Predicting genes associated with Alzheimer's Disease: A biological networks approach

Sarrah Hakimjee[1], Kyla Levin[1], Mary-Joy Sidhom[1], Mert Erden[1], Kapil Devkota[1], Lenore J. Cowen[1]

[1] Department of Computer Science, Tufts University, Medford MA

MOTIVATION

Alzheimer's Disease (AD) is a neurological degenerative disorder that is the most common cause of dementia in elderly people. There currently exists no treatment to slow the progression of AD. However, many studies have been conducted to understand AD and learn the genes associated with it. The aim of this study is to **use different network propagation methods to determine more genes that are associated with AD given the genes that we already know to be associated with AD**. This information will be instrumental for understanding the mechanics of AD and for further research on what we can do to slow its progression.

The two main network propagation methods that were studied and and utilized are Random Walk with Restart (RWR) and GLIDER. These methods were run on a set of genes associated with Alzheimer's to determine more genes associated with it. A hybrid method that involves modifying these network propagation methods will be devised to increase the efficiency and accuracy of the algorithm.

This project aims to present all genes associated with the brain in a network of nodes and given a set of genes that are known to be related to AD find more genes that are associated with AD. The goal is to not only have a larger set of known genes related to AD but also to have a ranked order of the genes.

DATASETS

HURI: (HuRi Link Here)

DREAM networks: (RG Link here)

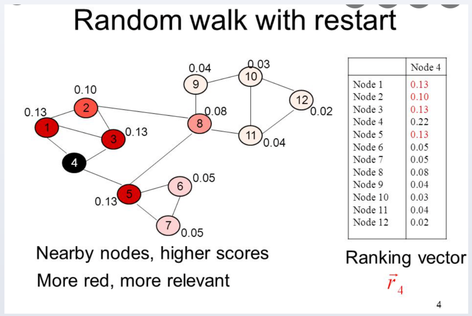
P. StringDB network: (StrindDB Link here)

METHODS

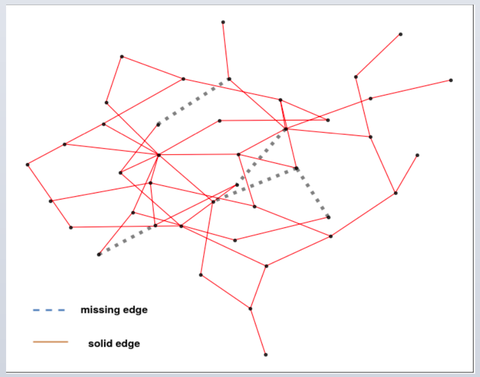
How we came about the datasets:
Huri was used because it gave the most wholistic network data, other sources of data are biased towards genes that have already been discovered to have associations with harmful diseases

How we figured out what the known genes are:
We explored GWAS studies on Alzheimer's disease and picked genes with significant p-values

Random Walk with restart:
Random walk with restart is a prediction based algorithm that traverses the graph from source node with a certain probability for restart. The most visited nodes are then recorded as the possible related nodes



GLIDER
Glider is an unknown link prediction algorithm that runs global prediction algorithms and local prediction algorithms depending on where it is in the biological cluster. It takes in a graph and predicts new edges with scores for confidence



Validation
- Leave one out cross validation
- AUROC: Area under the receiver operating characteristic

RESULTS

The top five genes that were predicted by random walk with restart are displayed on the left, most of them show promising literature evidence of an association with AD.

Top 5 predicted genes for AD (Random Walk with restart)	RELEVANT LITERATURE STUDIES
FYN	Fyn kinase induces synaptic and cognitive impairments in a transgenic mouse model of Alzheimer's disease.
RPS27A	Rps27a might act as a controller of microglia activation in triggering neurodegenerative diseases
ITPKB	Inositol trisphosphate 3-kinase B is increased in human Alzheimer brain and exacerbates mouse Alzheimer pathology
UBA52	Walking the tightrope: proteostasis and neurodegenerative disease
UBC	The polyubiquitin Ubc gene modulates histone H2A monoubiquitylation in the R6/2 mouse model of Huntington's disease

KNOWN GENES FOR AD

Genes associated with AD		
CR1	CLU	ADAMTS4
BIN1	SPI1	HESX1
INPP5D	MS4A2	CLNK
HLA-DRB1	PICALM	ZCWPW1
TREM2	SORL1	CNTNAP2
CD2AP	FERMT2	ECHDC3
NYAP1	SLC24A4	MS4A6A
EPHA1	ABCA7	ADAM10
PTK2B	APOE	APH1B
ALPK2	CASS4	KAT8
ABI3	CD33	SCIMP
		AC074212.3

FURTHER RESEARCH

To move this investigation further, it would be interesting to score the predicted genes to determine the best algorithms. In addition, it would be interesting to implement something like RUOIC to score the final genes based on the best algorithm.

ACKNOWLEDGEMENTS

1. Laidlaw foundation
2. Cowen Lab, Tufts University
3. Tufts Bioinformatics and computational Biology research group

REFERENCES

Luck, Katja et al. "A Reference Map Of The Human Binary Protein Interactome". *Nature*, vol 580, no. 7803, 2020, pp. 402-408. *Springer Science And Business Media LLC*, https://doi.org/10.1038/s41586-020-2188-x. Accessed 5 Nov 2021.

Magger, Oded et al. "Enhancing The Prioritization Of Disease-Causing Genes Through Tissue Specific Protein Interaction Networks". *Plos Computational Biology*, vol 8, no. 9, 2012, p. e1002690. *Public Library Of Science (Plos)*, https://doi.org/10.1371/journal.pcbi.1002690. Accessed 5 Nov 2021.

Vanunu, Oron et al. "Associating Genes And Protein Complexes With Disease Via Network Propagation". *Plos Computational Biology*, vol 6, no. 1, 2010, p. e1000641. *Public Library Of Science (Plos)*, https://doi.org/10.1371/journal.pcbi.1000641.

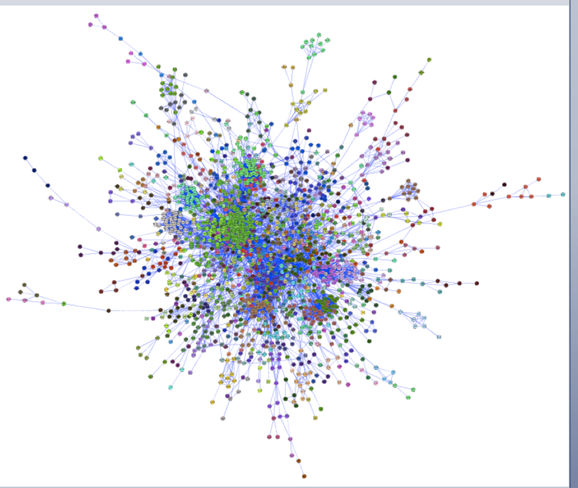
A Network-Based Approach to Disease Gene Prediction for Parkinson's Disease

Kyla Levin[1], Sarrah Hakimjee[1], Mary-Joy Sidhom[1], Mert Erden[1], Kapil Devkota[1], Lenore J. Cowen[1]

[1] Department of Computer Science, Tufts University, Medford, MA

MOTIVATION

Parkinson's Disease, or PD, is a long-term, debilitating degeneration of neurons in the central **nervous system, often first causing symptoms such as tremors or loss of balance before progressing into insomnia, depression, and other cognitive impairments. Although common in the U.S., only a small list of genes are known to directly cause PD through their mutations. Discovering more disease genes involved with PD could potentially lead to earlier diagnoses of PD or refined genetic treatments to PD with increased efficacy. Because the known disease genes exist in a large protein-protein interaction network, or PPI network, this network can be read as data for a graph with nodes and edges. This data can then be traversed by algorithms that will evaluate the likelihood of other genes to also be PD-causing based upon their known interactions. Therefore, we applied two label-predicting algorithms to a union of existing PPI network datasets, evaluating them by their accuracy in predicting already known disease genes, as well as verifying their disease gene predictions through existing literature.**



DATASETS

Huri

http://www.interactome-atlas.org/

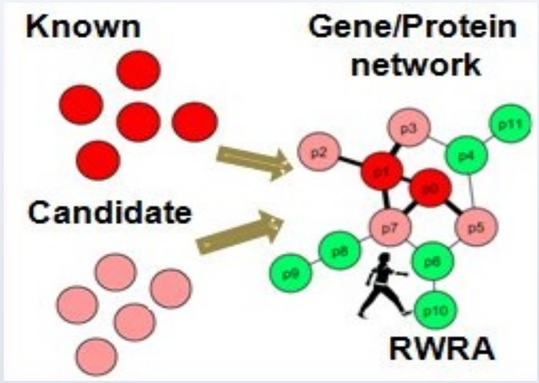
StringDB

https://string-db.org/

Dream 1, Dream 2

https://dreamchallenges.org/

METHODS



Evaluating the Algorithms

Methods were validated using leave-one-out cross validation:

Each known gene is excluded from the list of "known" genes. An accurate algorithm is then expected to produce each missing gene high in its ranked predictions for candidate disease genes.

Random Walk with Restart

Based upon the idea that a traverser "steps" away from each known gene, and with each step, there is a chance that they were reappear back at the node where they started.

Candidate genes are those that have been visited the most often due to close proximity/high interactions with the known disease genes.

Calibrated using alpha (0-1 chance of restarting) and a k-value (number of highest-ranking genes to report).

Altered using DaDa, which adjusted the results to account for a bias towards more well-studied and thus better-connected nodes.

Glider

Better suited towards PPI networks, which are incomplete in their edges and are sparse, but with very dense centers.

Local methods work better for sparse neighborhoods, while global methods work better in highly-connected clusters.

Glider runs one of two algorithms—either a local or global one—depending on network density.

Local method: triatic closure. Based on the idea that nods with multiple common neighbors must also be related.

Glider reproduces the network's edges with scores, where each edge has a confidence score. However, Glider does not accommodate for false positives.

- Glider decides where to go based off common friends, not just direct edges

RESULTS

Results from the methods show that the five genes, RPS27A, UBA52, UBB, UBC, and ITPKB were predicted as the most likely to have a direct involvement with causing Parkinson's Disease. While no literature confirming the involvement of UBA52 or UBB in PD could be found, evidence shows that the other top five genes have been studied with respect to other neurodegenerative diseases. RPS27A[2] has been shown to contain a switch for neurodegenerative diseases in its mutations, while UBC[3] has been studied in mice with Huntington's Disease and ITPKB[4][5] has a direct correlation to the symptoms of both PD and Alzheimer's Disease.

The benefits of this study are to help future research narrow in on which genes to possibly look into for a more specific causal relationship with a given patient's disease. The hope is that these list of confidence ratings can guide research into developing more specific PD treatments.

Gene(name='RPS27A'	labels=None) 0.0008875247934060238
Gene(name='UBA52'	labels=None) 0.0008778151212065721
Gene(name='UBB'	labels=None) 0.0008680038732019332
Gene(name='UBC'	labels=None) 0.0008611809359110106
Gene(name='ITPKB'	labels=None) 0.0004866017667798674

Top 5 unknown genes of highest confidences derived from the known disease modules for PD using Random Walk with Restart

Parkinson's Disease Genes[1]

Gene	Confidence	GWAS-Nominated
SNCA	Very high	Yes
PRKN	Very high	No
PARK7	Very high	No
LRRK2	Very high	Yes
PINK1	Very high	No
ATP13A2	Very high	No
FBXO7	Very high	No
GBA	Very high	Yes
PLA2G6	Very high	No
VPS35	Very high	No
POLG	High	No
DNAJC6	High	No
SYNJ1	High	No
VPS13C	High	Yes

Confidences derived from existing literature, mutations in the genes being present in families in long-term PD studies, and the ability to replicate these findings.

ACKNOWLEDGEMENTS

Thank you to Prof. Cowen for accepting me onto this project and for agreeing to be my research advisor. And thank you to the Laidlaw Program for helping to sponsor my research.

REFERENCES

1. Blauwendraat, Cornelis, et al. "The Genetic Architecture of Parkinson's Disease." The Lancet Neurology, vol. 19, 2020, pp. 170-78, doi:10.1016/S1474-4422(19)30287-X.
2. Khayer, Nasibeh, et al. "Rps27a Might Act as a Controller of Microglia Activation in Triggering Neurodegenerative Diseases." PLOS ONE, edited by Santosh R D'Mello, vol. 15, no. 9, 2020, p. e0239219. Crossref, doi:10.1371/journal.pone.0239219.
3. Bett, J.S., Benn, C.L., Ryu, K.-Y., Kopito, R.R. and Bates, G.P. (2009), The polyubiquitin Ubc gene modulates histone H2A monoubiquitylation in the R6/2 mouse model of Huntington's disease. Journal of Cellular and Molecular Medicine, 13: 2645-2657. https://doi.org/10.1111/j.1582-4934.2008.00543.x
4. Apicco, Daniel J., et al. "The Parkinson's Disease-Associated Gene ITPKB Protects against α-Synuclein Aggregation by Regulating ER-to-Mitochondria Calcium Release." Proceedings of the National Academy of Sciences, vol. 118, no. 1, 2020, p. e2006476118. Crossref, doi:10.1073/pnas.2006476118.
5. Zhang, Yizhi, et al. "Microglia-Derived Extracellular Vesicles Carrying MiR-711 Alleviate Neurodegeneration in a Murine Alzheimer's Disease Model by Binding to Itpkb." Frontiers in Cell and Developmental Biology, vol. 8, 2020. Crossref, doi:10.3389/fcell.2020.566530.