Justin Patterson, Jared Coffsky, Farhib Ahsan

Problem Description

Type II Diabetes Mellitus (T2D) is caused by both hereditary and environmental factors.(1) Among the hereditary factors, some are almost certainly encoded in the genome because concordance rates for the disease differ between monozygotic and dizygotic twin pairs.(2–4) These widely varying rates are seen in the United States, for example, where it is estimated that nearly 50% of Pima Indians have T2D(5) while the prevalence among the entire US population is thought to be closer to 10%.(6)

While the complete etiology of T2D remains only partially understood, various biological networks have been implicated in the development of the disease and in the response of the body to various anti-diabetic drugs. However, it is unknown how genetic diversity across human populations contribute to biological network differences that subsequently increase susceptibility to the disease. Previous work has mainly focused on mutations in single genes and has not addressed the problem on a network level.

Focusing specifically on those networks whose nodes are genetic loci and proteins, we ask the question: Can the variation in T2D prevalence by genetic ancestry be explained by variations in the biological networks of different human populations? We hypothesize that entire biological networks associated with T2D have evolved differently in different human populations. This diversity on the network level can perhaps explain the variation in T2D prevalence by genetic ancestry.

Related Work

Predicting ethnicity from genetic data using machine learning approaches is a relatively novel, but not completely unheard of, problem. Lee et al. (7) utilized Support Vector Machines to infer course ethnic information from mitochondrial DNA sequences; Yuan et al (8) utilized logistic regression with an elastic net penalty (GLMNET) to predict ethnicity from placental DNA methylation data, while also investigating approaches using nearest shrunken centroids (NSC), K-nearest neighbors (KNN) and support vector machines (SVM); and Govender et al (9) applied a hybrid approach combining SVM, Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), and Random Forest (RF) to predict genetic relatedness using mtDNA hypervariable region 1 sequences.

Our approach seeks to predict ethnicity from a much broader dataset derived from the 1000 Genomes Project (10) formatted in a graphical structure to encode interdependence of features. To this end, we aim to use a Message Passing Graphical Neural Network approach, a combination of approaches pioneered by Duvenaud et al. (11), Li et al. (12), and Gilmer et al. (13). We also aim to provide explainability to our model utilizing Captum, a library aimed at improving interpretability of machine learning approaches developed by Kokhlikyan et al. (14).

Justin Patterson, Farhib Ahsan, Jared Coffsky

Data

Data will be collected in three stages:

Stage 1

A reasonable number of genetic loci and proteins associated with T2D will be mined from the literature. Particular emphasis will be given to entities that the authors hypothesize are component causes of T2D and are not merely associated with its development or aftereffects. The entities should also be involved in a small number of pathways rather than scattered among many different pathways. Primary focus will be given to entities associated with oxidative phosphorylation such as mitochondrial complex I and the gene INSR which codes for insulin receptor protein. This literature search will make use of tools such as the BioGRID (15), WikiPathways (16), NCBI (17), EMBL-EBI (18), and Open Targets (19). All of the entities that meet the shortlist will be coded into a graph data structure suitable for machine learning with the genetic loci and proteins as nodes and the interactions among them as edges. Some of the edges may be directed and some may be undirected. Also, some of the edges may be weighted to show the strength/evidence for the interaction.

Stage 2

Personal genomic data will be obtained from the publicly available 1000 Genomes (20) Phase 3 Reanalysis with DRAGEN 3.7 located on the Registry of Open Data on AWS. (21) In this reanalysis, Illumina uses hg38 as its human reference genome and alt-aware mapping designed to increase the sensitivity and specificity of variant calls. (22) Figure 5 shows an example of the files that are available for each participant. The hard-filtered VCF files (shown in Figure 6) will constitute the bulk of our data.

Because the collection of data is very large, the files will be downloaded and put through a pipeline one at a time for each of the participants.

Stage 3

Each VCF will be annotated with gene names so that variants can be linked back to nodes in the graph. The annotations will include GC ratios for the ref and alt alleles, type of variation, and clinical relevance of the variation. Then, all of these annotations will be summarized for each gene. For example, the number of variants in introns, the number variants in exons, the harmonic mean of the GC ratios, and the total number of pathogenic variants, benign variants, and variants of uncertain significance will be calculated. These summary metrics will be used as features associated with the nodes in our graph. The final result of our data curation efforts will be graph data structures suitable for PyG.

Methods

The graphs will be embedded in $\Re^d$ using PyG. These embeddings will be classified with a message-passing GNN, with the ethnicity of the participant as the target. The specific details of

the model - parameters, depth, etc - we will determine holistically, driven by lessons learned from processing our data. We will utilize the Captum Pytorch library to improve the interpretability/explainability of our model at prediction time.

Evaluation Plan

The main criteria we will be evaluating is the effectiveness of the classification algorithm. To do this, we will designate some of the biological networks in our dataset as “test data” and predict their ethnicities using our model. We will then apply a variety of measures to quantify the effectiveness: the percent of correct classifications within each ethnicity, along with precision, recall, and the F1 score. Further, we will analyze the data through visualizations, generating confusion matrices and potentially even scatter plots along some of the most important features (potentially revealing clusters within the dataset). It is important to note that to be successful, not only must the model be built effectively with well-tuned parameters, but also “ethnicity” must create meaningful differences between data entries. As such, if accuracy is low, we will also create confusion matrices using other attributes to label the data.

Contributions

Justin will be responsible for retrieving and formatting the data into a form suitable for PyG. He will also contribute code on GitHub, help with the interpretation of the results, and help with the final report. Farhib will be mainly responsible for building the neural network. Jared will also help with constructing the neural network, along with analyzing the results and writing the report.

Back-up Plan

If necessary, we will dispense with the graph structure. All of the features in the nodes will be combined across graphs. Ethnicity will be predicted using a non-graphical neural network. The hope is that the graph structure will provide additional information that will perform better than using this back-up approach.

Justin Patterson, Farhib Ahsan, Jared Coffsky

Bibliography

1. Prasad RB, Groop L. Genetics of type 2 diabetes-pitfalls and possibilities. Genes (Basel). 2015 Mar 12;6(1):87–123.
2. Poulsen P, Kyvik KO, Vaag A, Beck-Nielsen H. Heritability of type II (non-insulin-dependent) diabetes mellitus and abnormal glucose tolerance--a population-based twin study. Diabetologia. 1999 Feb;42(2):139–45.
3. Medici F, Hawa M, Ianari A, Pyke DA, Leslie RD. Concordance rate for type II diabetes mellitus in monozygotic twins: actuarial analysis. Diabetologia. 1999 Feb;42(2):146–50.
4. Florez JC, Hirschhorn J. The inherited basis of diabetes mellitus: implications for the genetic analysis of complex traits. Annual review of genomics …. 2003;
5. Bennett PH, Burch TA, Miller M. Diabetes mellitus in American (Pima) Indians. Lancet. 1971 Jul 17;2(7716):125–8.
6. CDC. Prevalence of Diagnosed Diabetes | Diabetes | CDC [Internet]. Prevalence of Diagnosed Diabetes. 2019 [cited 2023 Mar 31]. Available from: https://www.cdc.gov/diabetes/data/statistics-report/diagnosed-diabetes.html
7. Lee C, Măndoiu II, Nelson CE. Inferring ethnicity from mitochondrial DNA sequence. BMC Proc. 2011 May 28;5 Suppl 2:S11.
8. Yuan V, Price EM, Del Gobbo G, Mostafavi S, Cox B, Binder AM, et al. Accurate ethnicity prediction from placental DNA methylation data. Epigenetics Chromatin. 2019 Aug 9;12(1):51.
9. Govender P, Fashoto SG, Maharaj L, Adeleke MA, Mbunge E, Olamijuwon J, et al. The application of machine learning to predict genetic relatedness using human mtDNA hypervariable region I sequences. PLoS ONE. 2022 Feb 18;17(2):e0263790.
10. Fairley S, Lowy-Gallego E, Perry E, Flicek P. The International Genome Sample Resource (IGSR) collection of open human genomic variation resources. Nucleic Acids Res. 2020 Jan 8;48(D1):D941–7.
11. Duvenaud D, Maclaurin D, Aguilera-Iparraguirre J, Gómez-Bombarelli R, Hirzel T, Aspuru-Guzik A, et al. Convolutional Networks on Graphs for Learning Molecular Fingerprints. arXiv. 2015;
12. Li Y, Tarlow D, Brockschmidt M, Zemel R. Gated Graph Sequence Neural Networks. arXiv. 2015;
13. Gilmer J, Schoenholz SS, Riley PF, Vinyals O, Dahl GE. Neural Message Passing for Quantum Chemistry. arXiv. 2017;
14. Kokhlikyan N, Miglani V, Martin M, Wang E, Alsallakh B, Reynolds J, et al. Captum: A unified and generic model interpretability library for PyTorch. arXiv. 2020;
15. Stark C, Breitkreutz B-J, Reguly T, Boucher L, Breitkreutz A, Tyers M. BioGRID: a general repository for interaction datasets. Nucleic Acids Res. 2006 Jan 1;34(Database issue):D535-9.

16. Martens M, Ammar A, Riutta A, Waagmeester A, Slenter DN, Hanspers K, et al. WikiPathways: connecting communities. Nucleic Acids Res. 2021 Jan 8;49(D1):D613–21.

17. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2006 Jan 1;34(Database issue):D173-80.

18. Kanz C, Aldebert P, Althorpe N, Baker W, Baldwin A, Bates K, et al. The EMBL nucleotide sequence database. Nucleic Acids Res. 2005 Jan 1;33(Database issue):D29-33.

19. Ochoa D, Hercules A, Carmona M, Suveges D, Baker J, Malangone C, et al. The next-generation Open Targets Platform: reimagined, redesigned, rebuilt. Nucleic Acids Res. 2023 Jan 6;51(D1):D1353–9.

20. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. Nature. 2015 Oct 1;526(7571):68–74.

21. Illumina. 1000 Genomes Phase 3 Reanalysis with DRAGEN 3.5 and 3.7 - Registry of Open Data on AWS [Internet]. 1000 Genomes Phase 3 Reanalysis with DRAGEN 3.5 and 3.7. 2021 [cited 2023 Mar 31]. Available from: https://registry.opendata.aws/ilmn-dragen-1kgp/

22. Illumina. Illumina DRAGEN Bio- IT Platform 3.7 [Internet]. Illumina DRAGEN Bio- IT Platform 3.7 User Guide. 2020 [cited 2023 Mar 31]. Available from: https://support.illumina.com/content/dam/illumina-support/documents/documentation/software_documentation/dragen-bio-it/Illumina-DRAGEN-Bio-IT-Platform-User-Guide-1000000141465-00.pdf