Predicting Genomic Anomalies with Generational Genomic Signals

1st Nathan Theng
Department of Computer Science
California State University, Fresno
Fresno, California
theng_nathan@mail.fresnostate.edu

1st Hassoon Sarwar
Department of Biology
California State University, Fresno
Fresno, California
hsarwar@mail.fresnostate.edu

2nd Mario Banuelos
Department of Mathematics
California State University, Fresno
Fresno, California
mbanuelos22@mail.fresnostate.edu

The human genome consists of sequences of nucleotides (A,T,G,C) spanning billions of base pairs. Due to the evolutionary processes of mutation, variation may be present between different genomics samples in the form of single nucleotide variants (SNVs) or observable differences spanning a larger region of DNA known as structural variants (SVs). SVs are found in the form of deletions, inversions, insertions, and duplications. Although the prevalence of such genomic variation promotes genetic diversity, it can also be associated with genetic diseases such Autism and Alzheimer's disease. SVs can potentially impact gene expression, resulting in altered phenotypes and possibly cancer, making it very important to develop accurate detection models.

In our work, we implement both neural network and binary classification machine learning models and assess their ability to identify the presence of a true deletion within an individual's genome. We compare the predictive analytics of a Convolutional Neural Network (CNN) and Logistic Regression in identifying such variation using genomics signals of related family members. We utilize a low-quality dataset that holds information about the number of DNA fragments at specific genomics locations. The data used for our study was obtained from the 1000 Genomes Project and includes information about a three-generation, 17-member family.

The dataset consists of the observation signals, the number of fragments supporting a potential SV, at various locations as well as the binary truth signals that correspond to each respective genomic location. Based on the genetic alteration, the truth signal indicates if a true variation is present as seen in Figure 1.

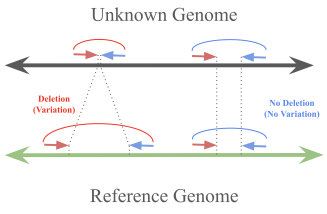

Fig. 1. Illustration of Structural Variants, where the left reads do not align to the reference genome (indicating a potential deletion) and the right reads align with the reference genome.

Our work utilized the observation signals of nine different individuals at a specific location to predict an individual's variation and compared it to the truth signal at the same specific location. We created three different cases that included different combinations of the nine related individuals to explore generational relationships. Case A included only children (Individuals 79, 80, 81, 82, 83, 86, 87, 88, 93). Case B included two parents and seven children (Individuals 77, 78, 79, 80, 82, 83, 86, 87, 88). Case C included four grandparents, two parents, and three children (Individuals 89, 90,91, 92, 77, 78, 82, 83, 87). An example of Case C is seen in Figure 2. Each case was used to predict a target individual: a grandparent (Individual 90), a parent (Individual 77), or a child (Individual 84).

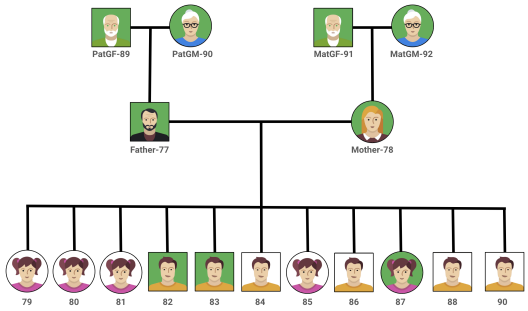

Fig. 2. Pedigree of the family sequenced for this study (CEPH pedigree 1463). This pedigree highlights Case C, individuals from all three generations (in green) we used to make predictions for our models.

As there are roughly 900,000 genomic locations for each individual, but less than 1% are true variations, undersampling was implemented to balance the dataset to achieve unbiased testing. Since SVs are rare, we undersample the signal and only consider 538 genomic locations for each individuals. An example of this is seen for the Paternal Grandmother (Individual 90) in Figure 3. A logistic regression model and a convolutional neural network (CNN) model were created via sklearn and PyTorch, respectively. For the CNN, we converted the observational signals into grayscale 3x3 images, seen in Figure 4.

Each image, representing one potential SV location, goes through a series of convolutions to produce a binary classifi-

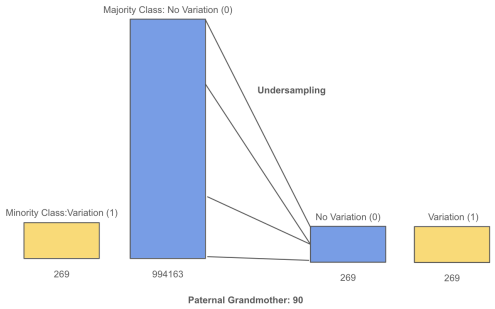

Fig. 3. Undersampling for Paternal Grandmother (90)

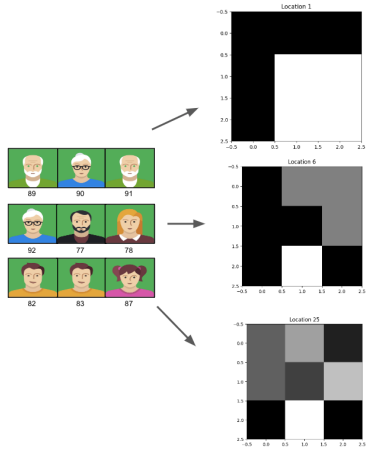

Fig. 4. Conversion of Observation Signals from Set C to 3x3 Images

cation prediction that is compared to the original truth value. The CNN follows three convolutional layers set then to a 24-node dense layer and finally to one output node. This can be seen in the schematic seen in Figure 5.

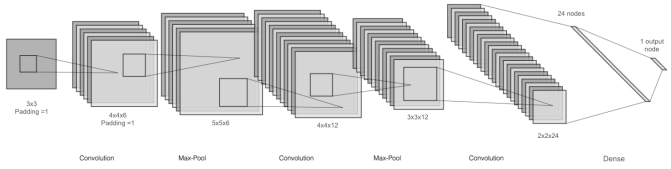

Fig. 5. Schematic of Convolutional Neural Network

Both models ran independently for three different sets: Case A, Case B, and Case C. Within each set, the model predicted three separate individuals (Individuals 90, 77, and 84). The CNN model ran 5 trials for each individual in each case and the average metrics were calculated. The accuracy for both models was usually within 70% to 80%, except for the CNN model for Case A – Individual 84 prediction. The AUC value for the logistic regression model was consistently greater than that of the CNN. This indicates that the logistic regression model was better at predicting the presence of a true genomic variation when using familial data. Below are the results from Case C which was most accurate case of the three indicating that multi-generational data is more predictive than using any one single generation.

Case C	Accuracy	Precision	Recall	F1	AUC
Pat-GM (90)	0.7963	0.7941	0.6426	0.7105	0.8409
Father (77)	0.8056	0.7838	0.6905	0.7342	0.8494
Son (84)	0.7778	0.7368	0.6667	0.7000	0.8586

TABLE I
CASE C LOGISTIC REGRESSION RESULTS

Case C	Accuracy	Precision	Recall	F1	AUC
Pat-GM (90)	0.7614	0.7095	0.6523	0.6786	0.7498
Father (77)	0.7907	0.7721	0.6619	0.7112	0.7673
Son (84)	0.7703	0.7791	0.6523	0.6884	0.7489

TABLE II
CASE C CONVOLUTIONAL NEURAL NETWORK RESULTS

Our future work will focus on optimizing the hyperparameters and including a validation split for the CNN model as well as exploring a transformer architecture. We will enhance our logistic regression model to simultaneously predict SVs for multiple individuals. We will also explore different grouping of individuals to further understand inheritance patterns of SVs.