A Pangenome Graph-based Approach for Predicting Alzheimer's Disease

Prathibhamol C P[a], Chandrakiran J[a], Sethulakshmi Santhosh[a], Meenakshi M[a], Akash S Menon[a], Manjusha Nair[*, b]

[a]Department of Computer Science and Engineering,Amrita School of Computing, Amrita Vishwa Vidyapeetham, Amritapuri, India.
[b]Department of Computer science and Applications, Amrita School of Computing, Amrita Vishwa Vidyapeetham, Amritapuri, India.
email:manjushanair@am.amrita.edu

ARTICLE HISTORY
Compiled June 11, 2024

ABSTRACT
Pangenomes offer novel approaches to disease prediction and a comprehensive picture of genomic variation. For studying complex diseases like Alzheimer's, understanding genetic variation across populations is crucial for elucidating disease mechanisms. Pangenome graphs are gaining popularity in genomic research due to their ability to overcome the limitations of linear reference genomes. Pangenome graphs are used in this study to extract genomic variations which are used to train a Random forest machine learning model to predict Alzheimer's disease. This study focuses on four major genes associated with the disease: APP, APOE, PSEN1, and PSEN2. This approach incorporates genetic data from these genes and provides molecular insights into illness risk. This makes earlier diagnosis possible than with more conventional techniques like MRI scans[1]. Our results show that pangenome graphs and machine learning can be used to forecast Alzheimer's disease. As we advance, improved feature selection methods and machine learning techniques can be adopted along with more precise genetic markers to improve the system further.

KEYWORDS
Pangenome graph, Alzheimer's disease, Random Forest, Machine Learning, Seqwish, Minimap, Gfautil

1. Introduction

Millions of people worldwide suffer from Alzheimer's disease - a neurological condition marked by increasing cognitive decline, memory loss, and behavioral abnormalities. Alzheimer's disease has a complex pathophysiology that involves an intricate interplay between environmental and genetic factors. Even after much research, it is still difficult to understand the complex genetic roots of Alzheimer's disease. Several important genes have been identified as major contributors to the pathophysiology of Alzheimer's disease. These consist of the Apolipoprotein E (APOE) gene, Amyloid Precursor Protein (APP) gene, Presenilin 1 (PSEN1), and Presenilin 2 (PSEN2) [2].Apoprotein E, which is essential for lipid metabolism, is encoded by the APOE gene. An important genetic risk factor for late-onset Alzheimer's disease is the APOE $\epsilon 4$ allele, which is linked to increased brain plaque deposition and decreased Aβ clearance.The

γ-secretase complex, which is necessary for cleaving amyloid precursor protein (APP) into amyloid beta (Aβ) peptides, which form amyloid plaques, the defining feature of Alzheimer's disease, is encoded by the PSEN1 and PSEN2 genes. Mutations in these genes impair the processing of APP, which increases the synthesis of harmful Aβ peptides and exacerbates neuronal dysfunction as well as cognitive decline. The precursor of Aβ peptides is encoded by the APP gene, and mutations in this gene can increase the production of Aβ and so promote plaque accumulation. Based on hereditary propensity, these genetic insights guide prospective therapies by illuminating the molecular pathways causing Alzheimer's disease [3].

Pangenome graphs have become a potent method for comprehensively displaying genomic diversity, disease mechanisms, and ultimately improve diagnosis and treatment strategies by combining data from several genomes[4–8]. Despite computing constraints, researchers are increasingly turning to pangenome graphs for their potential to enhance analysis compared to traditional linear references [4]. By capturing intricate genomic patterns such as structural variants, insertions, deletions, and single nucleotide polymorphisms (SNPs), these graphs provide a more detailed picture of the genetic landscapes linked to disorders like Alzheimer's. A promising aspect of pangenome graphs is the identification of frequented regions (FRs), which hold potential for advancing disease classification. By analyzing patterns of genetic variation within these regions, researchers can gain deeper insights into the genetic underpinnings of diseases[7]. Efforts such as GenomicKB aim to automate the analysis of genetic data, but challenges remain in integrating disparate datasets effectively [8]. Similarly, tools like DGLinker offer valuable approaches for disease gene prediction, but their accuracy relies heavily on access to comprehensive genomic data [5]. This study uses pangenome graphs to extract genetic variations which are then used by machine learning models for disease prediction. The pipeline for acquiring, processing, and evaluating genetic variation connected to Alzheimer's disease is presented in the following sections of this paper.

2. Proposed Methodology

This study uses pangenome graphs on the genetic landscape of Alzheimer's disease. The approach involves first building pangenome graphs [9]for important genes related to Alzheimer's disease (PSEN1, PSEN2, APP, and APOE) using Pairwise Alignment Format (PAF) alignment files produced by the minimap2 program. Then the tool seqwish is used to create Graphical Fragment Assembly (GFA) files. These GFA files provide comprehensive representations of genetic diversity within each gene by capturing common and unusual variants in a range of populations. Pangenome graphs are used to construct Variant Call Format (VCF) files, which include detailed information about genetic variations, including their allele frequencies and genomic positions. The extracted genetic variants from these VCF files are then used to train a Random Forest model—to predict illness status based on input DNA sequences.

The software tools used in this project include:

- Minimap2: A sequence alignment tool that aligns the DNA sequences against a reference sequence and produces PAF files.
- Seqwish: A variation graph inducer that generates pangenome graphs (GFA files) from the PAF files and input sequences.
- Gfatools: A set of tools for manipulating GFA files, such as extracting variations,

converting formats, and visualizing graphs.

- Gfautil: A command line tool for various operations on GFA and related files.
- PySAM: A Python library for working with SAM/BAM/VCF/ BCF/GFF/GTF/FASTA/FASTQ files.
- Scikit-learn: A Python library for machine learning, providing various algorithms, models, and metrics.
- Bandage: A software for visualizing and exploring pangenome graphs.

2.1. Dataset

The datasets used in this project comprise DNA sequences in FASTA format obtained from the National Center for Biotechnology Information (NCBI) [10]. For the gene PSEN1, the following accession numbers were utilized: NM_000021.4, NM_007318.3, KT120066.1, AK312531.1, DN998975.1, LC756953.1, and AF205592.1. For the gene PSEN2, the accession numbers user are: NG_007381.2, BC006365.2, DQ893826.2, NM_012486.3, NM_000447.3, and KT120068.1. The dataset for the gene APP includes the following accession numbers: NG_007376.2, AY919674.1, KT120069.1, NM_000484.4, NM_001136131.3, and NM_001385253.1. Lastly, for the gene APOE, the dataset incorporates the following accession numbers: NG_007084.2, AY077451.1, DJ359647.1, NM_001302688.2, NM_001302689.2, and NM_000041.4. All these datasets include affected and unaffected variants to capture genetic variation across different individuals and populations.

2.2. Module I: GFA and VCF File Generation

For the PSEN1, PSEN2, APP, and APOE genes, this module is first intended to provide the necessary Genome Feature Annotation (GFA) files using PAF file. The process begins with the alignment of sequences to a reference genome using the Minimap2 [11] program which produces the PAF files. Sequence alignments to a reference genome are described in detail in PAF files. This contains the locations, alignment quality scores, names of the target and query sequences, as well as other pertinent information, and hence is the basis for logging genomic variants within the targeted genes. The tool Seqwish [12] uses the produced PAF files, to create Genome Feature Annotation (GFA) files.The structural subtleties of genes like PSEN1, PSEN2, APP, and APOE are captured in these files and visual representations of pangenome graphs are also performed by this tool. Variant Call Format (VCF) files serve as a standardized format in bioinformatics, encapsulating information about genetic variants such as SNPs, insertions, deletions, and structural variations. The Gfautil tool [13] , available on GitHub, is employed for this task .

2.3. Module II: Random Forest Model Training

Feature extraction begins with a comprehensive examination of three pivotal genetic variants: SNPs (single-nucleotide polymorphisms), Indels (insertions and deletions), and CNVs (copy number variations). SNPs serve as invaluable markers, offering insights into individual base variations, including their chromosomal location, reference and alternate alleles, and quality ratings. Indels, on the other hand, categorize genetic alterations based on their chromosomal position, reference and alternate alleles, size, and classification as insertions or deletions.CNVs unveil alterations in copy numbers by elucidating genomic regions, log2 copy ratios, and other pertinent metrics.

After feature extraction, a Random Forest model [14] [15]was trained[16] as shown in Algorithm 1.

Algorithm 1 Random Forest Algorithm

```
Input: Training data - features and labels from pangenome graphs (VCF files)
Output: Random Forest model for predicting labels for new data
Algorithm Random Forest:
Create bootstrap samples by randomly selecting subsets of training data with replacement
for each bootstrap sample do
    Build a decision tree
end for
Make predictions:
Pass new data point to each decision tree in the forest
Average predictions from all decision trees
end Random Forest Algorithm
```

3. Results & Discussion

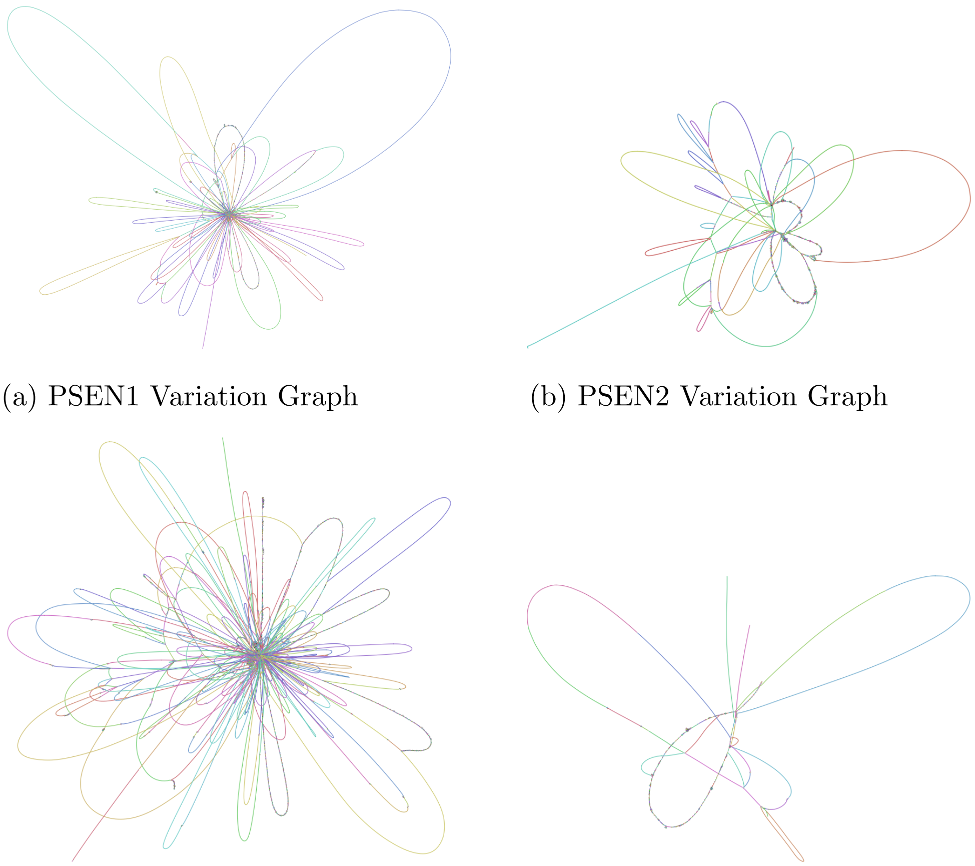
3.1. Pangenome Graph Visualization

The GFA files, encapsulating variations associated with each gene, are visualized through Bandage software, providing a comprehensive overview of the pangenome graphs for PSEN1, PSEN2, APP, and APOE. Here, each node consists of a sequence of base pairs and the edges provide the path for the connection of each node. This graphical representation(Figure 1) enhances our understanding of genetic variations within each gene and their potential implications for disease prediction.

3.2. Model Evaluation

Following the training of the Random Forest model, we conducted a comprehensive evaluation on a test set to assess its predictive performance.The training data utilized genetic variation features extracted from VCF files for the PSEN1, PSEN2, APP, and APOE genes, paired with corresponding labels denoting Alzheimer's disease status. These features encompassed information on SNPs, CNVs, and indels. The model exhibited an overall accuracy of 86%, indicating its proficiency in correctly classifying samples from the testing data.

Upon evaluation, the model demonstrated high precision, recall, and F1-score for predicting Class 0(without Alzheimer's disease), signifying its effectiveness in identifying unaffected individuals. The macro-average precision, recall, and F1-score were 92%, 75%, and 79% respectively, while the weighted average precision, recall, and F1-score were 88%, 86%, and 84% respectively, providing a comprehensive overview of the model's performance across both classes. Notably, the precision for Class 1(with Alzheimer's disease) was better, indicating that when the model predicted an individual to have Alzheimer's disease, it was almost correct. However, the recall and F1-score for Class 1 were lower, indicating that the model might need further refinement



(a) PSEN1 Variation Graph

(b) PSEN2 Variation Graph

(c) APP Variation Graph

(d) APOE Variation Graph

Figure 1. Variation Graphs Derived with Bandage

4. Conclusion

With an emphasis on Alzheimer's disease, the study combined genomic data, pangenome graphs, and machine learning to forecast genetic illnesses. With each phase, a more thorough understanding of the genetic variants linked to the disease was developed. This approach proved to be beneficial compared to the traditional way of predicting the disease using MRI scans which could be potentially used for detecting latter stages of the disease. In the ongoing study, we focus on creating a web-based tool that predicts genetic diseases using frameworks like Flask for the frontend and Vue.js for the backend [17]. With the use of this technique, researchers and physicians could enter genetic data and receive disease predictions based on pangenome graphs and machine learning models in an intuitive interface. Furthermore, by utilizing pre-trained models on sizable genomic datasets, the application of transfer learning[18] techniques can be employed to improve the performance.

References

[1] S. Sivaranjani , et al., 'Diabetes Prediction using Machine Learning Algorithms with Feature Selection and Dimensionality Reduction,' 2021 7th International Conference on

Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 2021, pp. 141-146, doi: 10.1109/ICACCS51430.2021.9441935.

[2] Lanoiselée HM, et al.; collaborators of the CNR-MAJ project. ‘APP, PSEN1, and PSEN2 mutations in early-onset Alzheimer disease: A genetic screening study of familial and sporadic cases’. PLoS Med. 2017 Mar 28;14(3):e1002270. doi: 10.1371/journal.pmed.1002270. PMID: 28350801; PMCID: PMC5370101.

[3] Andrade-Guerrero, et al. ‘Alzheimer’s Disease: An Updated Overview of Its Genetics.’ International Journal of Molecular Sciences 24, no. 4 (February 13, 2023): 3754. https://doi.org/10.3390/ijms240437.

[4] Jordan M. Eizenga, et al., ‘Annual Review of Genomics and Human Genetics’ 2020 21:1, 139-16.

[5] Harling-Lee, J.D., Gorzynski, J., Yebra, G. et al. ‘A graph-based approach for the visualisation and analysis of bacterial pangenomes’. BMC Bioinformatics 23, 416 (2022). https://doi.org/10.1186/s12859-022-0489.

[6] ‘Disease association with frequented regions of genotype graphs’. Samuel Hokin, Alan Cleary, Joann Mudge medRxiv 2020.09.25.20201640; doi: https://doi.org/10.1101/2020.09.25.20201640.

[7] Fan Feng, et al., ‘GenomicKB: a knowledge graph for the human genome’, Nucleic Acids Research, Volume 51, Issue D1, 6 January 2023, Pages D950–D956, https://doi.org/10.1093/nar/gkac957.

[8] Jiajing Hu, et al, ‘DGLinker: flexible knowledge-graph prediction of disease–gene associations’, Nucleic Acids Research, Volume 49, Issue W1, 2 July 2021, Pages W153–W161, https://doi.org/10.1093/nar/gkab449.

[9] Erik Garrison, Andrea Guarracino, ‘Unbiased pangenome graphs’, Bioinformatics, Volume 39, Issue 1, January 2023, btac743, https://doi.org/10.1093/bioinformatics.

[10] National Center for Biotechnology Information. Accessed March 27, 2024. https://www.ncbi.nlm.nih.gov.

[11] Heng Li, ‘Minimap2: pairwise alignment for nucleotide sequences’, Bioinformatics, Volume 34, Issue 18, September 2018, Pages 3094–3100, https://doi.org/10.1093/bioinformatics/bty1.

[12] Ekg. ‘EKG/Seqwish: Alignment to Variation Graph Inducer.’ ’GitHub, November 1, 2019. https://github.com/ekg/seqwi.

[13] Fischer, Christian. ‘GFAUTIL - Rust Utility.’ gfautil - Rust utility //, February 23, 2021. https://lib.rs/crates/gfauti.

[14] Ani, R., Augustine, A., Akhil, N.C., Deepa, O.S. (2016). ‘Random Forest Ensemble Classifier to Predict the Coronary Heart Disease Using Risk Factors’. In: Suresh, L., Panigrahi, B. (eds) Proceedings of the International Conference on Soft Computing Systems. Advances in Intelligent Systems and Computing, vol 397. Springer, New Delhi. https://doi.org/10.1007/978-81-322-2671-.

[15] C. Jose and G. Gopakumar, ‘An Improved Random Forest Algorithm for classification in an imbalanced dataset,’ 2019 URSI Asia-Pacific Radio Science Conference (AP-RASC), New Delhi, India, 2019, pp. 1-4, doi: 10.23919/URSIAP-RASC.2019.8738.

[16] Pattanayak, S., Singh, T. (2022). ‘Cardiovascular disease classification based on machine learning algorithms using GridSearchCV, cross validation and stacked ensemble methods’. In Communications in computer and information science (pp. 219–230). https://doi.org/10.1007/978-3-031-12638-.

[17] A. Hebbale, G. Vinay, B. V. Krishna and J. Shah, ‘IoT and Machine Learning based Self Care System for Diabetes Monitoring and Prediction,’ 2021 2nd Global Conference for Advancement in Technology (GCAT), Bangalore, India, 2021, pp. 1-7, doi: 10.1109/GCAT52182.2021.958.

[18] S. S. Rajeswari and M. Nair, ‘A Transfer Learning Approach for Predicting Alzheimer’s Disease,’ 2021 4th Biennial International Conference on Nascent Technologies in Engineering (ICNTE), NaviMumbai, India, 2021, pp. 1-5, doi: 10.1109/ICNTE51185.2021.9487.