1. Applicant's name: Junqing Peng

2. Title of proposed research project:

Multi-risk Prediction from Large Genome-Wide Data with Graph Attention Network

3. Summary of research

This research aims to build a graph-attention network model used to predict muti-risk based on large genome-wide data. Genome plays a critical role in the biological study and it is an important aspect of disease prediction. Since multi-risk disease prediction has not been studied, utilizing the advantages of the graph-attention network can improve the effectiveness and accuracy of multi-risk disease prediction using large genomic data. This research will provide more comprehensive treatment plans, leading to a higher survival rate for humans.

4. Introduction

The genome plays a critical role in the biological study, including health management (Krantz et al., 2021), genetic analysis (Martin-Trujillo et al., 2022), and disease prediction (Pan et al., 2021) and there are complete databases for genome such as the UK Biobank Datasets. In addition, thanks to the Human Genome Project, a mature study of the human genome was conducted extensively, which can be used for prediction.

However, there are great deficiencies and defects in the ability to predict human diseases. It is not entirely true that there is always a one-to-one correlation between genetic defects and diseases because there may be hidden relationships between these two. For example, neonatal inflammatory skin and bowel disease 1 (NISBD1) with complete loss of ADAM17 expression lead to erythroderma, atrichia, nail dystrophy, oesophageal strictures, and other diseases reported by a young female individual with NISBD1 (Samuelov et al., 2022). This is because of a complex compound heterozygous defect and it also consists of a large genomic deletion. Thus, these complex relationships make risk prediction using genetic defects extremely difficult.

Since most of the existing studies are targeting a single genetic defect and a single disease each time, they do not reflect all the effects of a single genetic defect on human diseases. Study like Individualized Coherent Absolute Risk Estimation models (ICARE) is used to encourage the development of a more advanced risk-prediction model, allowing more comprehensive examination of different diseases, including risk-stratified breast cancer (Pal Choudhury et al., 2019). Moreover, both genome and disease are cross-mixing relationships, showing that the evaluation of multiple diseases based on a large number of genomes is necessary.

Graph attention network is designed for graph-structured data using attention mechanism, allowing the model to use the most essential part of the graph when making the prediction. GEHGAN uses graph embedding and a heterogeneous graph attention network to predict circRNA-disease correlations effectively and efficiently by using random walks with the jump and stay strategies (Wang & Lu, 2024). Using the CircR2Diseasev2.0 database, this model has incredible performance with an AUC score of 0.9829 and an AUPR value of 0.9815. However, this is just the study for circRNA-disease instead of the large genome-wide datasets. Thus, a graph attention network can be used to solve the problem of prediction of multiple disease risks using a large number of genomes. Because of the attention mechanism, we can conduct the muti-risk prediction based on different features extracted from large genome-wide datasets.

5. Objectives and hypotheses to be tested

Aim: To address inefficiency of existing one-to-one and many-to-one risk-prediction methods, using a graph attention network, this proposal aims to develop a novel muti-risk prediction method from large genome-wide data. The objectives are:

(1) Acquisition and processing of large genome-wide data with great efficiency;

(2) Development of a new graph model based on large genome-wide data;

(3) Construction of a graph attention-neutral network method for muti-risk prediction based on the developed graph model;

(4) Verification of the effectiveness of the proposed method on multiple datasets.

Hypothesis:

(1) Large genome-wide datasets have complex graphical relationships;

(2) There are hidden effects of the relationship between genes and diseases.

6. Literature review

Large genome has been widely used in many studies. The PANTHER classification system uses genomes, gene function classifications, pathways, and statistical analysis, to make the analysis of large-scale genome-wide datasets possible (Mi et al., 2019). It uses 313 complete genomes grouped by gene families and subfamilies, evolutionary relationships between genes generated by phylogenetic trees, multiple sequence alignments, and statistical models. In addition, the study of the evolution of Phasmatodea's genome shows that the evolution is because of the increase in repetitive regions and intron elongation (Wu et al., 2017). With the use of a de novo genome assembly of a female C. hookeri, it shows candidate genes linked to gamete production and progress in both females and males. Thus, large genome is very essential to many biological studies. Because of those large numbers of genomes and different genetic variants, the risk of certain diseases will exist.

Nowadays, the risk prediction of disease can be categorized into two types of methods namely traditional statistic-based method and machine learning-based method. Using the traditional statistic-based method, an individual's phenotype from their DNA can be predicted accurately using an SBayesR model (Lloyd-Jones et al., 2019). Genome-wide association studies utilize the summary statistics to generate this model based on a Bayesian multiple regression model. Recently, built on traditional statistical models, machine learning is advancing rapidly. For instance, using the Lasso regression method, based on the AUC values, a model with 5 remaining predictors is generated to predict cardiovascular prognoses for the patient with NDD-CKD (Li et al., 2022).

However, traditional machine learning methods are unable to carry out deep specific extraction of large data sets, including human genome data. In that case, using deep-learning methods to predict disease from genome data has been rapidly developed. Amyotrophic lateral sclerosis is a heterogeneous neurodegenerative disease, which can be predicted using three deep learning models based on different architectures (Pancotti et al., 2022). These models use ALS progression based on PRO-ACT data to get better performance. In addition, by applying de novo genome reconstruction, taxonomic profiling, and deep learning-based functional annotations from DeepFRI, a new metagenome analysis workflow is introduced. (Maranga et al., 2023). It allows for a new understanding with human gut micorbiome functional features in health and disease and sets a solid foundation for future metagenomics studies.

But, these deep learning methods are unable to assess and predict diseases with complex relationships like the relationship between genomes, leading to the introduction of some learning methods based on graph modeling used to solve disease-risk prediction. These methods will consider these inherent relationships between different biological entities to produce more accurate predictions. For example, a graph theoretic-based gene selection method can find a subset of genes with the least inner similarity and the most essential aspects to the target class (Azadifar et al., 2022). It has better performance than renowned filter-based gene selection approaches. In addition, SpaGCN is a graph convolutional network that combines gene expression, spatial location, and histology in SRT data analysis (Hu et al., 2021). It has faster performance and platform independence in different SRT studies that enable the comprehensive characterization of gene expression patterns in terms of tissue microenvironment. Moreover, CellVGAE is an example of a graph neural network for the unsupervised exploration of scRNA-seq data by pulling out essential features from the data and it uses the means to visualize and interpret different features of the model (Buterez et al., 2021). It has better interpretability than scRNA-seq variational architectures and is more advanced than other competing methods. Thus, the use of learning methods based on graph modeling allows the advancement in performance and interpretability of some special characteristics and complexity of data.

However, graphs can not easily handle hidden and indeterminate graph data structures which large genome-wide data always has. Since multi-risk disease prediction has not been studied, this proposal makes use of the dynamic modeling advantages of graph-attention networks to improve the effectiveness and accuracy of multi-risk disease prediction using large genome-wide data.

7. Materials and methods

This proposal develops a graph attention network for muti-risk prediction with large genome-wide datasets including graph modeling, model training, and model evaluation. The whole architecture is illustrated as follows:

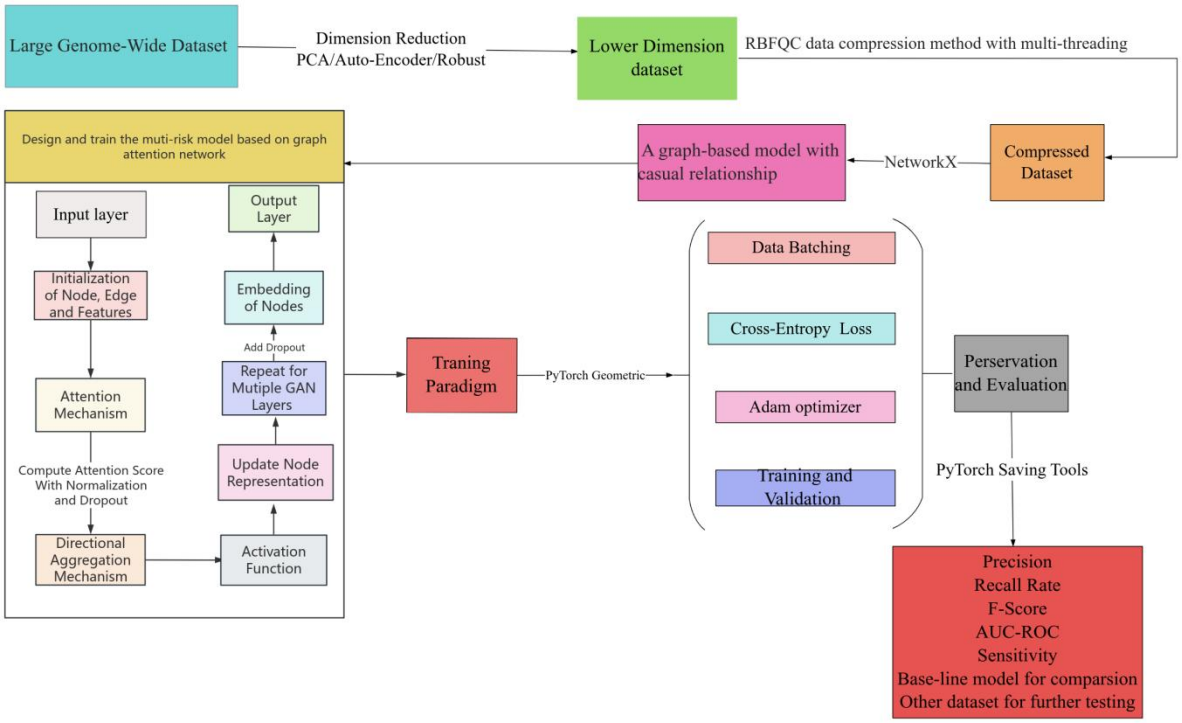

Figure 1 The overview of graph attention network method used for muti-risk prediction

A. Data Collection and Processing

There are some large genome-wide datasets, including the Genome-Wide Association Studies (GWAS), the Biological interaction data, and the UK Biobank Datasets. The UK Biobank Datasets (https://www.ukbiobank.ac.uk) can be used in our research since the datasets have useful information including chronic conditions and genetic data. These data are measured from people with different sex, ages (40-69), and locations. Moreover, there are many risks including genetic risk, cardiovascular risk, and cancer risk in the datasets.

There are many challenges when using these large genome-wide datasets so data processing is required. UK Biobank Datasets are high dimensional data, with about 500,000 participants, 800,000 directly genotyped SNPs per person, and millions of genetic variants. It will cause over-fitting issues and difficulties in identifying genetic variations. Hence, reducing the dimension of the datasets is critical. These unsupervised learning methods, including autoencoders, robust, and classical principal component analyses can be used for dimension reduction (Hassan et al., 2023).

In addition, the sizes of these data files are very big, making them difficult to read and save. RBFQC data compression method is suggested since it is 10-25% faster than domain-specific FastQ file referential genome compression techniques to reduce time, data processing, transmission, and storage costs (Kumar et al., 2023). Moreover, to ensure performance, multi-threading computing can be used (Zou et al., 2021).

B. Building a Genome-Wide Graph-based Model

Using NetworkX, which is a Python library, a graph-based model can be designed (NetworkX, 2024). Moreover, to generate a useful and efficient graph-based model, the causal relationship between different aspects of large genomes and different diseases should be defined precisely. The relationship within the graph including nodes, edges, attributes, subgraphs, and direct graph type should be defined clearly. Moreover, use a suitable algorithm to generate a suitable graph, and an appropriate visualization. For instance, the local genetic analysis results show that there is a cause-effect relationship between GIT (gastrointestinal tract) and cognitive traits in the genome (Adewuyi et al., 2022). GIT and cognitive traits can be defined as the nodes with weights and the edges represent the cause-effect relationship with direction.

C. Design and Train the Multi-risk Prediction Model based on Graph Attention Network

Since the muti-risk prediction model has been represented as a graph, each aspect of the graph should represent meaningful information. Zhu showed a detailed setup of directional graph attention networks (Zhu, 2023) and these ideas can be adapted to this problem. The input of each node represents a genomic feature, such as CDS and mRNA and edges represent the relationships using weight scores to measure strength. In addition, the output should use a suitable representation, such as 0 and 1, with a suitable threshold, to indicate the presence of diseases and the risk probabilities.

Different parameters should be used to make the model work. Graph Attention Network (GAT) model utilizes the muti-head attention approach so that number of the attention heads should be defined. With the use of an

attention mechanism with normalization and dropout, a directional aggregation mechanism, and an activation function such as Relu, the node representation can be updated and we can also repeat the process for many graph attention network layers. After embedding the nodes, the output layers can be generated for further classification.

In addition, a loss function such as a cross-entropy loss function for each prediction will be used, and using an Adam optimizer can make training efficient with a default learning rate. Meanwhile, using sub-graphs as mini-batches in order to achieve more efficient memory usage, more advanced generalization, better parallelism, and stochasticity (GeeksforGeeks, 2024). Moreover, the datasets are categorized into training sets, validation sets ,and test sets, and during the training, use the early stopping strategy to avoid overfitting. PyTorch geometric is a deep learning package on graphs using different methods and by using an advanced GPU, it makes the operations of mini-batch loaders on many small and single giant graphs possible (PyG, 2024). To save the model, we choose the model with the best performance and no overfitting, and pyTorch has its saving tools.

D. Model Evaluation

Designing appropriate evaluation metrics and methods is required, which includes precision, recall rate, the choice of baseline models, and the choice of the other datasets. The model with a high precision and a high recall rate is preferred. Moreover, the visualization of the loss curves will let us have a better understanding of the performance of the model. In addition, the expected model should have a high F-score, a reasonable sensitivity, and an AUC-ROC which is close to 1.

Moreover, baseline models with strong performance when dealing with complex datasets like Bayes should be used in order to compare the results. Also, we need to use other large datasets to test the model with the same metrics as above. Furthermore, those models can be used to predict the disease using a validation dataset such as a subset of UK Biobank Datasets. Then, we use the models to predict the disease and compare the accuracy. We expect our model to have higher accuracy than the baseline models.

To evaluate the model based on its robustness and generalizability, we can add noise such as Gaussian noise to the model. Also, the noise can also be applied to the node and the data. Moreover, we can compare the performance of our model with noise and the performance of the baseline model with noise. We expect the performance of our model to be better even under noisy conditions.

8. Anticipated outcome and value of the research

A graph-attention model is built to predict multi-risk based on large genome-wide data with great efficiency and accuracy. This model can predict multi-risk diseases, which will provide more comprehensive treatment plans, leading to a higher survival rate. In addition, this model also provides the probability of each disease based on different genome features.

This model provides multi-risk prediction which is more advanced compared to previous studies. First of all, this model provides a solution that will deal with the complex relationship of large genome-wide datasets and the hidden effects of the relationship between genes and diseases. At the same time, during the testing, an efficient way to reduce the dimension of large genome-wide data will be found, speeding up the future workflow. Moreover, this model will allow us to have a better understanding of the multiple effects of a single gene, leading to further study and testing. In the future, our model will continue to improve its accuracy and implement it into different medical hardware in about 3 to 5 years.

Based on the research proposed, I would like to make a breakthrough in the large-scale genome-wide data analysis and risk prediction for humankind's health. The data and case studies involve diverse genders, ages, and districts, helping to understand the culture and human behavior diversities in depth. The outcome of the research has the potential to be published in top-tier journals such as Nature, Science, etc. Besides, my study and living background across from China mainland and the USA will contribute to the research and development communities in Hong Kong, hence, bringing cultural diversity and research collaboration among different districts.

9. Literature cited

Krantz, I. D., Medne, L., Weatherly, J. M., Wild, K. T., Biswas, S., Devkota, B., Hartman, T., Brunelli, L. Fishler, K. P., Abdul-Rahman, O., Euteneuer, J. C., Hoover, D., Dimmock, D., Cleary, J., Farnaes, L., Knight, J., Schwarz, A. J., Vargas-Shiraishi, O. M., Wigby, K., … Taft, R. J. (2021). Effect of whole-genome sequencing on the clinical management of acutely ill infants with suspected genetic disease. JAMA Pediatrics, 175(12), 1218. https://doi.org/10.1001/jamapediatrics.2021.3496

Martin-Trujillo, A., Garg, P., Patel, N., Jadhav, B., & Sharp, A. J. (2022). Genome-wide evaluation of the effect of short tandem repeat variation on local DNA methylation. Genome Research, 33(2), 184–196. https://doi.org/10.1101/gr.277057.122

Pan, Z., Yao, Y., Yin, H., Cai, Z., Wang, Y., Bai, L., Kern, C., Halstead, M., Chanthavixay, G., Trakooljul, N., Wimmers, K., Sahana, G., Su, G., Lund, M. S., Fredholm, M., Karlskov-Mortensen, P., Ernst, C. W., Ross, P., Tuggle, C. K., … Zhou, H. (2021). Pig genome functional annotation enhances the biological interpretation of complex traits and human disease. Nature Communications, 12(1).

https://doi.org/10.1038/s41467-021-26153-7
Samuelov, L., Sarig, O., Malovitski, K., Bergson, S., Meijers, O., Shouval, D. S., & Sprecher, E. (2022). Neonatal inflammatory skin and bowel disease type 1 caused by a complex genetic defect and responsive to combined anti-tumour necrosis factor-α and interleukin-12/23 blockade. British Journal of Dermatology, 186(6), 1026–1029. https://doi.org/10.1111/bjd.20978
Pal Choudhury, P., Wilcox, A. N., Brook, M. N., Zhang, Y., Ahearn, T., Orr, N., Coulson, P., Schoemaker, M. J., Jones, M. E., Gail, M. H., Swerdlow, A. J., Chatterjee, N., & Garcia-Closas, M. (2019). Comparative validation of breast cancer risk prediction models and projections for future risk stratification. JNCI: Journal of the National Cancer Institute, 112(3), 278–285. https://doi.org/10.1093/jnci/djz113
Wang, Y., & Lu, P. (2024). GEHGAN: CircRNA–disease association prediction via graph embedding and heterogeneous graph attention network. Computational Biology and Chemistry, 110, 108079. https://doi.org/10.1016/j.compbiolchem.2024.108079
Mi, H., Muruganujan, A., Huang, X., Ebert, D., Mills, C., Guo, X., & Thomas, P. D. (2019). Protocol update for large-scale genome and gene function analysis with the Panther Classification System (v.14.0). Nature Protocols, 14(3), 703–721. https://doi.org/10.1038/s41596-019-0128-8
Wu, C., Twort, V. G., Crowhurst, R. N., Newcomb, R. D., & Buckley, T. R. (2017). Assembling large genomes: Analysis of the Stick Insect (clitarchus hookeri) genome reveals a high repeat content and sex-biased genes associated with reproduction. BMC Genomics, 18(1). https://doi.org/10.1186/s12864-017-4245-x
Lloyd-Jones, L. R., Zeng, J., Sidorenko, J., Yengo, L., Moser, G., Kemper, K. E., Wang, H., Zheng, Z., Magi, R., Esko, T., Metspalu, A., Wray, N. R., Goddard, M. E., Yang, J., & Visscher, P. M. (2019). Improved polygenic prediction by bayesian multiple regression on summary statistics. Nature Communications, 10(1). https://doi.org/10.1038/s41467-019-12653-0
Li, N., Wang, Z., Yang, X., Xie, H., Gu, Q., Guo, J., & Li, Z. (2022). Development and validation of a cardiovascular disease risk prediction model for patients with non-dialysis-dependent chronic kidney diseases based on the nomogram. Kidney and Blood Pressure Research, 48(1), 7–17. https://doi.org/10.1159/000527856
Pancotti, C., Birolo, G., Rollo, C., Sanavia, T., Di Camillo, B., Manera, U., Chiò, A., & Fariselli, P. (2022). Deep learning methods to predict amyotrophic lateral sclerosis disease progression. Scientific Reports, 12(1). https://doi.org/10.1038/s41598-022-17805-9
Maranga, M., Szczerbiak, P., Bezshapkin, V., Gligorijevic, V., Chandler, C., Bonneau, R., Xavier, R. J., Vatanen, T., & Kosciolek, T. (2023). Comprehensive functional annotation of Metagenomes and microbial genomes using a deep learning-based method. mSystems, 8(2). https://doi.org/10.1128/msystems.01178-22
Azadifar, S., Rostami, M., Berahmand, K., Moradi, P., & Oussalah, M. (2022). Graph-based relevancy-redundancy gene selection method for cancer diagnosis. Computers in Biology and Medicine, 147, 105766. https://doi.org/10.1016/j.compbiomed.2022.105766
Hu, J., Li, X., Coleman, K., Schroeder, A., Ma, N., Irwin, D. J., Lee, E. B., Shinohara, R. T., & Li, M. (2021). SPAGCN: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. Nature Methods, 18(11), 1342–1351. https://doi.org/10.1038/s41592-021-01255-8
Buterez, D., Bica, I., Tariq, I., Andrés-Terré, H., & Liò, P. (2021). Cellvgae: An unsupervised scrna-seq analysis workflow with graph attention networks. Bioinformatics, 38(5), 1277–1286. https://doi.org/10.1093/bioinformatics/btab804
Hassan, A. Z., Ward, H. N., Rahman, M., Billmann, M., Lee, Y., & Myers, C. L. (2023). Dimensionality reduction methods for extracting functional networks from large-scale crispr screens. Molecular Systems Biology, 19(11). https://doi.org/10.15252/msb.202311657
Kumar, S., Singh, M. P., Nayak, S. R., Khan, A. U., Jain, A. K., Singh, P., Diwakar, M., & Soujanya, T. (2023). A new efficient referential genome compression technique for FASTQ files. Functional & Integrative Genomics, 23(4). https://doi.org/10.1007/s10142-023-01259-x
Zou, Y., Zhu, Y., Li, Y., Wu, F.-X., & Wang, J. (2021). Parallel Computing for Genome Sequence Processing. Briefings in Bioinformatics, 22(5). https://doi.org/10.1093/bib/bbab070
NetworkX documentation. NetworkX. (n.d.). https://networkx.org/
Bean, D. M., Heimbach, J., Ficorella, L., Micklem, G., Oliver, S. G., & Favrin, G. (2019). Correction: Esyn: Network building, sharing and publishing. PLOS ONE, 14(1). https://doi.org/10.1371/journal.pone.0204058
Adewuyi, E. O., O'Brien, E. K., Porter, T., & Laws, S. M. (2022). Relationship of Cognition and Alzheimer's Disease with Gastrointestinal Tract Disorders: A Large-Scale Genetic Overlap and Mendelian Randomisation Analysis. https://doi.org/10.21203/rs.3.rs-2191133/v1
Zhu, J. (2023). Directional Graph Attention Networks. https://doi.org/10.20944/preprints202309.0962.v1
GeeksforGeeks. (2024, February 16). Why mini batch size is better than one single "batch" with all training data? https://www.geeksforgeeks.org/why-mini-batch-size-is-better-than-one-single-batch-with-all-training-data/
PyG. PyG Documentation - pytorch_geometric documentation. (2024). https://pytorch-geometric.readthedocs.io/en/latest/