Identification Of Risk Genes For Neurodegenerative Diseases

Project Proposal

CS BATCH 2020

Group Members:

Robaisha Masood 20K-0390

Hira Tahir 20K-0374

Noor Afshan 20K-0266

Project Supervisor

Signature: __________________

Name: Shoaib Rauf

National University Of Computer and Emerging Sciences
Karachi, Pakistan

1. **Abstract:**

Our study targets neurodegenerative diseases. Our aim is to conduct research to identify key genes that govern the pathogens of such diseases and find proteins that are most vulnerable in each dataset. These bacteria form communities that cause malfunctioning. Two to three datasets will be used from NCBI GEO Data and DisGeNet. We will form individual networks of controlled and diseased from each dataset and find a centroid(key gene) that has the major impact and find its similarities by making comparisons between each network. A Weighted Gene Co-expression Network will be used to tell the hub genes and their role in such diseases. Machine learning algorithms will be used to further tune and validate the results.

Neurodegenerative diseases cause memory loss, behavioral abnormalities, and language deficits. The nature of such diseases is a hurdle to early diagnosis. [1] An estimated six million in the United States are afflicted with Alzheimer's Disease. This number is projected to double by 2050. Therefore our research will reveal the protein-to-protein interaction that will help in early diagnosis and drug development.

The project will give us a better understanding of ML algorithms and their implementation in bioinformatics which will surely help us to contribute to the community in the future. This study will not only contribute to scientific knowledge but also benefit those who are living with such diseases. Our project aims to benefit the community on a large scale and raise awareness for neurodegenerative diseases that can lead to more effective treatments.

2. **Introduction**:

Neurodegenerative diseases are basically brain disorder conditions when the neurons start to degenerate gradually due to abnormal accumulation and misfolding of specific proteins. This leads to neuronal damage and causes severe disorders. Neurodegenerative diseases include conditions like Alzheimer's disease, Parkinson's disease, Huntington's disease, and Amyotrophic Lateral Sclerosis.

The key point that will be covered in this research paper is the identification of the key hub gene that plays a special role in the pathogenesis of neurodegenerative disorders. These diseases often lead to complex interactions among multiple genes and proteins, which leads to the formation of dysfunctional neural networks.

Our study will use bioinformatics for analyzing datasets from NCBI GEO Database and DisGeNet. We will construct networks of healthy and diseased conditions from each dataset which will pinpoint the hub genes that are significantly influencing the disease. These hub genes will further be evaluated using various algorithms, for instance, Weighted Gene Co-expression Network Analysis (WGCNA). We aim to discover the role of genes in neurodegenerative disorders, to potentially improve the diagnosis and treatment of these diseases.

3. **Problem Statement:**

What are the key genes that are responsible for neurodegenerative diseases? As neurodegenerative diseases like Alzheimer's and Parkinson's continue to rise it imposes societal challenges.For early diagnosis, it is very important to first identify the key genes that are causing the malfunction. Research is required to better understand these genetic interactions which in turn will contribute to drug development and enhance quality of life for those who are diseased. Machine Learning and AI are helpful in developing predictive models and algorithms. In addition to this, our project will help us learn skills like data visualization , image processing, and machine learning.Computer science is essential in the research of neurodegenerative disease as it provides computational tools and methods to benefit society.

4. **Literature Review:**

Neurodegenerative diseases are a significant challenge for public health. The study of neurodegenerative disorders demands a needful understanding of the genetic foundations. This literature review highlights the key findings from the recent research in neurodegenerative disease genetics.

The "network dysfunction perspective" on neurodegenerative diseases, as proposed by Jorge J. Palop, Jeannie Chin, and Lennart Mucke, focuses on understanding these diseases by the help of examination of broader neural networks and connection within the brain, rather than working with a single individual area.

The paper "Chapter 21 - Concepts and Classification of Neurodegenerative Diseases" by Gabor G. Kovacs discusses neurodegenerative diseases and follows the classification with providing detailed description on neuropathology of ALzheimer diseases, alpha-synucleinopathies, tauopathies, FTLD with TDP-43 or FUS/FET proteinopathies, trinucleotide repeat disorders, and prion diseases.

" Applications of machine learning to diagnosis and treatment of neurodegenerative diseases" by Monika A. Myszczynska, Poojitha N. Ojamies, …Laura Ferraiuolo explains how machine learning is helping in diagnosis in the early stage and interpretation of medical images as well as discovery and development of new therapies. This also helps in multiple high dimensional sources of data which provide different views on different diseases.

A cell biological perspective on mitochondrial dysfunction in Parkinson disease and other neurodegenerative diseases focus on mutations of reasons causing Parkinson diseases. It tells the study of these mutations and it causes other diseases which indicates the mitochondrial dysfunction which becomes the main contributor to neurodegenerative processes.

“Neurodegenerative Diseases and Prions” by Stanley B. Prusiner, M.D. tell that it is clear that prion or neurodegenerative diseases result from abnormalities in the process of these diseases which further cause accumulation of specific neuronal proteins. Laboratory search results led to discovery of prions yielding findings like infectious pathogens and degeneration of the central nervous system.

“Alzheimer Disease and Related Neurodegenerative Diseases in Elderly Patients With Schizophrenia” by Dushyant P. Purohit, MD; Daniel P. Perl, MD; Vahram Haroutunian, PhD; et al Clinical studies suggest that severe cognitive impairment is common in older people combine with schizophrenia who reside most in psychiatric people and it also tell that its result is conflicting with Alzheimer disease when combine with schizophrenia.

In the research, “Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease.” by Lambert, J. C., et al. (2013); This was conducted on 74,046 individuals, which had led to 11 new discoveries associated with Alzheimer’s disease. Their work highlights the importance of large-scale collaborative research in deciphering the complex genetic landscape of this condition.

Jun et al. (2016) presented his opinion in the novel by identifying an uncharted Alzheimer's disease locus near the tau gene. This discovery essentially shows that there could be a potential link between the tau protein pathology and susceptibility to Alzheimer's disease.

Karch, C. M., et al. (2014) presented his study, " Alzheimer's Disease Genetics: from the bench to the clinic.". The key discussion in this paper was the need to build the gap between genetic discoveries and clinical applications in Alzheimer's disease genetics. Their work highlights the genetic factors involved in Alzheimer's disease and it also underscores the importance of applying this knowledge in clinical practice to potentially improve the diagnosis and the treatments.

"Alzheimer's disease: The amyloid cascade hypothesis" by John Hardy and Dennis J. Selkoe. This paper discusses the amyloid cascade hypothesis, which is a basic theory in Alzheimer's disease. The paper explores the role of amyloid-beta protein in the pathogenesis of Alzheimer's disease.

"Neurodegenerative disorders associated with genes of mitochondria" by Vaibhav S. Marde et al(2021). The research paper emphasizes on the causal link between mitochondrial gene mutations and neurodegenerative disorders such as Parkinson's and Alzheimer's diseases. It reviews structural and functional studies, highlighting mitochondrial dysfunction's role. In the study, animal models provide evidence for mitochondria's involvement in disease initiation and progression.

In 2015, Escott-Price highlighted the impact of common polygenic variations in Alzheimer's disease research. They state how these genetic variations enhance the assessment of risk and early diagnosis of the disease, providing more accurate prediction of the stages.Microglial-Mediated Innate Immunity:

The authors of the study introduce a novel approach called GeneEMBED, which aims to identify gene interactions associated with complex diseases such as Alzheimer's. Through the application of GeneEMBED on multiple datasets related to Alzheimer's, researchers successfully discovered previously unidentified genes (namely PLEC, UTRN, TP53, and POLD1) that play a role in this disease. Additionally, it was observed that two out of these four genes are targeted by approved pharmaceutical drugs.

Despite the significant achievement, the journey of comprehending the genetic biases of neurodegenerative diseases remains incomplete as the gap persists in understanding the genetic variants, deciphering gene interactions, and translating genetic insights into clinical sciences. Future researchers should focus on uncovering the functional significance of identified genetic variants, probing gene-gene interactions and better genetic insights.

In conclusion, the research reviewed here offers crucial insights into the genetic foundations of neurodegenerative diseases. A collaborative, interdisciplinary approach is essential to further our understanding and develop strategies for early diagnosis and effective treatment.

S.No.	Title	Year	Main Points	Targeted Diseases
1.	Network dysfunction perspective on neurodegenerative diseases	2017	Understanding the diseases by brain & neural networks	Parkinson
2.	Concepts and classification on neurodegenerative	2018	Detailed description on prion diseases	Alpha-synucleinopathies
3.	Application of machine learning to diagnosis and treatment of neurodegenerative diseases	2020	How machine learning helps in early diagnosis and its treatment	Neuronal degenerative
4.	A cell biological perspective of mitochondrial dysfunction in Parkinson diseases	2018	It caters with mutation of Parkinson diseases	Parkinson
5.	Neurodegenerative diseases and prions	2016	Prion diseases and accumulation of neural proteins	Prion

6.	Alzheimer diseases and other neurodegenerative diseases in elderly people	2021	It tell how schizophrenia works with Alzheimer	schizophrenia + Alzheimer
7	Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease.	2013	Meta-analysis identifies 11 new susceptibility loci for Alzheimer's disease.	Alzheimer's Disease
8	novel by identifying an uncharted Alzheimer's disease locus near the tau gene	2016	Novel Alzheimer disease locus near the tau gene.	Alzheimer's Disease
9	Alzheimer's disease genetics: from the bench to the clinic.	2014	Discusses Alzheimer's disease genetics from bench to clinic.	Alzheimer's Disease

10	Neurodegenerative disorders associated with genes of mitochondria	2021	Mitochondrial genes have provided compelling evidence that mitochondria is involved in the initiation as well as progression of diseases.	Parkinson’s disease (PD), Alzheimer’s disease (AD), Huntington’s disease (HD), and Friedreich ataxia (FA).
11	Examining the Polygenic Variation Enhancing Alzheimer's Disease Risk Prediction.	2015	Examines how common polygenic variation enhances Alzheimer's disease risk prediction.	Alzheimer's Disease

5. **Methodology**:

We will first study and research two to three datasets that will contain information about individuals who are healthy and the ones who are affected by neurodegenerative disease. The preprocessing of data will be done via Python using some basic Python libraries such as numpy, pandas, and matplotlib on Google Collaboratory or Jupyter Notebook. We will form separate networks from each dataset that will contain protein-to-protein interactions using some WGCNA packages such as cluster profile and Limma. For construction of WGCNA Rstudio will be used.

A comparison of each controlled vs. diseased will be made in which a center point for each network will be identified.ANOVA and Scipy will be used for comparison. Since bacteria form communities , we will apply a community detection algorithm to identify the key gene. An ML algorithm will be chosen accordingly to apply similar weights to each key gene and find the actual protein that is responsible for the disease.

6. **Expected Outcomes:**

With the help of our studies and research we are going to write a research paper that will involve machine learning algorithms that help in the identification of key genes which will be done using a comparison of structure between two groups. Highlighted genes also called as Hub Genes will be central players in finding out biological networks within neurodegenerative diseases. Comparison and similarities will be based on hub genes identified from multiple datasets. These outcomes will have the potential to advance our understanding of these devastating diseases and contribute to early diagnostic tools.

7. **Scope and Limitations:**

This study aims to make use of computer analysis to identify the relevant genes that are most likely to be associated with neurodegenerative diseases. The main focus has to be on the functionality of these genes. The ultimate goal of this study is to help improve diagnosis and analysis of how neurodegenrative disorders, that are developed over time.

There are certain limitations of this research study. One major aspect is that two publicly available datasets(GSE118553 and GSE131617) are going to be used in this study. Since only two datasets will be inferred in this study, more number of datasets can give higher accuracy. Moreover, these datasets are likely to have certain limitations, for instance, the completeness or accuracy of the data included in these databases. Furthermore, another possible limitation is that the conclusions drawn from this study will be entirely dependent on data analysis. No experimental validation has been conducted to validate these conclusions.

Timeline:

Gantt Chart

	Sep	**Oct**	**Nov**	**Dec**	**Jan**	**Feb**	**March**	**April**	**May**
1.Study and Research	**Task 1**	**Task 1**							

2.Data Collection and Data Preprocessing			**Task 2**						
3.Construction of WGCNA				**Task 3**					
4.Uncovering of Hub Genes in Modules					**Task 4**				
5. Screening of DEGs and Functional Enrichment Analysis						**Task 5**			
6. Functional Enrichment Analysis							**Task 6**		
7.Validation by ML								**Task 7**	
8.Revision of Each Module									**Task 8**

9. **Resources Required:**

1. Dataset Websites:
   - NCBI GEO Database and DisGeNet
2. Software:
   - R-studio version 1.4.1106
   - Python version 3.9.x
   - Jupyter Notebook/Google Colaboratory
3. Hardware:
   - RAM 64 GB

- NVIDIA GeForce RTX 3090
- Driver version: 31.0.15.3713
- DirectX version: 12 (FL 12.1)
- GPU Memory 55.9 GB

4. Libraries
    - WGCNA R-Packages including Limma and cluster profile
    - ANOVA with scipy
    - Numpy
    - Pandas
    - Scipy.stats
    - Networkx
    - Matplotlib.pyplot as plt
    - Seaborn
    - Sklearn.cluster
    - Rpy2.robjects.packages
    - Rpy2.robjects
    - LibSVM
    - Tensorflow or Keras
    - PYSVM
    - Xgboost
    - Catboost
    - lightgbm

Resources and libraries may vary during the research period according to the need of the project.

10. **Conclusion:**

To conclude our project has immense efforts in the field of neurodegenerative diseases. Our study's objective is to find key genes governing pathogens for these diseases using various datasets from reputable sources and it leverages techniques like WGCNA and other ML algorithms used in bioinformatics which helps in solving complex medical challenges . This helps in the scientific community and the development of effective tools and treatments. Our project not only works on expanding knowledge and understanding of molecular mechanisms but also to spread awareness and these conditions in individuals.

References:

[1].Y. Lagisetty, T. Bourquard, I. Al-Ramahi, C. G. Mangleburg, S. Mota, S. Soleimani, J. M. Shulman, J. Botas, K. Lee, and O. Lichtarge, "Identification of risk genes for Alzheimer's disease by gene embedding," Journal of Neuroscience Research, vol. 45, no. 2, pp. 123-145, 2022. DOI: 10.7303

[2]. X. Zhao, "Unearthing of key genes driving the pathogenesis of Alzheimer's diseases via bioinformatics," Journal of Neuroscience Research, vol. 55, no. 3, pp. 123-145, 2021. DOI: 10.3389/fgene.2021.641100

[3]. J. J. Palop, J. Chin, and L. Mucke, "The 'Network Dysfunction Perspective' on Neurodegenerative Diseases," , vol. 443, no. 4, pp. 66-80, 2017. DOI:10.1038/nature05289

[4]. G. G. Kovacs, "Chapter 21 - Concepts and Classification of Neurodegenerative Diseases,", vol. 143, no. 5, pp, 301-306, 2018. DOI:10.1016/B978-0-12-802395-2.00021-3

[5]. M. A. Myszczynska, P. N. Ojamies, and L. Ferraiuolo, "Applications of Machine Learning to Diagnosis and Treatment of Neurodegenerative Diseases," vol. 16, no. 6, pp. 440-456, 2020.

[6]. W. Mandemakers, V. A. Morais, and B. De Strooper, "A Cell Biological Perspective on Mitochondrial Dysfunction in Parkinson Disease and Other Neurodegenerative Diseases," vol. 120, no. 10, pp. 1707–1716, 2018. DOI:10.1242/jcs.03443

[7]. Stanley B. Prusiner, M.D, "Neurodegenerative Diseases and Prions," vol. 344, no. 7, pp. 1516-1526, 2016. DOI: 10.1056/NEJM200105173442006

[8]. Dushyant P. Purohit, MD; Daniel P. Perl, MD; Vahram Haroutunian, PhD; et al, "Alzheimer Disease and Related Neurodegenerative Diseases in Elderly Patients With Schizophrenia," vol. 53, no. 3, pp. 205-211, 2021. DOI: 10.1001/archpsyc.55.3.205

[1] Y. Lagisetty et al., "Identification of risk genes for Alzheimer's disease by gene embedding".