Final Report

Erin Ledford

BIOI 4870

Gene to Protein: Human Mitochondria Genes and Their Proteins

**Abstract**

The project is currently a small database with a DDL and DML of human mitochondria genes and proteins. The objective of this project is to learn how NCBI and other databases create user friendly databases that show genes and proteins [2, 5, 8]. The goal of this project is to make a small, user friendly database with only human mitochondria genes and proteins. The research question this database answers is how many variants of mitochondria genes in the database are related to diseases. This is answered by querying the database created for the variants and the highest number of variants related to a disease.

**Introduction / Background**

The human mitochondrial genome is important to understand for those in the health field. There are many diseases associated with mitochondrial genes and proteins [3, 7]. Mitochondrial DNA is inherited differently from compared to nuclear DNA [3, 4, 6, 7]. Mitochondria DNA is inherited only maternally [3, 4, 6, 7]. This makes it harder for those that need to find information about the diseases caused by those genes as they are less researched compared to nuclear DNA. This project is needed in the field to provide a smaller, more specific database to search for the genes or proteins that may cause these diseases or to allow users to study specific genes and proteins.

A version of this project exists elsewhere, in NCBI, as a giant database that can be hard for users to search [2]. The difference between NCBI and my project is the both the size (mine is far smaller) and that my project will be an aggregation of all reputable sources of human mitochondrial genes and proteins sequences from NCBI, Uniprot, and Ensembl [2, 5, 8].

This project is significant as it shows just how much variation in genes can cause disease [3, 7]. There are many variations that can be associated with disease [3, 7]. This project has fourteen genes in the database and all of these genes contain variants that lead to disease, with over two thousand variants that lead to disease in only fourteen genes [5].

**Database Diagram**

The Database diagram is within the github for my code. Here is the link: https://github.com/bearyboo126/Course-Project-Undergraduate-BIOI-4870.

**Research Questions**

This project will store and display information and sequences of human mitochondria genes and their proteins to find out how many genes have proteins for the mitochondria. This project also stores and displays information on what disease is linked to the biggest number of variants.

**Methods**

**Code**

All code is contained within my github. Here is the link: <https://github.com/bearyboo126/Course-Project-Undergraduate-BIOI-4870>.

**Data Dictionary**

The contents of my database are human mitochondria genes and proteins. It will be a dummy database as Odin cannot handle all the data for human mitochondria genes and proteins.

The database is split into two tables, genes and proteins. The genes table describes the mitochondria genes in the database. The proteins table describes the mitochondria proteins in the database.

There are many variables that I had to add to my database as I created the code. The gene ID is a number based off NCBI and is the primary key [2]. Gene symbol is the alias that gene is known as, usually based off the full name of the gene [2]. The full name of the gene is what it is known as in NCBI [2]. The organism tag is always the same organism but is kept for clarity to make sure that the user knows where this gene and its sequence come from [2]. Last\_updated variable is grabbed from NCBI or Uniprot as to when the gene or protein was updated [2, 8]. Location for genes is where on a chromosome is it located. bp\_length is the length of the DNA sequence for the gene [2]. Gene type describes where the gene ends up, either protein-coding or something else like tRNA [2]. The associated\_protein variable describes whether the gene has a corresponding protein, usually this depends on its gene\_type [8]. The num\_associated\_variants describe the total amount of variants associated to disease for that gene and is grabbed from Ensembl [5]. The num\_highest\_variant is the highest number of variants related to one disease [5]. The highest\_disease-variant is the disease with the highest number of variants that are associated to that disease [5].

For the protein table, the primary key is the accession\_number of the protein, which gives it an easy and unique identifier [8]. The full\_name of the protein is grabbed from NCBI or Uniprot to give a better explanation of the protein [2, 8]. Every protein must come from a gene so there is a variable called associated\_geneSymbol used to refer to the protein's gene [2, 8]. The organism variable is the same as in the genes table, just used for clarity [2, 8]. Last\_update in the proteins table is the same as the genes table [2, 8]. Location is used to describe where the protein resides in the mitochondria [8]. The aa\_length variable is used to describe protein sequence length [8]. The sequence variable is the sequence of the protein [8].

**License**

This project falls under the MIT License that only requires preservation of copyright and license notices.

MIT License

Copyright (c) [2020] [Erin Ledford]

Permission is hereby granted, free of charge, to any person obtaining a copy of this software and associated documentation files (the "Software"), to deal in the Software without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the Software, and to permit persons to whom the Software is furnished to do so, subject to the following conditions:

The above copyright notice and this permission notice shall be included in all copies or substantial portions of the Software.

THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.

**Data Provenance**

The data for this project was gathered by hand. Data was gathered from NCBI, Ensembl, and Uniprot [2, 5, 8]. All webpages where data was gathered from for this database is located within the SQL code to populate my database. NCBI genes search was used to find the protein coding genes for the full human mitochondria genome, after entering human mitochondria into the search bar on the home page [2]. After the genes were gathered from NCBI, those genes were used to search the Uniprot database to find the associated proteins [2, 8]. The same gene symbols gathered from NCBI were also used to find the corresponding genes in the Ensembl database [2, 5]. Each gene page has a phenotypes page that shows what diseases are associated with each gene [5]. Every page used for data gathering is within my SQL code.

**Webpage**

Sample queries are located within the FAQ on my front-end. Here are some queries to try: 4519, 4541, 4508, 4512.

Here is the link to my front-end: http://odin.unomaha.edu/~eledford/course\_project\_web.php

**Results**

The result of my research is a searchable front-end website that allows the user to search for a gene and the result it gives are the gene and its associated protein. The research page of my front end does not take a search term but does show each gene in the database and how many variants there are for each gene and what disease has the highest number of variants associated with it. For all genes within the database, the disease with the highest number of variants associated with it is leigh syndrome [4-6]. This may mean that it is more likely that a variation in mitochondria genes will lead to leigh syndrome. These answer the research question of which disease has the highest number of variants. The home page of the front end answers my other research question of how many protein-coding mitochondria genes I can fit into a database successfully, which is all 13 protein coding genes within human mitochondria [3, 7].

**Discussion and Conclusions**

The implications of this final project and database shows that there are many variants associated with disease for these genes within the database [3, 4, 6, 7]. This has been done before, but it does not mean that it is less important. There was no aggregation of both sequences and highest variants associated with diseases that I could find on one website. The result of my analysis shows that the highest number of variants associated with a single disease for every gene within the database is leigh syndrome [4-6]. According to NIH, leigh syndrome is a neurodegenerative disorder that develops in infancy or childhood. Most of the causes of Leigh syndrome are point mutations [4, 6]. This disorder may cause symptoms of vomiting, delayed development, seizures, muscle weakness, problems with movement, heart disease, kidney problems, and difficulty breathing [4, 6]. The mitochondrial type of leigh syndrome can only be inherited maternally as only females can pass on mitochondrial associated DNA [4, 6].

**Challenges encountered and their solutions**

I encountered challenges with making a search bar specifically to search for proteins. I could not get this fixed before PHP broke on Odin. As such, the search bar gives results for both the gene and its protein. The only way to search this search bar is to put in a gene ID. The gene IDs for the database can be found within the FAQ page.

**References**

[1] E. A. Schon, S. DiMauro, and M. Hirano, “Human mitochondrial DNA: roles of inherited and somatic mutations,” Nature Reviews Genetics, vol. 13, no. 12, pp. 878–890, 2012.

[2] E. W. Sayers, J. Beck, E. E. Bolton, D. Bourexis, J. R. Brister, K. Canese, D. C. Comeau, K. Funk, S. Kim, W. Klimke, A. Marchler-Bauer, M. Landrum, S. Lathrop, Z. Lu, T. L. Madden, N. O’Leary, L. Phan, S. H. Rangwala, V. A. Schneider, Y. Skripchenko, J. Wang, J. Ye, B. W. Trawick, K. D. Pruitt, and S. T. Sherry, “Database resources of the National Center for Biotechnology Information,” OUP Academic, 23-Oct-2020. [Online]. Available: https://academic.oup.com/nar/article/49/D1/D10/5937080. [Accessed: 02-May-2021].

[3] H. Li, J. Slone, L. Fei, and T. Huang, “Mitochondrial DNA Variants and Common Diseases: A Mathematical Model for the Diversity of Age-Related mtDNA Mutations,” Cells, vol. 8, no. 6, p. 608, Jun. 2019.

[4] NIH, “Mitochondrial DNA-associated Leigh syndrome,” Genetic and Rare Diseases Information Center. [Online]. Available: https://rarediseases.info.nih.gov/diseases/3671/maternally-inherited-leigh-syndrome. [Accessed: 02-May-2021].

[5] S. E. Hunt, W. McLaren, L. Gil, A. Thormann, H. Schuilenburg, D. Sheppard, A. Parton, I. M. Armean, S. J. Trevanion, P. Flicek, and F. Cunningham, “Ensembl variation resources,” Database, vol. 2018, 2018.

[6] S. Rahman, R. B. Blok, H.-H. M. Dahl, D. M. Danks, D. M. Kirby, C. W. Chow, J. Christodoulou, and D. R. Thorburn, “Leigh syndrome: Clinical features and biochemical and DNA abnormalities,” Annals of Neurology, vol. 39, no. 3, pp. 343–351, Mar. 1996.

[7] T. R. W. T. DM; “Mitochondrial DNA mutations in human disease,” Nature reviews. Genetics, 06-May-2005. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/15861210/. [Accessed: 02-May-2021].

[8] The Uniprot Consortium, “UniProt: the universal protein knowledgebase in 2021,” Nucleic Acids Research, 25-Nov-2020. [Online]. Available: https://academic.oup.com/nar/article/49/D1/D480/6006196. [Accessed: 02-May-2021].