uKIN Combines New and Prior Information with Guided Network Propagation to Accurately Identify Disease Genes

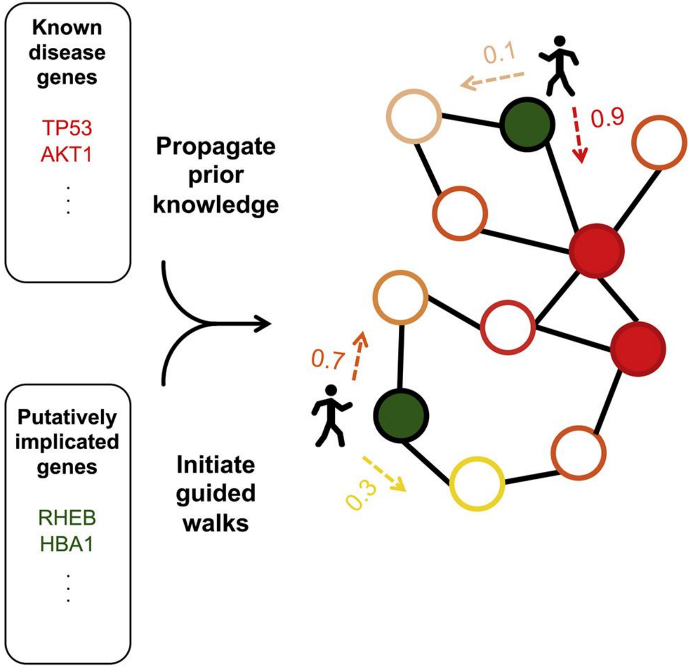
Borislav H Hristov , Bernard Chazelle , Mona Singh

CSE 590C - 11/23/20
Alyssa La Fleur

uKIN (using knowledge in networks)

Highlights:

- Guided network propagation method for discovery of disease-relevant genes
- Uses known disease genes to guide random walks initiated at newly implicated genes
- The guided walks allow for network-based integration of prior and new data
- Effectiveness of method shown on cancer genomics and genome-wide association data



Background: biological networks

- Large amount of variant data now available for healthy and disease genomes, but understanding the genetic basis underlying complex human diseases is difficult
- Biological networks provide a framework for identifying disease genes:
  - Disease genes tend to cluster in networks
  - If some genes are known to be causal for a disease, nearby genes in the network could also be disease relevant
- Two dominant network propagation techniques to uncover more disease genes
  1. Spreading signal from well-established, annotated genes
  2. Spreading signal from genes with new evidence of being disease relevant

Background: Random walks with restarts

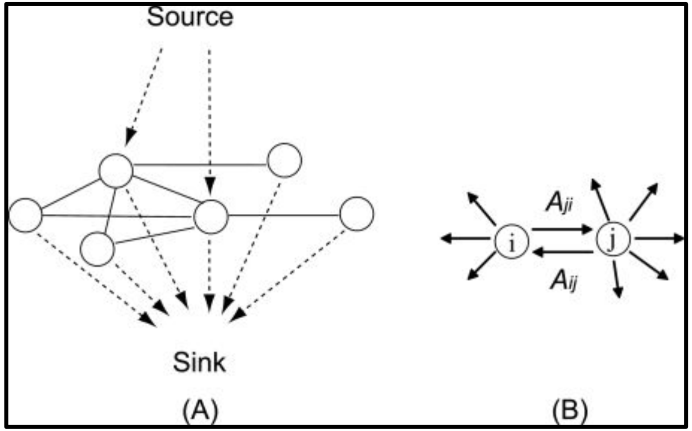
$$\mathbf{p}^{t+1} = (1-r)\mathbf{W}\mathbf{p}^{t} + r\mathbf{p}^{0}$$

- $\mathbf{p}^t$:vector where i-th element is the probability of being at node i at time step t
- $\mathbf{p}^0$:start probability vector
- r :restart probability
- $\mathbf{W}$:Column normalized adjacency matrix of the graph

Kohler, S.; Baur, S.; Horn, D.; Robinson, P.N. Walking the Interactome for Prioritization of Candidate Disease Genes. *AJHG,* **2008,** *82(4):* 949-958.

Background: Diffusion & diffusion kernels

- A ‘fluid’ is pumped into the graph to an initial set of nodes
- Fluid spreads over the edges of the graph
- Fluid is allowed to leak out from each node to a sink



$$\dot{p}_i(t) = \sum_j A_{ij}p_j(t) - \{\gamma + \sum_j A_{ji}\}p_i(t) + b_i u(t),$$

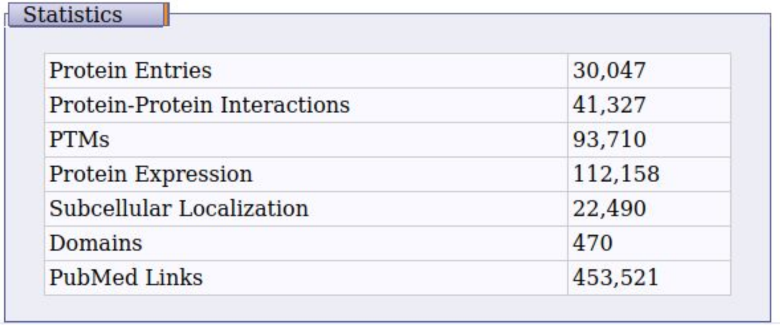
$$\dot{\vec{p}}(t) = (\mathbf{A} - \mathbf{S} - \gamma\mathbf{I})\vec{p}(t) + \vec{b}u(t) \qquad \mathbf{L} = -(\mathbf{A} - \mathbf{S} - \gamma\mathbf{I})$$

$$\vec{p}(t) = \int_{t'=0}^{t} e^{-\mathbf{L}(t-t')}\vec{b}u(t')dt'. \qquad \overrightarrow{p}_{SS} = \mathbf{L}^{-1}\overrightarrow{b}$$

Qi, S; Suhail, Y.; Lin, Y.; Boeke, J.D.; Bader, J.S. Finding friends and enemies in an enemies-only network: A graph diffusion kernel for predicting novel genetic interactions and co-complex membership from yeast genetic interactions. *Genome Res,* **2008,** *18(12)*: 1991-2004

Background: PPI Networks

- Human Protein Reference Database (HPRD): database of curated proteomic information



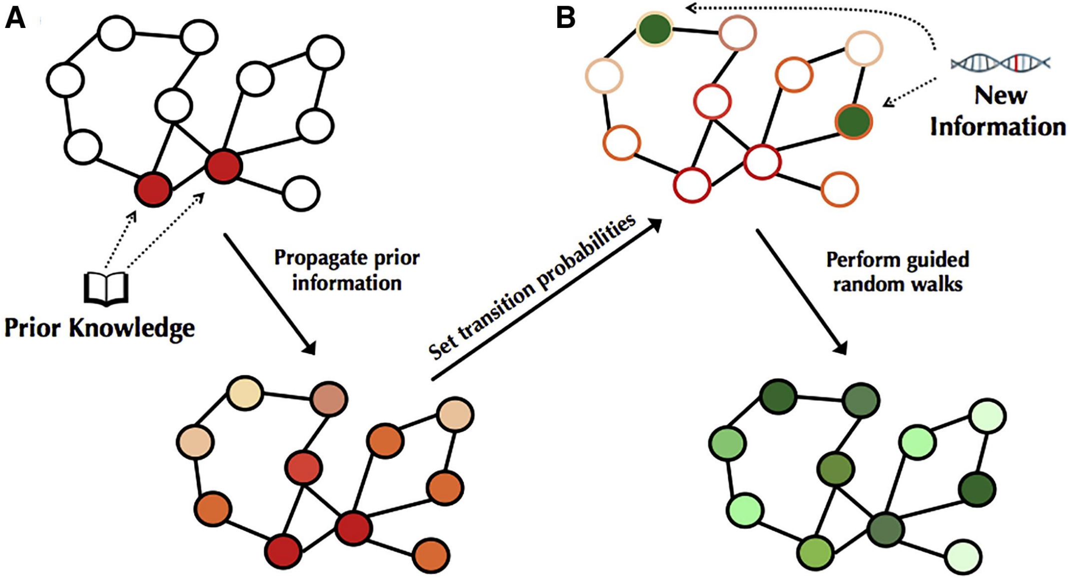
Statistics

Protein Entries	30,047
Protein-Protein Interactions	41,327
PTMs	93,710
Protein Expression	112,158
Subcellular Localization	22,490
Domains	470
PubMed Links	453,521

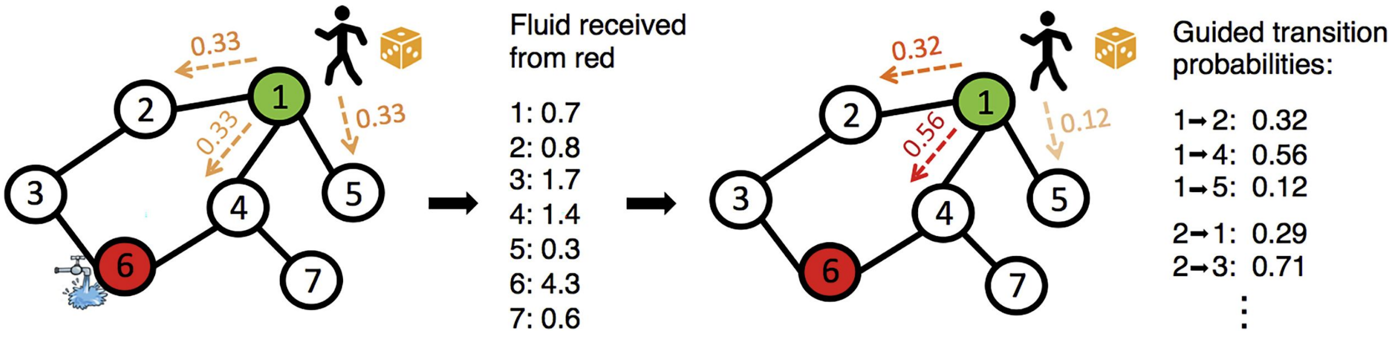
- Last release, release 9, was 4/13/2010
- Filtered network with 9,379 proteins and 36,638 interactions used for uKIN

Prasad, T. S. K., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., Balakrishnan, L., Marimuthu, A., Banerjee, S., Somanathan, D. S., Sebastian, A., Rani, S., Ray, S., Kishore, C. J. H., Kanth, S., Ahmed, M., Kashyap, M., Mohmood, R., Ramachandra, Y. L., Krishna, V., Rahiman, A. B., Mohan, S., Ranganathan, P., Ramabadran, S., Chaerkady, R. and Pandey, A. Human Protein Reference Database - 2009 update. Nucleic Acids Research. 37, D767-D772.

uKIN Method (overview)


A
B
Prior Knowledge
Propagate prior information
Set transition probabilities
New Information
Perform guided random walks

uKIN Method



Fluid received from red

1: 0.7
2: 0.8
3: 1.7
4: 1.4
5: 0.3
6: 4.3
7: 0.6

Guided transition probabilities:

1→2: 0.32
1→4: 0.56
1→5: 0.12

2→1: 0.29
2→3: 0.71
⋮

uKIN Method

Graph:

$$G = (V, E)$$

$$K = \{k_1, k_2, \ldots, k_l\} \; M = \{m_1, m_2, \ldots, m_p\} \; F = \{f_{m_1}, f_{m_2}, \ldots, f_{m_p}\} \; K \subset V, M \subset V, K \cap M = \emptyset$$

Diffusion: $q = L^{-1}b$

RWR: $$p_{ij} = ((1 - \alpha)\delta_{ij})\frac{q_j}{\Sigma_{k \in N(i)} q_k} + \alpha \frac{f_j}{\Sigma_{k \in M} f_k}$$

$$\delta_{ij} = 1 if j \in N(i)$$

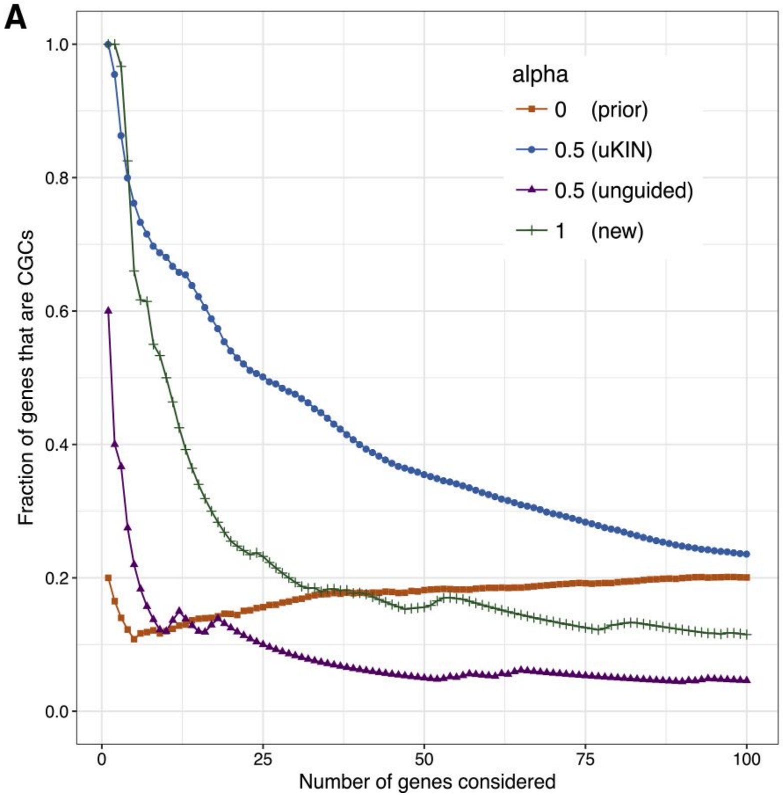
$$\pi P^t = \pi$$

Method comparisons:

- The Cancer Genome Atlas (TCGA) used for ‘new’ information, mutation frequency is the number of somatic and nonsense mutations per gene across tumor samples / # amino acids in the protein product
- 719 Cancer Gene Census genes that are labeled by COSMIC (version August 2018)
- 400 randomly drawn CGCs for a hidden set, H
- 20 CGC genes selected for K
- Ran uKIN 100 times drawing H and K, considered top 100 gene predictions for evaluations
- Metric 1: Fraction of top predicted genes in H
- Metric 2: AUPRC using H as the true labels, CGCs not in H as neutral, and all other genes as negatives. Used log2 of AUPRC between methods to compare them.

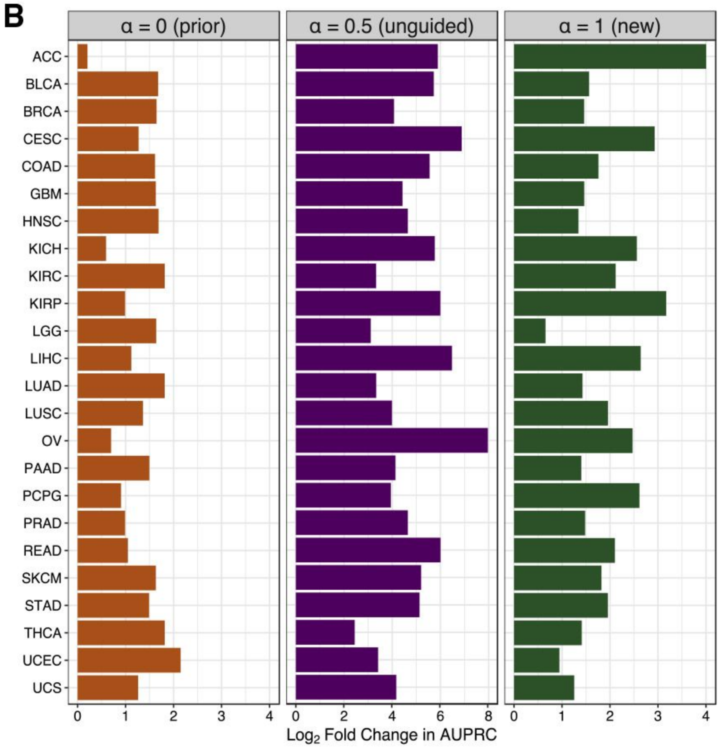
uKIN Example: glioblastoma multiforme GBM

- Unguided is RWR, but without diffusion component



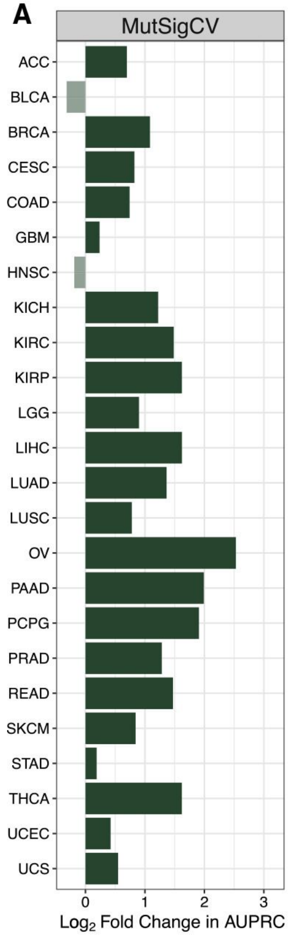
uKIN on all cancers

- Log change of AUPRC of uKIN compared to other methods for 24 cancers
- uKIN outperforms using only prior information and only new information in all cases.



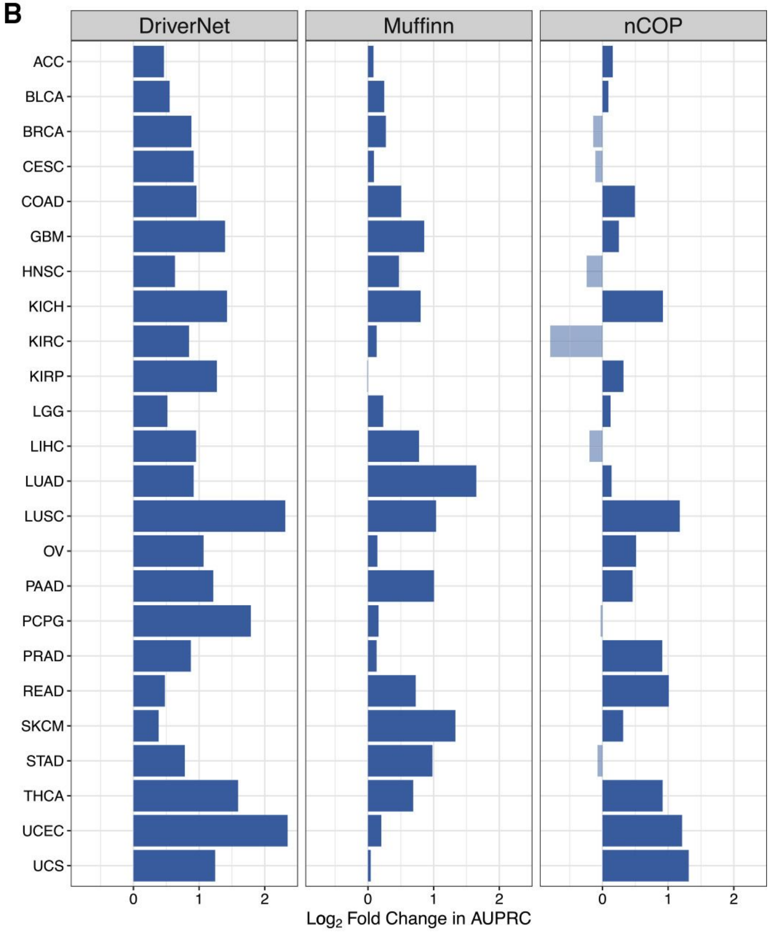
Comparing uKIN to other methods

- MutSigCV 2.0: mutation frequency based approach to identify cancer genes.
- uKIN had an increased AUPRC for 22/24 cancer types



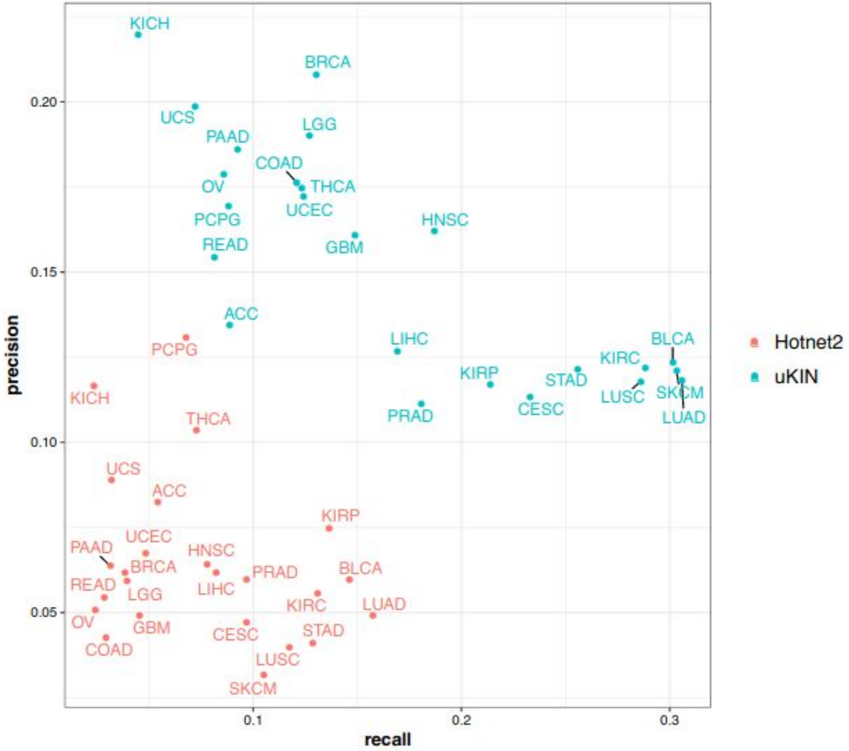
Comparing uKIN to other methods

- Muffinn: considers mutations in interacting genes. (uKIN outperforms on 23/24)
- DiverNet: finds driver genes by uncovering sets of somatically mutated genes lined to dysregulated genes. (uKIN outperforms on 24/24)
- nCOP: examines per-individual mutation profiles of cancer patients in a network (uKIN outperforms on 17/24)



Comparing uKIN to other methods

- Hotnet2: Diffusion kernel based method
- No ranked list of genes for output, instead outputs a list of genes predicted to be cancer relevant vs. not relevant
- Shows the benefit of using prior information & diffusion for uKIN

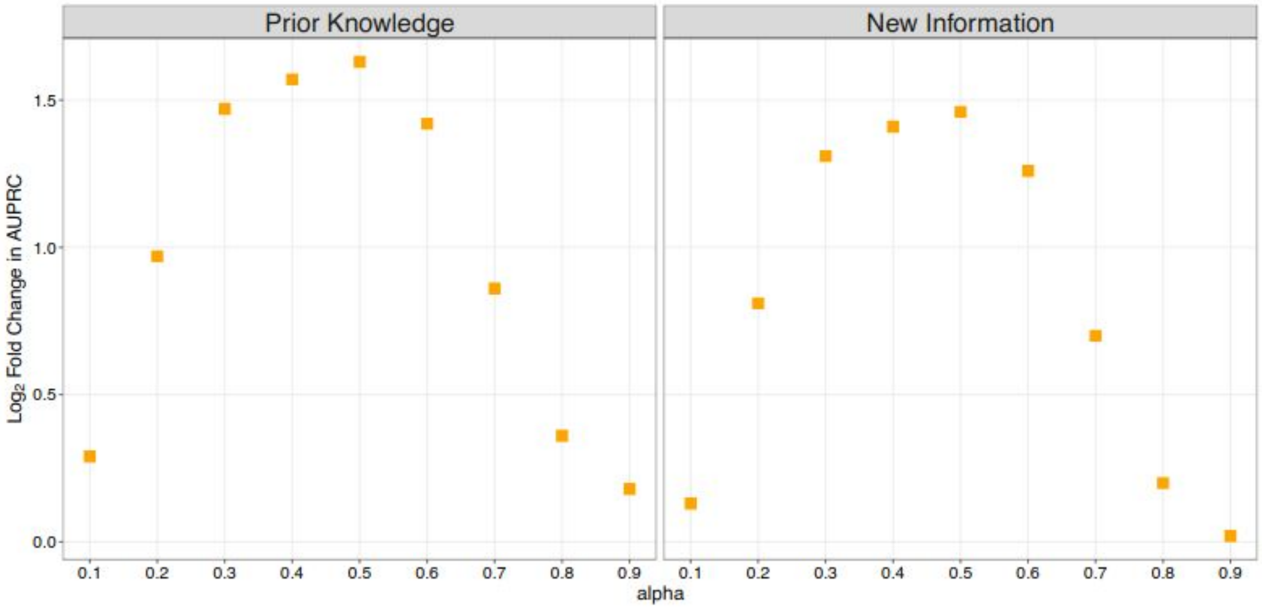


Robustness:

- Similar results for self and alternative method comparisons using non-Cancer Gene Census test set
- Similar results using the top 50 genes to compute AUPRCs instead of the top 100
- Similar results using biogrid PPI network instead of the HPRD
- Performance goes down with randomized PPI networks when using uKIN, as would be expected

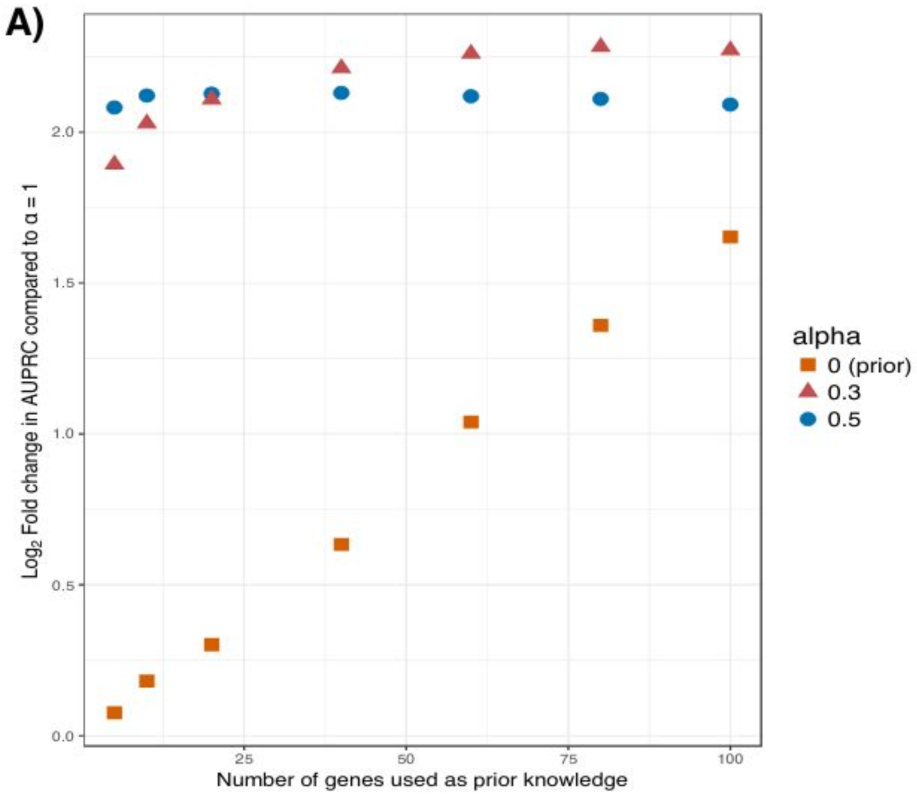
Varying alpha

- $0.1 \leq \alpha \leq 0.9$ were tested for GBM, with all values resulting in increased performance compared to α=0 and α=1 for uKIN



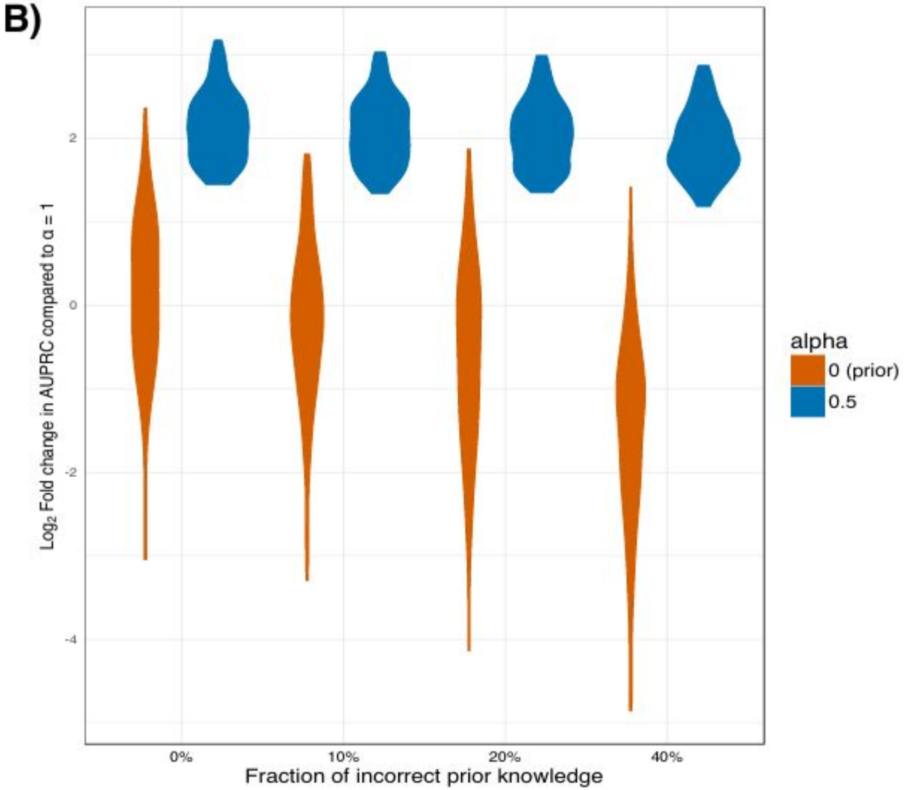
Varying prior knowledge:

- As few as 5 prior knowledge genes improves performance over ranking genes by mutational frequency

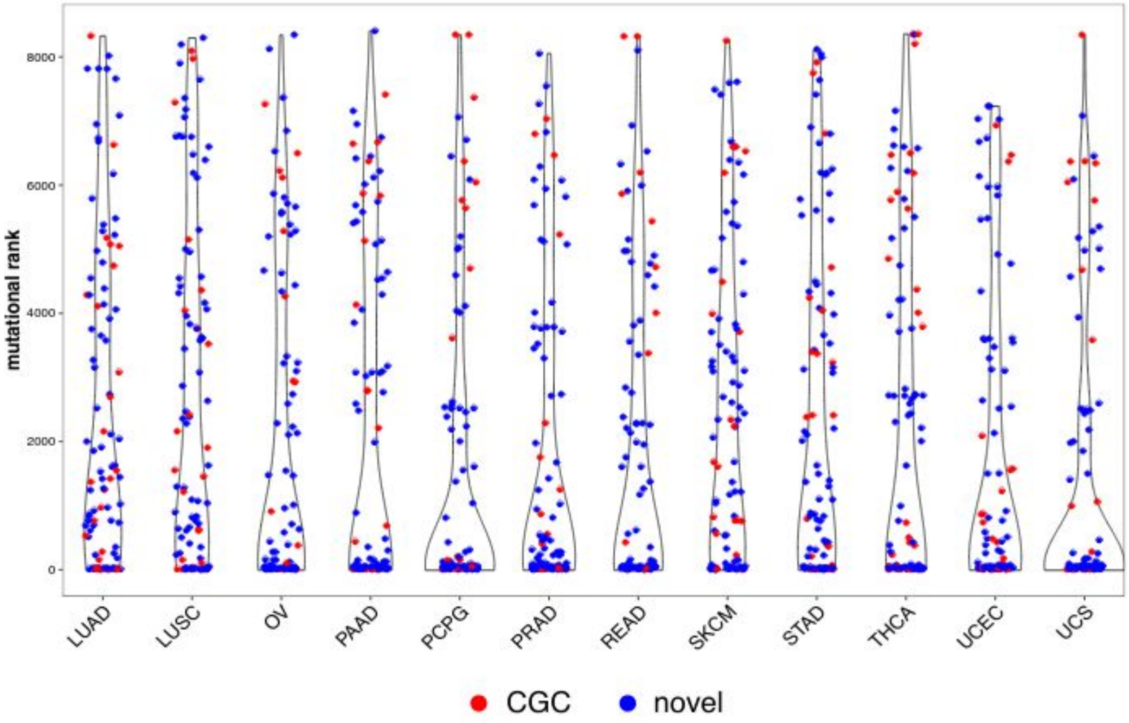


Incorrect prior knowledge

- uKIN with α=0.5 performs reasonably well with less than 20% incorrect annotations

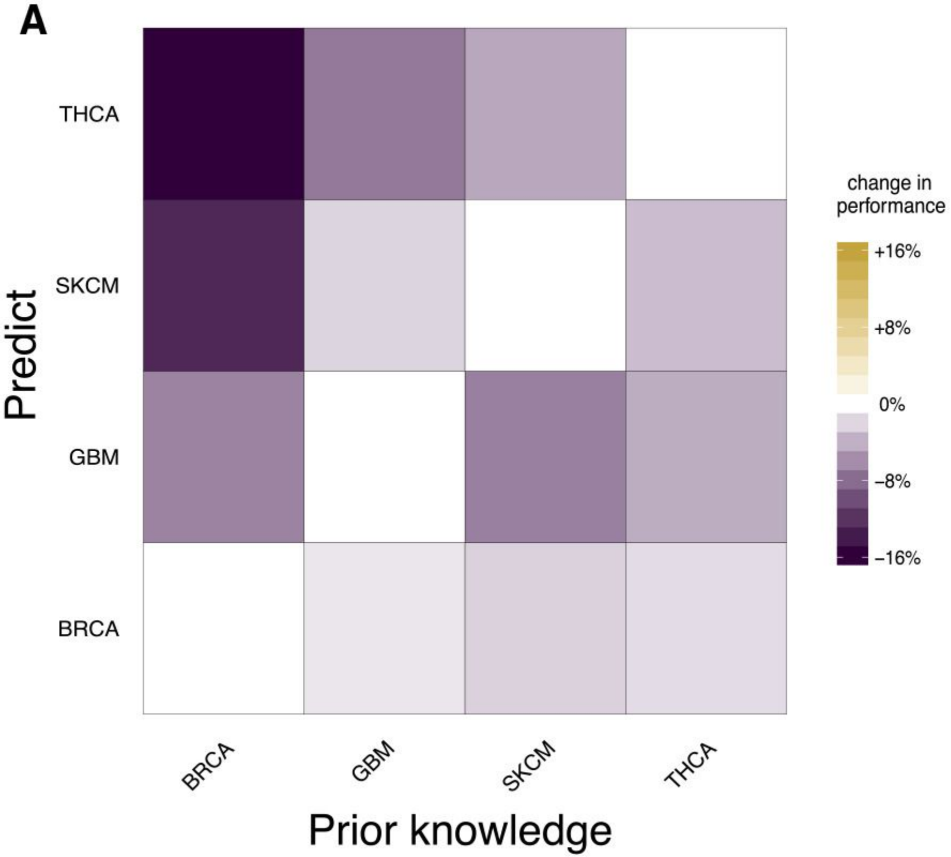


uKIN Highlights Infrequently Mutated Genes



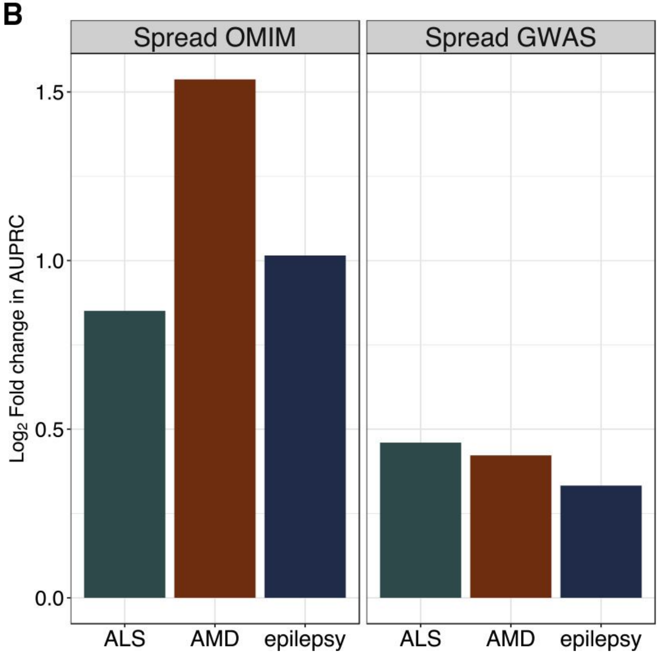
Cancer-specific prior knowledge

- Some CGC genes are annotated with the specific cancers they are drivers for
- glioblastoma multiforme (GBM) (33), breast invasive carcinoma (BRCA) (32), skin cutaneous carcinoma (SKCM) (42), and thyroid carcinoma (THCA) (29)

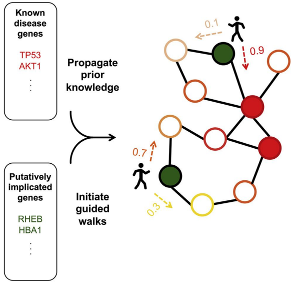


uKIN example: complex inherited disorders

- Amyotrophic lateral sclerosis (ALS), age-related macular degeneration (AMD), and epilepsy.
- Uses OMIM's disease associated list of genes for each disease for prior knowledge and hidden set to evaluate uKIN
- Sorting the genes by GWAS significance results in AUPRC 0 (uKIN with α=1)



Conclusions



- uKIN is effective, versatile, and robust.
- Because of using prior knowledge, it outperforms other state-of-the-art methods
- It can be used for cancer and other complex diseases
- Calibration of α does not seem to be necessary, but it could be varied with the amount of prior information available
- Extensions:
  - “Negative” knowledge of disease genes could be incorporated
  - Adding edge weights for interaction reliability
  - Scale starting probabilities using natural germline variation data
  - Use cancer subtype distinct information
  - uKIN could be applied to other biological network propagation problems (process prediction, drug target identification, etc.)

Discussion questions

- Mutational frequency is used for the RWR- what alternatives could be used for choosing where the random walks begin and restart from?
- How could PPI network quality affect uKIN performance? Some interactions are not as certain as others, and some interactions vary between cell types.
- Of the extensions, which seem the most promising?
    - "Negative" knowledge of disease genes could be incorporated
    - Adding edge weights for interaction reliability
    - Scale starting probabilities using natural germline variation data
    - Use cancer subtype distinct information
    - uKIN could be applied to other biological network propagation problems (process prediction, drug target identification, etc.)

Diffusion kernels:

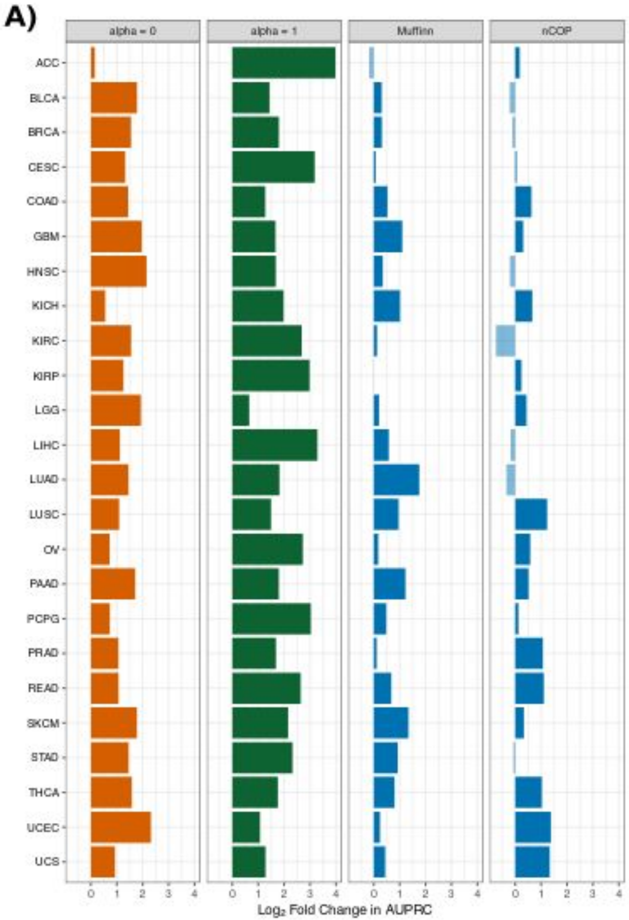
$$\dot{p}_i\,(t) = \sum_j A_{ij}p_j(t) - \{\gamma + \sum_j A_{ji}\}p_i(t) + b_i u(t),$$

$$\dot{\vec{p}}\,(t) = (\mathbf{A} - \mathbf{S} - \gamma\mathbf{I})\vec{p}(t) + \vec{b}u(t), \tag{2}$$

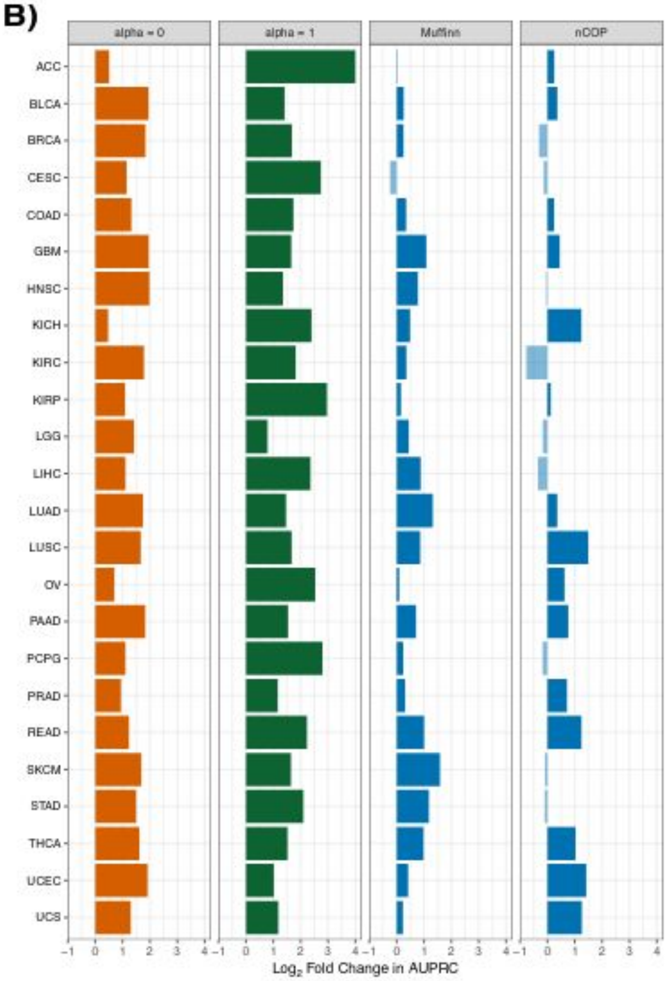
$$\vec{p}(t) = \int_{t'=0}^{t} e^{-\mathbf{L}(t-t')}\vec{b}u(t')dt'.$$

$$\vec{p}_{ss} = \lim_{s\to 0} s\,\frac{1}{s}\,(s\mathbf{I} + \mathbf{L})^{-1}\vec{b} = \mathbf{L}^{-1}\vec{b}. \tag{3}$$

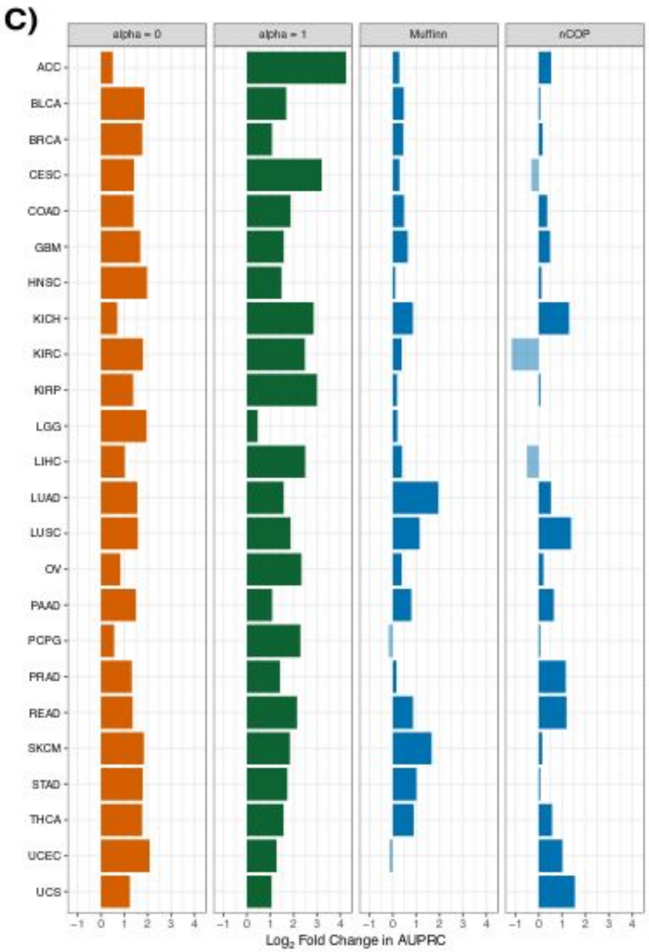
Different cancer gene labels:



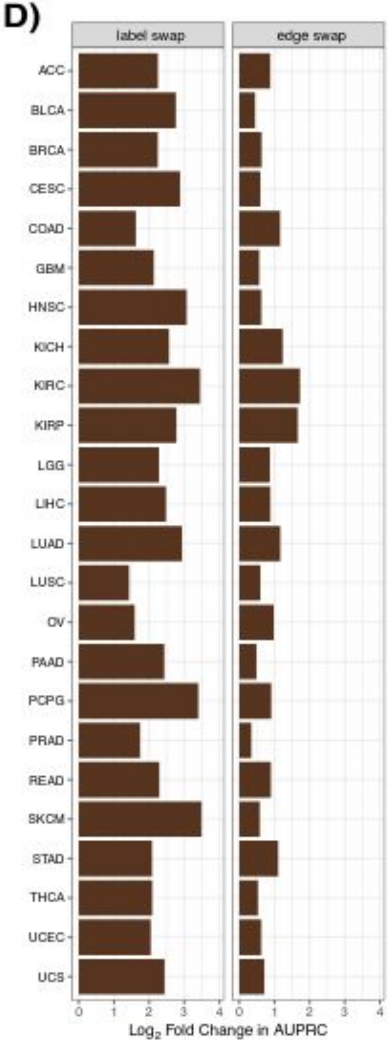
Different cutoff for AUPRC calculations:



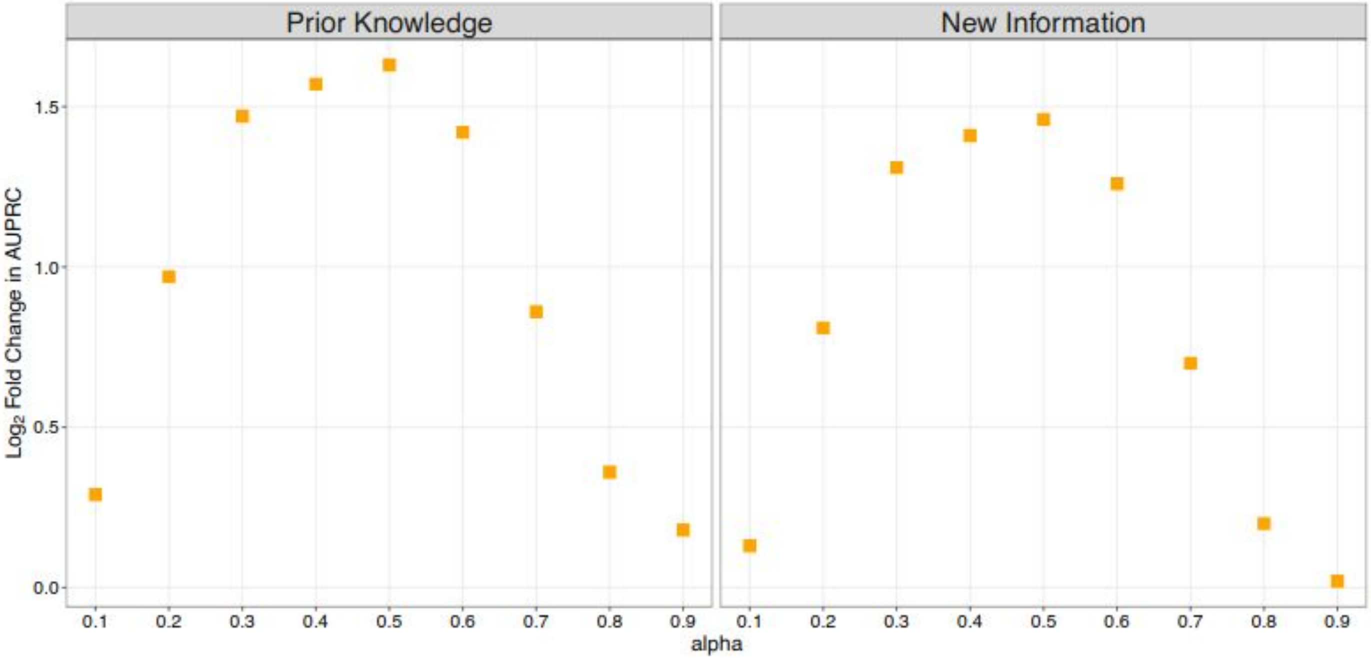
Different PPI network: Biogrid



Network shuffling:



GBM alpha value investigation


Prior Knowledge
New Information
Log$_2$ Fold Change in AUPRC
alpha