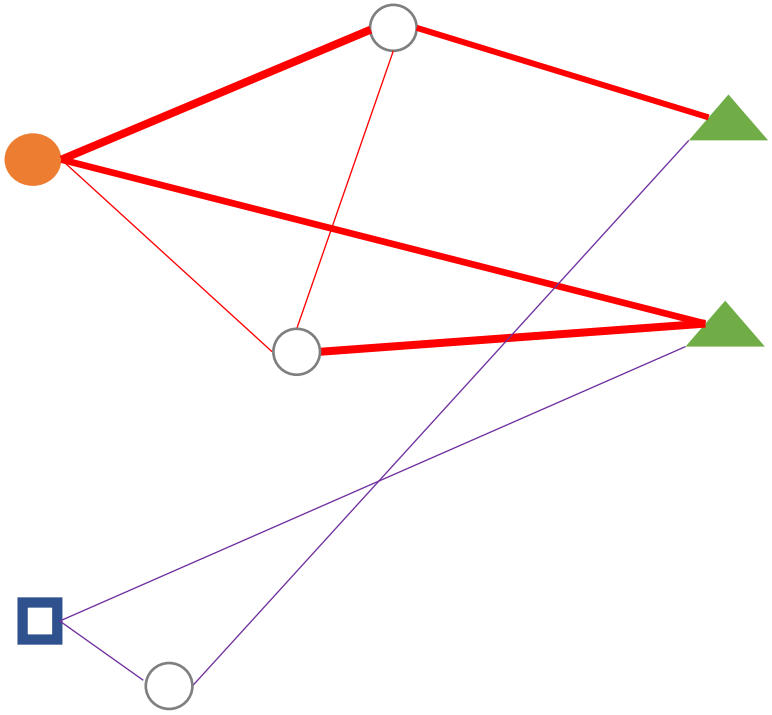
● Predictions

○ Other □ Random

[1] Dijkstra, **Numer Math** 1959

[2] Al-Ramahi *et al,* **Cell Syst.** 2018

Pitfalls of the shortest path length methods



Gold standard

Predictions

Other

Random

Problems with SPL:

- Do not consider multiple connected paths
- Computationally intensive

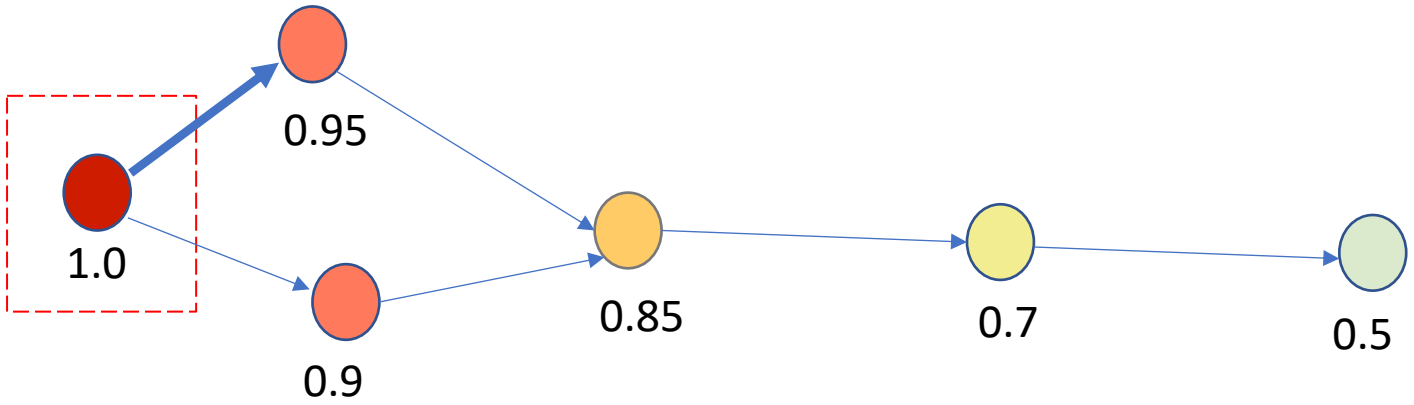
→ take a lot of time to compute [1]

- Integrating the edge weight makes it more computationally complicated

Graph-based information diffusion offers a solution

[1] Fredman *et al.*, **IEEE** 1984

Graph-based information diffusion approach [1]



- Efficiently combines both number of steps and edge weight
  → utilizing pathway interpretability and informational confidence
- Has extracted meaningful information in biological networks [2, 3]

Hypothesis: Diffusion method can prioritize genes with similar functions in biological networks better than SPL methods

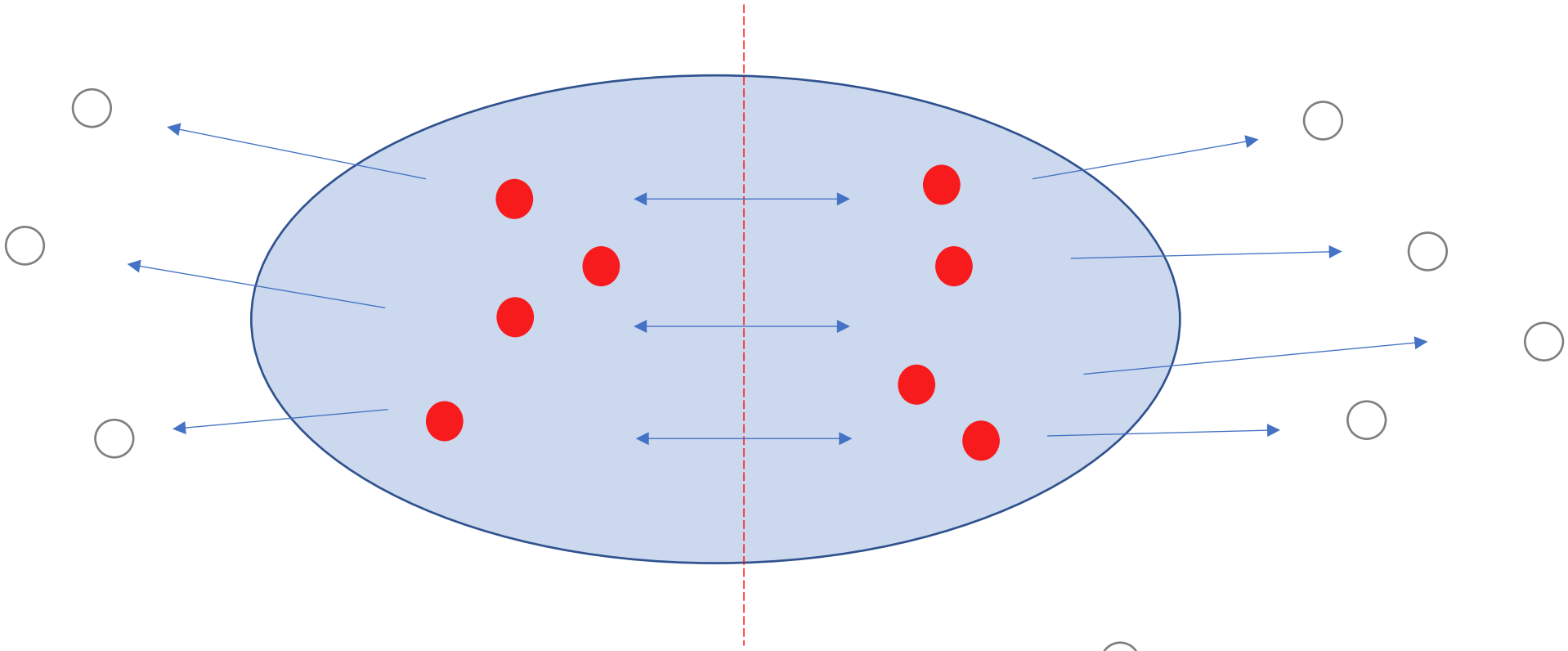
[1] Lisewski & Lichtarge, **Physica A** 2010
[2] Lisewski *et al.*, **Cell** 2014
[3] Venner *et al.*, **PLoS One** 2010

Questions

- Can diffusion method prioritize genes of same genetic/molecular processes?
- Is performance of diffusion method comparable to shortest path length methods?
- Can diffusion method prioritize genes associated with clinical phenotypes and do that better than SPL?

Hypothesis: Genes in the same pathways diffuse to each other more than to random genes in the network

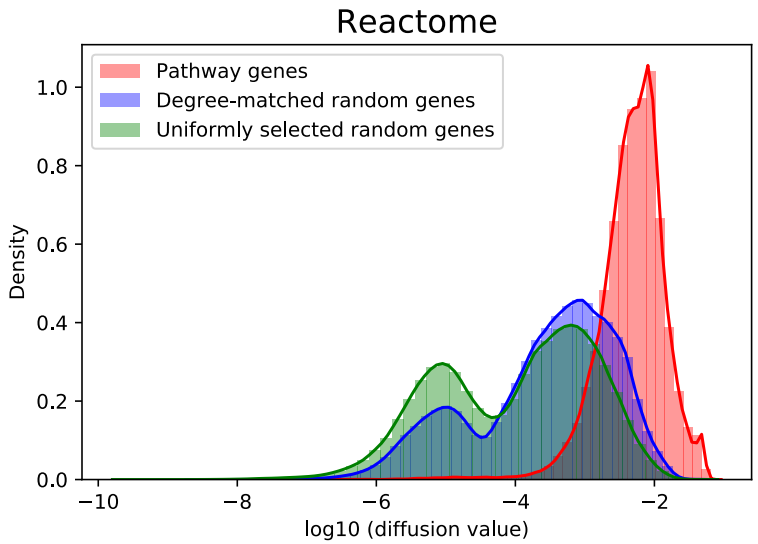
Methods: Genes in a pathway were split in half. Diffuse from one half to the other throughout a biological network (STRING PPI network)

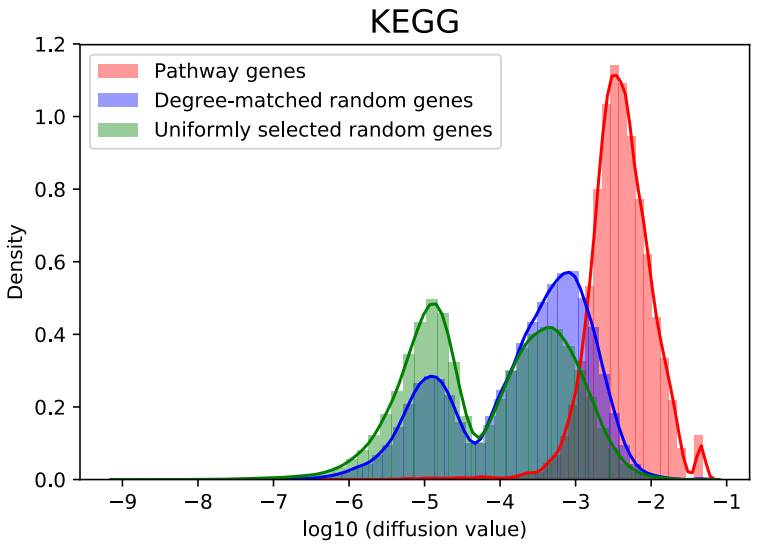


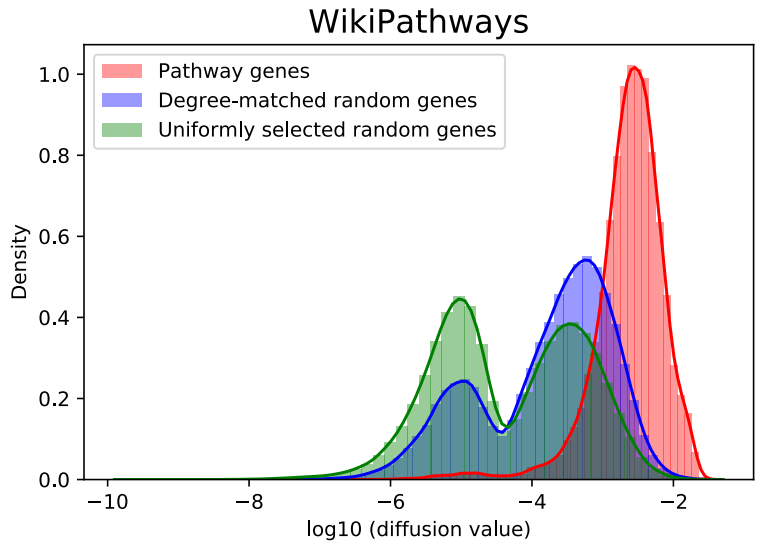
Other genes in the network

Genes in the same pathways diffuse to each other more than random ($p < 0.0001$)

Methods: Genes in a pathway were split in half. Diffuse from one half to the other

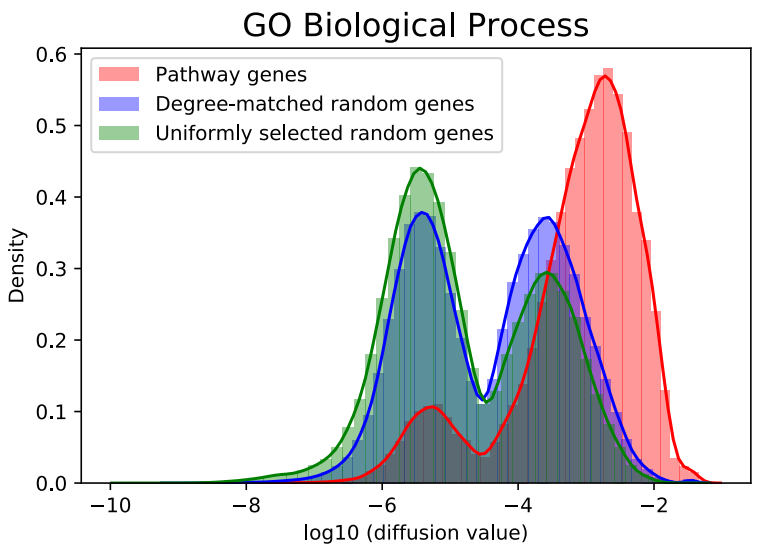


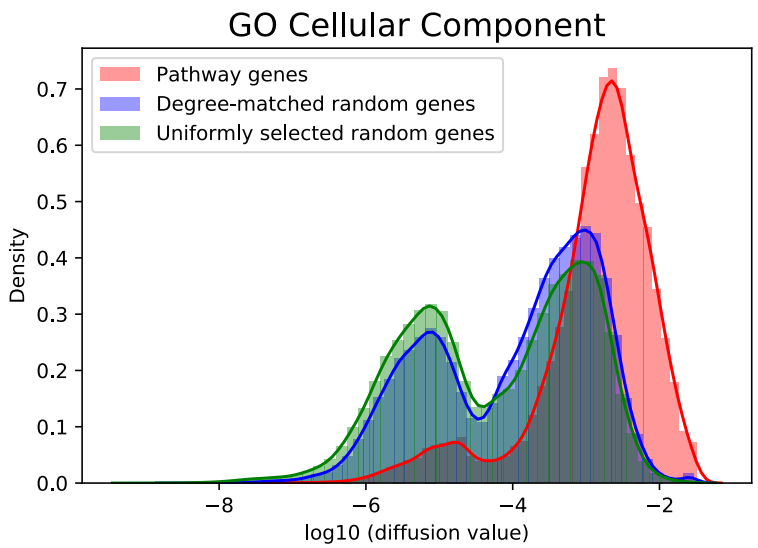


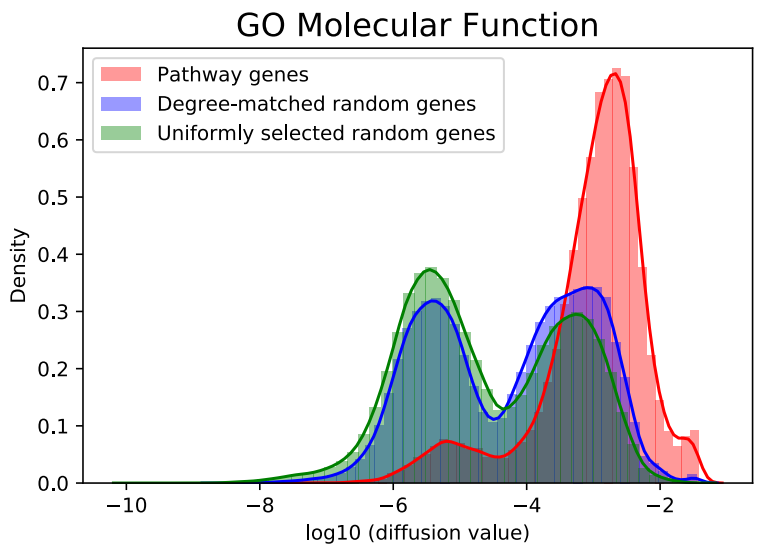


Genes in the same ontologies diffuse to each other more than random ($p < 0.0001$)

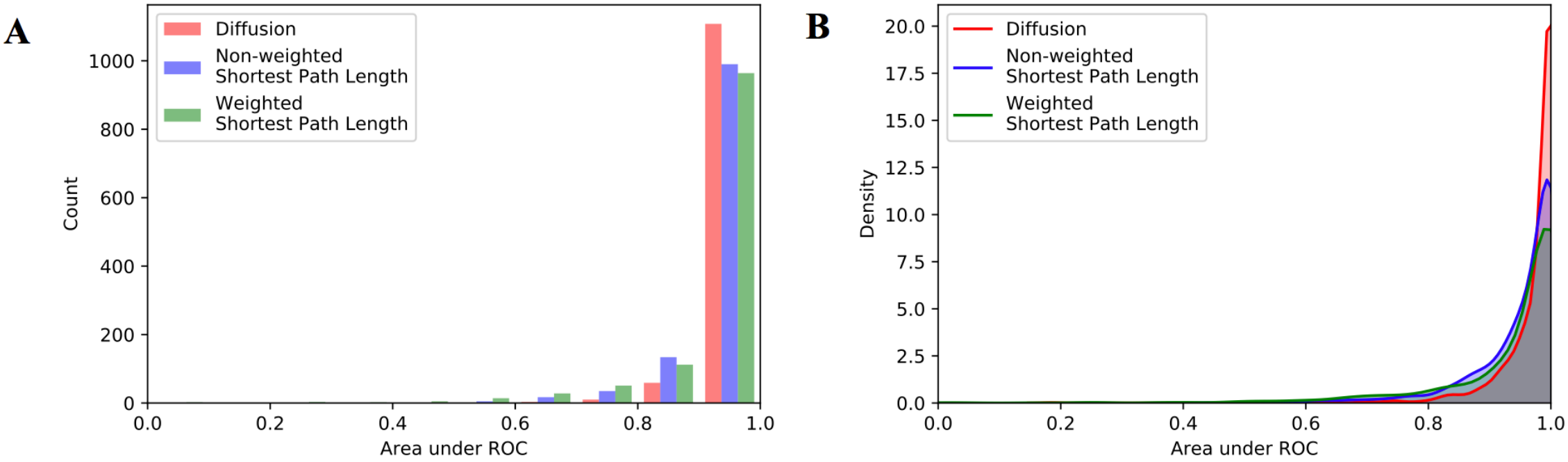
Methods: Genes in an ontology were split in half. Diffuse from one half to the other







Diffusion can detect genes in the same pathways better than shortest path length (SPL) methods



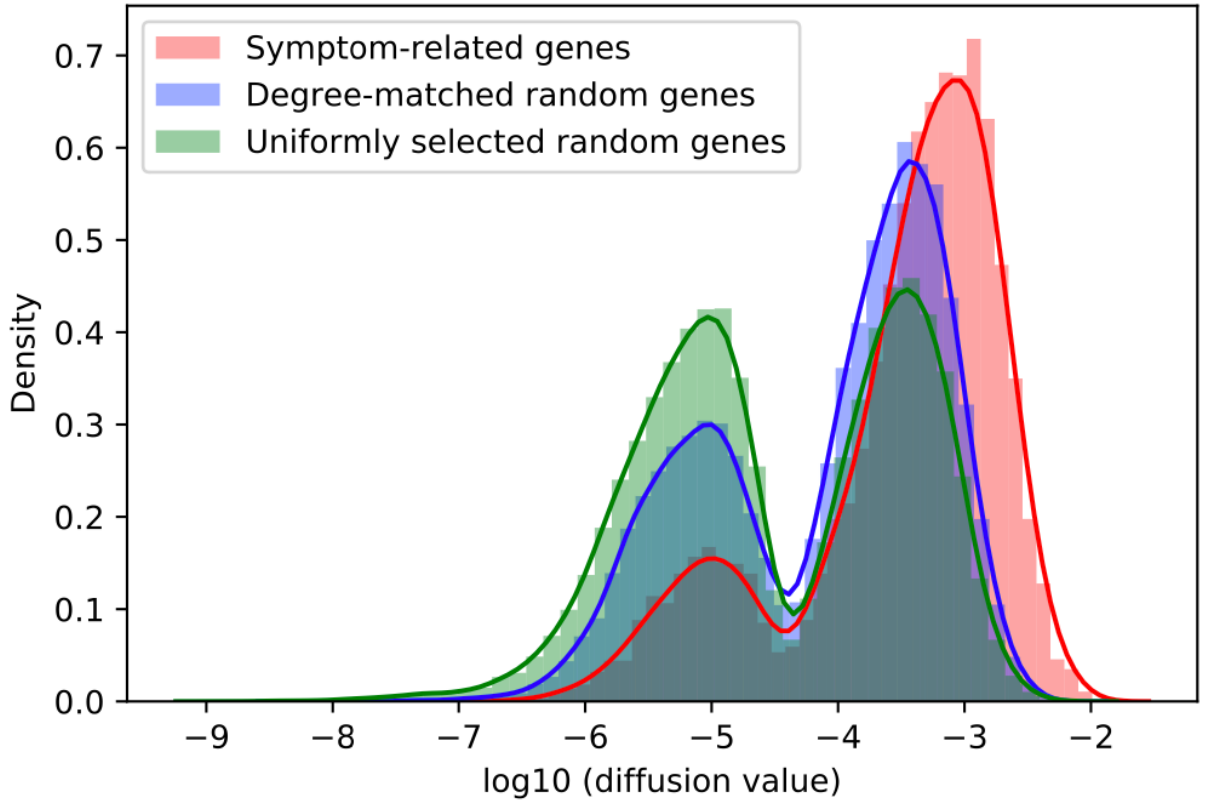
(KS test: $p_{\text{diffusion vs non-weighted SPL}} = 2.7e\text{-}28$, $p_{\text{diffusion vs weighted SPL}} = 2.8e\text{-}11$, $p_{\text{non-weighted vs weighted SPL}} = 2.7e\text{-}10$).

Genes associated with clinical phenotypes diffuse to each other more than random

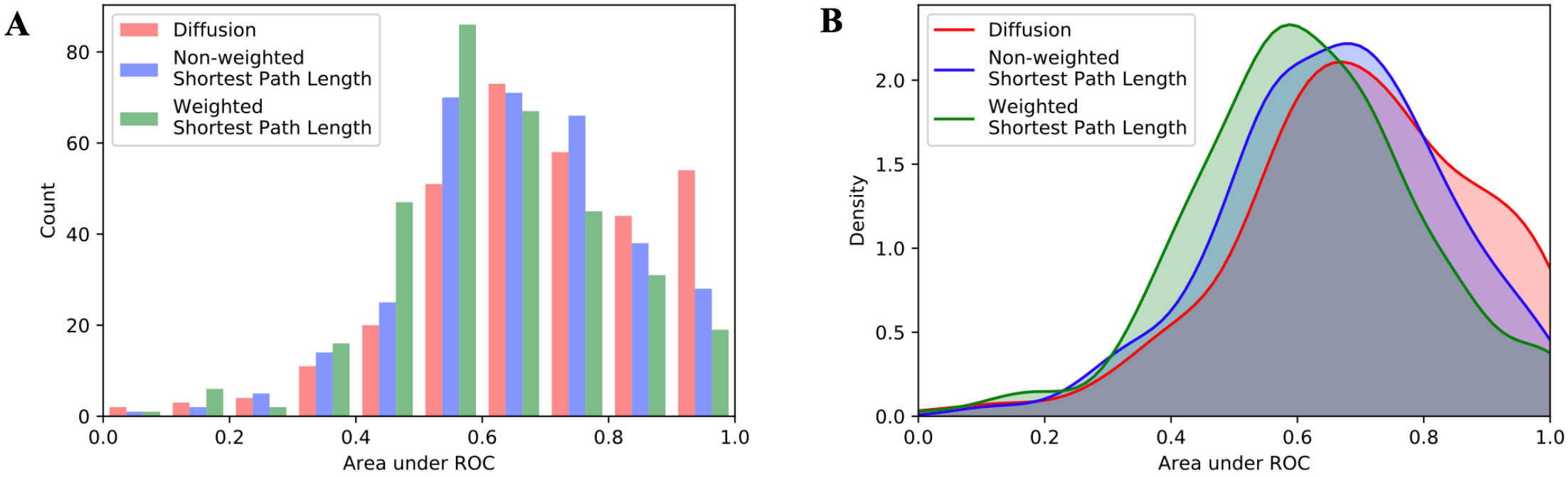
Gene KO = human-like clinical symptoms in mice

e.g. “parotid gland inflammation” and “joint swelling”

(from *Mouse Genomics Informatics* database)



Diffusion method can detect genes associated with clinical symptoms better than SPL methods



(KS test: $p_{\text{diffusion vs non-weighted SPL}} = 0.032$, $p_{\text{diffusion vs weighted SPL}} = 5.1\text{e-}07$, $p_{\text{non-weighted vs weighted SPL}} = 3.1\text{e-}03$).

Conclusions

- Diffusion method prioritized ***pathway-***, ***ontology-***, and ***clinical phenotype-*** specific genes more robustly than the shortest path length methods

- Diffusion method should be used routinely to validate, prioritize, and predict functionally related genes

MAHALO !!!

Baylor College of Medicine
Lichtarge Lab