Richard Truong-Chau
Capstone Project

Scenario

As technology advances, researchers are able to identify and characterize more genetic data from different organisms and viruses than ever. However, with the vast amount of data present, it is difficult to know exactly what type of genetic data you have. Some genetic codes may represent proteins, mitochondrial DNA, mutations, gene expressions, chloroplast DNA, Archaea, etc.

This database will be focusing on the underlying variants in the human genome that may or may not cause disease in an individual and is intended for people interested in their genetic make-up to make informed decisions. Once the genetic data is extracted from the individual the DNA is sequenced and inputted into the database. The database when completed would be able to characterize an individual's genotype and could be queried to find the probability of disease and the risk of passing down the disease to offspring.

First, it is important to record information about the individual such as their age, ethnicity, sex, occupation, address, phone and date of test. This information is vital as genetic data changes with age, environment, and other stimulus. If the individual does come back to retest, add the new date to date of test. Date of test should be a multivalued attribute. Family members who have been tested should also be recorded.

To characterize the genetic data, the genome must be recorded, and accuracy must be analyzed. If accuracy is not high, then there may be a mistake in the processing. If there is a known region in the genetic data, it can be classified as a genetic marker. Genetic markers are observable variants that arise from mutations or alterations and can be found in alleles. Markers also give the location in the genome where a gene exists.

Genes contain specific instructions on what products such as amino acids should be made, the amount, and regulation. Genes have genotypes that is heritable and can be expressed in a phenotype. However, genes in the DNA can be altered either by the environment or by random events. In general, there are two specialized groups of variations humans can receive in their DNA: Epigenetics and Mutations.

Mutations are changes to the nucleotides such as such as deletion, insertion, and insertions, while epigenetics is a change in gene expression due to modifications to histones. So, there is no actual change in nucleotides for epigenetics, yet these modifications can still be passed onto offspring.

The mutations lead to changes in codons, which in turn codes for different amino acid. Amino acids are the building blocks of protein, so different amino acids would affect the protein that is made leading to various repercussions in the body.

Markers describe regions of the gene or a sequence of DNA or RNA that has a known function. For each gene the name of the gene product, nucleotide composition, and position should be recorded.

Since markers give known locations, you can create genetic maps that can then be used to identify other genes and the distance between the two genes. Based on the distance between two genes you can rank how likely the genes would be passed down to the next generation.