# Predicting Deleterious Mutations

By: Jason Sy

University of Washington Bothell

B Bio 340 – Intro to Bioinformatics

## Abstract

Mutations are changes in DNA sequence of a particular gene in an organism that can be harmful, beneficial, or neutral in their effect on cell functions. Mutations that put an individual at risk of developing certain genetic disorders or diseases are known as deleterious mutations. Some common types of mutations are missense, frameshift and silent. The purpose of this paper is to understand more about genome sequencing techniques and how they can predict diseases. Current Methods include simulations and locating site of changes. This is done by using existing mapped genomes created as a reference genome that is roughly accurate for every human. The method used in this paper will include a FASTA parser, code to transcribe, translate, and having dictionaries for codon table, and chemical class. This is to test if differences in protein structure contribute to diseases. Results indicate that mutations such as missense and frameshift are more likely to be deleterious. For the missense testing of chemical classes, there were not any significant indicators that changes in chemical class contributed to more likelihood of diseases. However, there is yet to be a conclusion on the hypothesis due to limited sample size.

## Introduction

Mutation are random changes in the structure of a gene, that can result in variant forms transmitted to subsequent generations. They are random in the sense that they can't be predicted exactly when and where a specific mutation will occur. There are many different ways for mutation to occur and certain mutations are deemed deleterious. Deleterious mutations are changes in DNA sequences that puts an individual at risk of developing certain genetic disorder or disease, such as cancer. Which brings to question: To what extent are we able to predict deleterious mutations?

Charlesworth (2012) discusses mutation in evolution and the effects of deleterious mutations. Mutations happen all the time and can confer an advantage to the organism, at times the mutation is heritable, with the effect of increasing fitness and gets passed on. However, there are other mutations which cause a deleterious effect.

The article by Ruan, et al. discusses usage of simulations to predict traits and disease risk. Ruan, et al. goes into detail about how polygenic risk scores (PRS), have help measure the overall genetic liability to a trait or disease. Ruan, et al. cites that from simulation studies, PRS has improved cross population prediction over existing methods across traits with varying genetic architectures (Ruan et al 2022).

Current Methods include simulations and locating site of changes. This is done by using existing mapped genomes created as a reference genome that is roughly accurate for every human. Basically, it acts as a scaffold upon which you can place any individual sequence you might read. And by doing so you can sequence a person to tell if the person has any small mutations. Some genomes do not have many references, so scientist continue to assemble new reference genomes and reassemble the human genome to look for larger variants that don't agree with the original reference genome. (Collins 1995)

I predict that mutations such as frameshift and missense are more likely to have deleterious effects than mutations like silent mutations because the structure of the protein changes. To test if we're able to predict diseases based on changes in protein structure, we'll be using Python code to determine mutation type ranging from silent, frameshift and missense.

## Methods

For our code, we created a function that converts the nucleotide to protein utilizing a dictionary that contains a codon table. After converting nucleotide to protein, we can see what amino acids changed in the sequence, and from here we can call a function to get the mutation type. A similarity test is also performed on the protein to see the match percentage. Code includes:

- FASTA parser
- Code to transcribe DNA
- Code to translate RNA
- Code with dictionaries for codon tables, and chemical class.

Refer to https://github.com/Yellowbush/Project_2 for example codes.

Data gathered from: https://www.uniprot.org/uniprotkb?query=(taxonomy_id:9606)

From the large excel file gathered from UniProt, I narrowed down my search into an analyzing specific gene called: Breast Cancer Type 2 Susceptibility Protein because it is known to have an abundance of variants that cause deleterious effects. This gene is responsible for tumor suppressor which helps precent cells from growing and dividing too rapidly or in an uncontrolled way. So, mutations in this gene are often very deleterious.

Referring to the figure below, we can see using example sequences when proteins amino acid profile remains unchanged, it's determined to be silent, when an amino acid is deleted or inserted it can detect that a frameshift has occurred and is likely to be deleterious. Showing that if we can convert the nucleotides to amino acids, we can trace what kind of mutation has occurred and if that mutation is likely to be deleterious or not (Figure 1).

```
Mutation Test for Silent Mutation
     Nucleotides of original sequence:       AUGUUUUCUUAUUGUCUUCCUCAUCGUUGA
     Nucleotides of mutated sequence:        AUGUUCUCCUACUGCCUACCCCACCGGUAA
     Amino Profile of orignigal sequence:    Met-Phe-Ser-Tyr-Cys-Leu-Pro-His-Arg-STOP
     Amino Profile of mutated sequence:      Met-Phe-Ser-Tyr-Cys-Leu-Pro-His-Arg-STOP
     Mutation Type is:                       Silent : Not Deleterious
     Protein Similarity Percentage:          100.0 %


Mutation Test for Frameshift Deletion Mutation
     Nucleotides of original sequence:       AUGUUUUCUUAUUGUCUUCCUCAUCGUUGA
     Nucleotides of mutated sequence:        AUGUCUUAUUGUCUUCCUCAUCGUUGA
     Amino Profile of orignigal sequence:    Met-Phe-Ser-Tyr-Cys-Leu-Pro-His-Arg-STOP
     Amino Profile of mutated sequence:      Met-Ser-Tyr-Cys-Leu-Pro-His-Arg-STOP
     Mutation Type is:                       Frameshift Deletion : Likely Deleterious
     Protein Similarity Percentage:          94.74 %


Mutation Test for Frameshift Insertion Mutation
     Nucleotides of original sequence:       AUGUUUUCUUAUUGUCUUCCUCAUCGUUGA
     Nucleotides of mutated sequence:        AUGUUAUUUUCUUAUUGUCUUCCUCAUCGUUGA
     Amino Profile of orignigal sequence:    Met-Phe-Ser-Tyr-Cys-Leu-Pro-His-Arg-STOP
     Amino Profile of mutated sequence:      Met-Leu-Phe-Ser-Tyr-Cys-Leu-Pro-His-Arg-STOP
     Mutation Type is:                       Frameshift Insertion : Likely Deleterious
     Protein Similarity Percentage:          95.24 %


Mutation Test for Missense Mutation
     Nucleotides of original sequence:       AUGUUUUCUUAUUGUCUUCCUCAUCGUUGA
     Nucleotides of mutated sequence:        AUGCUCUCUUAUUGUCUUCCUCAUCGUUGA
     Amino Profile of orignigal sequence:    Met-Phe-Ser-Tyr-Cys-Leu-Pro-His-Arg-STOP
     Amino Profile of mutated sequence:      Met-Leu-Ser-Tyr-Cys-Leu-Pro-His-Arg-STOP
     Mutation Type is:                       Missense : Likely Deleterious
     Protein Similarity Percentage:          95.0 %
```

***Figure 1: Example of Reading of Nucleotide to Protein to Determine Mutation Type***

Based on my hypothesis on changes of protein structure. I predict Mutations such as frameshift and missense are more likely to have deleterious effects than mutations like silent mutations because the structure of the protein changes. And I also predict that if a missense were to happen, changes in chemical class are more significant than if a there isn't a change. Such as if a cysteine changes to a glycine, it wouldn't be as worst. But if a Cystine changed into a Histidine it would be a lot more significant (Figure 2).

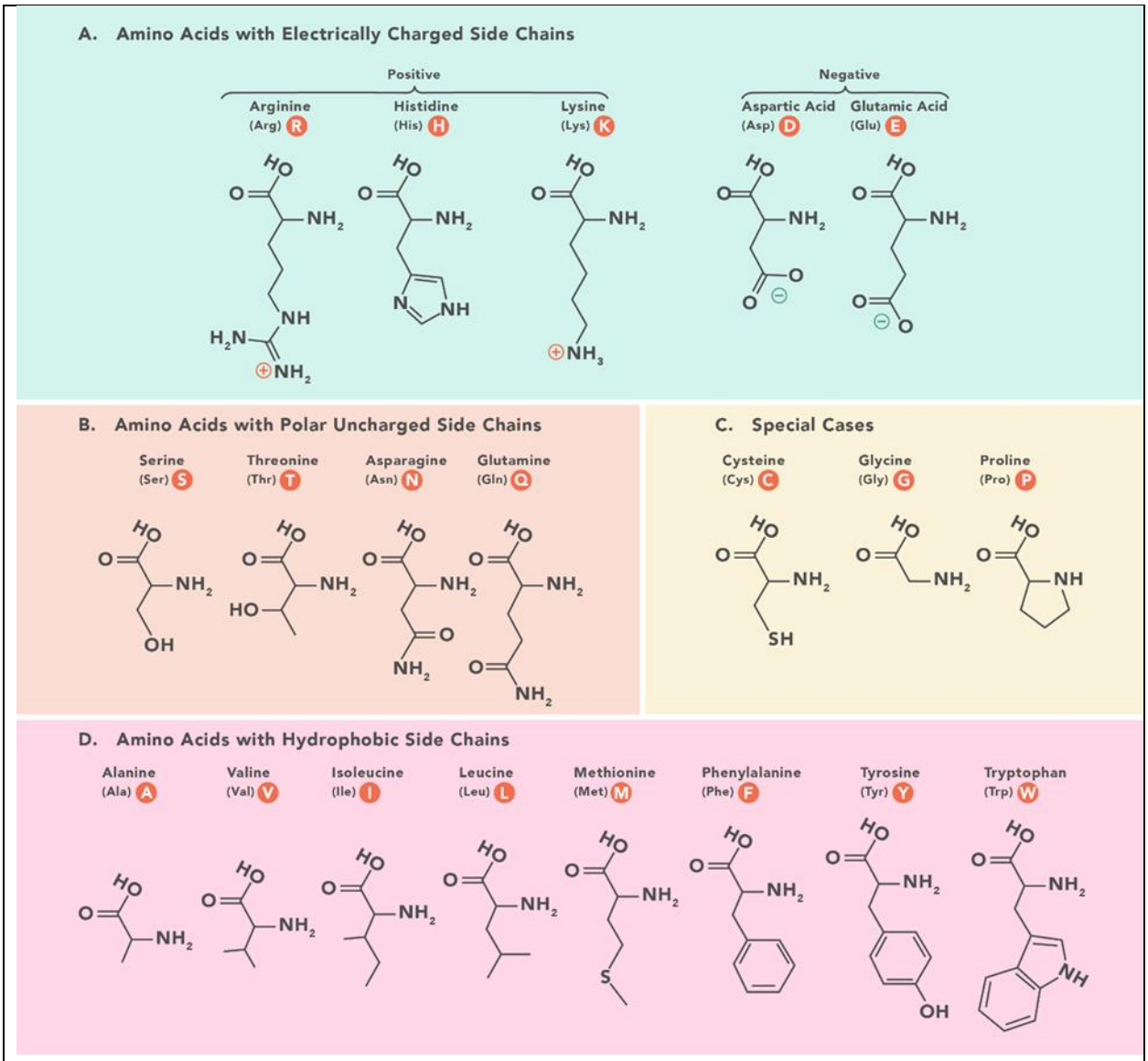



*Figure 2: Amino Acid Chemical Classes*

**Source:**

Steward, K. (2019, September 26). Amino acids – the building blocks of proteins. Applied Sciences from Technology Networks. Retrieved from https://www.technologynetworks.com/applied-sciences/articles/essential-amino-acids-chart-abbreviations-and-structure-324357

To test if the results match my hypothesis, I will run the FASTA files into my program to get the mutation type and likelihood of it being deleterious, then refer back to the database to see if my results match up with what is already known about the mutation. If the mutation is

missense, I will also perform a chemical class test to see if changes in chemical class will affect the likelihood of the mutation being deleterious.

## Results

From the database I selected 30 samples, 10 randomly that contain silent, missense, and frameshift mutation and compared them to the reference variant. I ran them into the program which predicts that missense and frameshift mutations will be more likely to be deleterious. Then I referred to the database and looked at the effects. Results from the database show that 0/10 silent mutations are deleterious, 4/10 missense are deleterious, and 8/10 frameshift mutation are deleterious. Refer to figure below for graph of results from database (Figure 3).

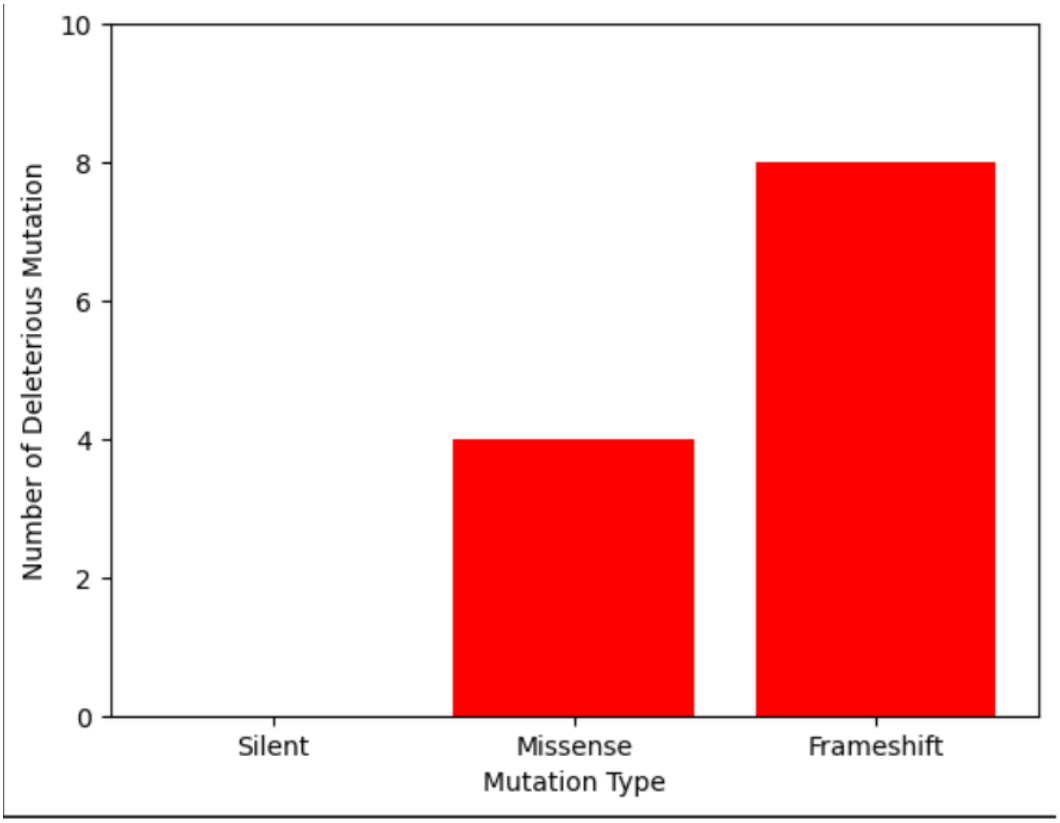



***Figure 3: Number of Deleterious Mutation by Mutation Type***

The program was able to detect the mutation type. However, it failed to accurately detect if the mutation is deleterious. By default, the program will detect 100% of missense and frameshift to be likely deleterious.

To test for significance of the mutation types, we will be performing t-test, to compare the means of two groups. As a baseline for the reference variant, we will be using the mean of 0 because no deleterious mutations are present. The t-test will help test our hypothesis to determine if changes in protein structure will have a significant effect on BRCA2's association with diseases.

***Table 1: T-test of Mutation Types***

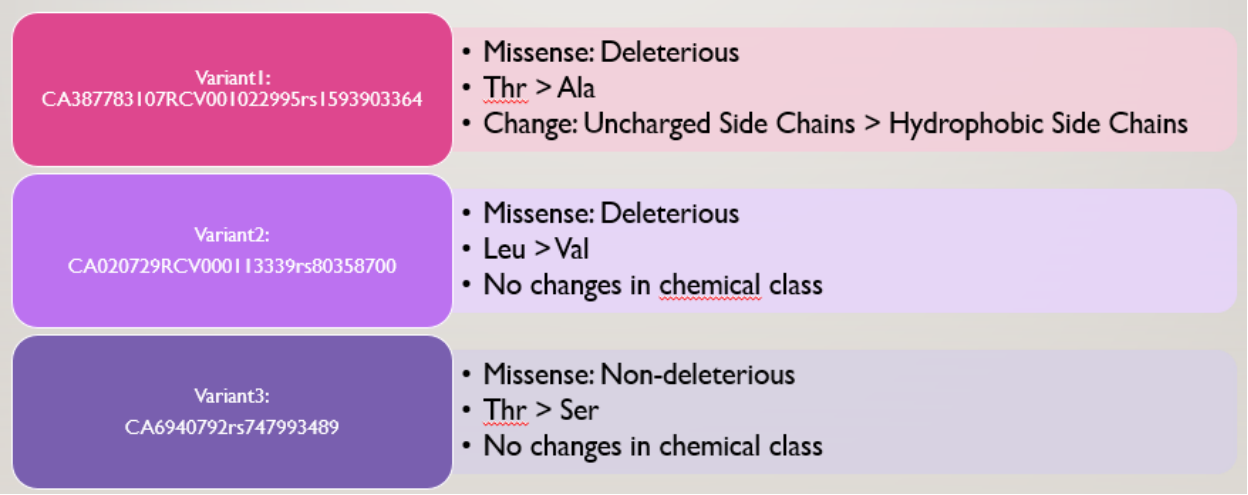
| Group | Silent | Missense | Frameshift |
|---|---|---|---|
| Mean | 0.00 | 0.40 | 0.80 |
| SD | 0.00 | 0.52 | 0.42 |
| SEM | 0.00 | 0.16 | 0.13 |
| N | 10 | 10 | 10 |
| P-Value | 1.000 | 0.0248 | 0.0001 |

The P-Value of 1.000 in Silent indicates that there are zero differences when compared to reference variant. While the P-Value of 0.0248 of missense mutation indicates that there is a statistically significant difference when comparing the deleterious nature of missense mutation and that of reference variant. The P-Value of 0.0001 of frameshift mutation also indicates a statically significant difference when comparing the deleterious nature of frameshift mutation and that of reference variant. (Table 1)

While the results do show that mutations that affect the structure of protein are significantly more likely to be deleterious, my original prediction is challenged, as using protein structure alone isn't 100% accurate predictor for deleterious mutations. As shown in the case of missense and frameshift.

For missense testing of chemical classes, I only analyzed three results because of limited time constraint. Referring to Figure 4 below, we see variant_1 has an entire chemical class change from uncharged side chains to hydrophobic side chains and is associated with a disease. While variant_2 doesn't have a chemical class change but is also associated with a disease. However, variant_3 doesn't have a chemical class change but is not associated with a disease. So, it doesn't seem like a very good indicator because even when there isn't a chemical class change effects can be deleterious.

***Figure 4: Missense Testing***




## Discussion

What this means for our hypothesis is that changes in protein structures could have a link to deleterious mutations. So, while my code wasn't able to accurately predict deleterious mutations,

the results did indicate that frameshift and missense have a significantly more association with diseases.

However, in terms of drawing a conclusion, I can't really draw a conclusion on the hypothesis because I do not have a large enough sample size. As it took a lot of manuals searching to get the variant_id and get the FASTA file from a database.

Specifically testing the Missense database, if I spend a bit more time on it, I can achieve better results. I made the chemical class program at the end so didn't have time to go over a large sample size. Thus, if I were to do this project again, I would start a lot earlier in the data gather phase.

Relating back to previous studies and current methods. It was pretty much the same method, where I used one variant as a reference point to test for mutations. When testing BRCA2, I read more studies into the gene and in one of the studies, a silent mutation where a guanine mutated into an adenine and still coded for the same lysine, resulted in BRCA2 having exon skipping which is associated with misaligned sections of genetic code (Hansen 2010). So even if there isn't a significant change in protein structure and the DNA codes for the same protein, there can still sometimes be a disease. So, while changes in protein structure can be an indicator, it isn't 100% accurate.

This leads us back to the current gaps in the current methods of testing because some genomes lack references, so if a mutation isn't in a database, we do not understand much of it and we are heavily reliant on reference genomes. So, scientists are still searching for more and more variants to build a database for more references to determine effects of mutations. My

project highlights this problem, as we are unable to accurately determine if a mutation is associated with a disease or not based on changes to protein structure alone.

## References


Charlesworth, B. (2012, January 1). The Effects of Deleterious Mutations on Evolution at Linked Sites. OUP Academic. Retrieved November 1, 2022, from https://doi.org/10.1534/genetics.111.134288

Collins, F. S., & Fink, L. (1995). *The Human Genome Project*. Retrieved December 1, 2022, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6875757/

Dunnen, J. T., & Antonarakis, S. E. (1999). Mutation nomenclature extensions and suggestions to describe complex mutations: A discussion. Human Mutation, 15(1), 7–12. https://doi.org/10.1002/(sici)1098-1004(200001)15:1<7::aid-humu4>3.0.co;2-n

Hansen, T. V. O., Nielsen, F. C., Ejlertsen, B., Andersen, M. K., Jønson, L., & Steffensen, A. Y. (2010, February). *The silent mutation nucleotide 744 g --> A, Lys172lys, in exon 6 of BRCA2 results in Exon skipping*. Breast cancer research and treatment. Retrieved December 1, 2022, from https://pubmed.ncbi.nlm.nih.gov/19267246/

Henn, B. M., Botigué, L. R., Bustamante, C. D., Clark, A. G., & Gravel, S. (2015, May 12). *Estimating the mutation load in human genomes*. Nature News. Retrieved November 15, 2022, from https://www.nature.com/articles/nrg3931

MULLER, H. J. (1950, June). Our Load of Mutations. American journal of human genetics. Retrieved November 1, 2022, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1716299/

Ruan, Y., & et al. (2022). *Improving polygenic prediction in ancestrally diverse populations*. Nature Genetics. Retrieved November 15, 2022, from https://scholar.harvard.edu/tge/publications/improving-polygenic-prediction-ancestrally-diverse-populations

Steward, K. (2019, September 26). Amino acids – the building blocks of proteins. Applied Sciences from Technology Networks. Retrieved November 15, 2022, from https://www.technologynetworks.com/applied-sciences/articles/essential-amino-acids-chart-abbreviations-and-structure-324357