Optimal Tool to Mine Missense Variants Pathogenicity Prediction Data From Deep Learning

Martin Min

Jericho High School 11th Grade

Abstract

MVP (Missense Variants Pathogenicity prediction) is a deep learning method that uses a residual neural network to indicate the pathogenicity of missense variants relative to all 76 million possible missense variants. Making queries efficient and convenient for large datasets such as that of MVP is crucial to researchers who will use the queries to compare results for projects similar to MVP. Here the creation of a database and website that acts as a median to access and query the data from MVP is described. Using MongoDB and node.js, the website was able to utilize the database and make queries that require several fields execute in a second or less.

Introduction

Improvement of predicting accurate pathogenicity of missense variants is crucial in genetic studies and more accurate interpretation in clinical genetic testing[1]. Missense variants are a major risk of common and rare diseases and have been shown to contribute to birth defects[2-3] and neurodevelopmental disorders[4-5], a small percentage of missense de novo mutations are pathogenic[4]. Thus, the ability to detect individual genes that have a risk of disease from missense variants is limited[6].

Machine learning has been used, yet there are problems such as a lack in the capability of leveraging training data[7] and a considerable frequency of false positives, which inflate the performance of the machine learning methods[8]. As a result, a new deep learning approach, Missense Variant Pathogenicity prediction (MVP)[1], was developed to address such issues and to determine pathogenicity through higher MVP scores, which are percentiles of raw sigmoid output relative to all missense variants from the HGMD[10], UniProt[11], and CliniVar[9].

Processing large training data in machine learning approaches such as MVP results in a sizable amount of rows with multiple fields to consider. Thus, databases are utilized as a solution to organize the data into sortable and manageable facilities of specific records that can be queried[12]. However, the queries used in the database discussed require multiple fields and severely inhibit the speed of the queries as a result. Once the data is analyzed through the residual neural network, it must be exhibited on a website as a reference. Thus, we created a database to organize the data into a publicly accessible website.

Materials and Methods

MVP data

The finalized data that's used in the database consists of missense variants from various sources that were used to order predictions based on correlation since highly correlated predictors are often clustered together. In addition, the missense variants' were described with the following fields: base pairs, gene, transcription ID, and MVP score(Fig. 1), which is the rank percentile of ResNet's raw sigmoid output to all 76 million missense variants[1]. The data was obtained through GitHub, a web-hosting service that uses Git.

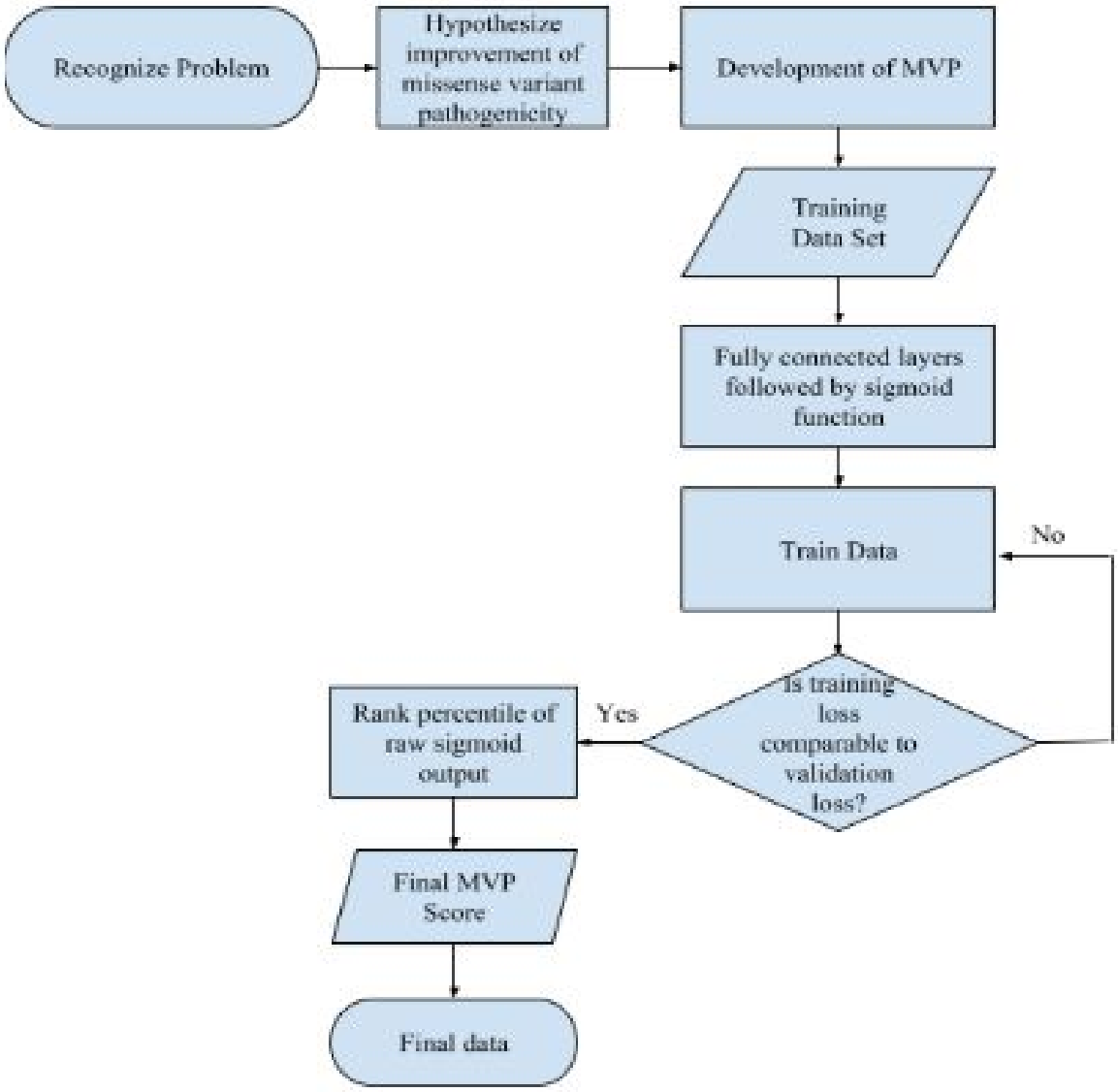

Figure 1. Workflow diagram of obtaining the finalized data for MVP.

Importing data into MongoDB

We utilized MongoDB Windows as a platform to load 76 million rows of data into a database that can be efficiently queried in minimal time. The final data was given as a text file and in order to import the data into MongoDB, the file had to be converted into a JSON format, so a python program that converts rows into JSON was created (Fig. 2). MongoDB Compass was used to import the JSON files into a database and allowed the creation of indexes, which are used to quickly locate data.

Creating a backend service

Node.js was used to create the backend service due to its efficiency in serverside API, ability to handle concurrency, speed from Google's V8 Javascript engine[13], and non-blocking I/O model that permits other processing before transmission. Furthermore, Node.js is an open source framework that contains several modules, specifically MongoDB[14]. Thus, MongoDB can easily be implemented into the backend service to query a database.

Building the website

In order to build the website, HTML was needed to implement front and back services that pull information from the database through the use of javascript. The main purpose of the website is to organize the data into queries, and in this case can be made with multiple criteria such as gene coordinate, gene symbol, and transcription ID. In addition, the query boxes that require different inputs allow for the search of missense variants based on specific fields.

Results

The final website was able to do just as intended, allow for queries that require specific fields: genomic coordinate, transcription ID, and gene (Fig. 3). Genomic coordinate is a concatenated field that uses four fields separated by dashes in order to create an efficient query. In addition, the data was queried into a table(Fig. 4), so all queries are organized by the fields that describe each of the 76 million rows of data. The website also contains a navigation bar that contains links to different pages that explain the background of MVP. Thus, the website was built for the convenience of the user and is able to query data in a second or less.

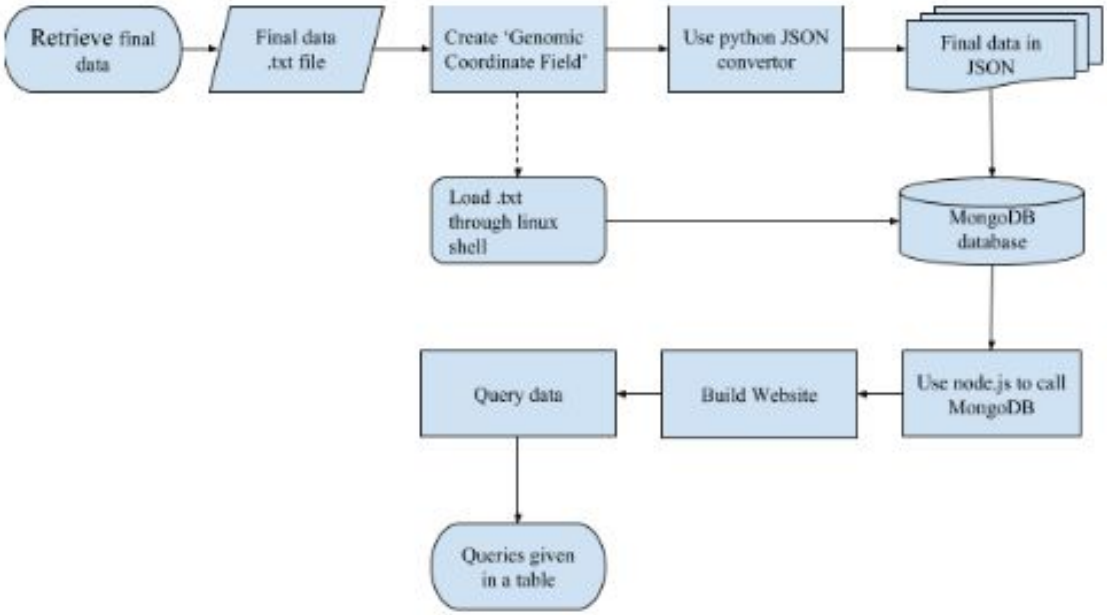

Figure 2. Workflow diagram of creating the website and querying data

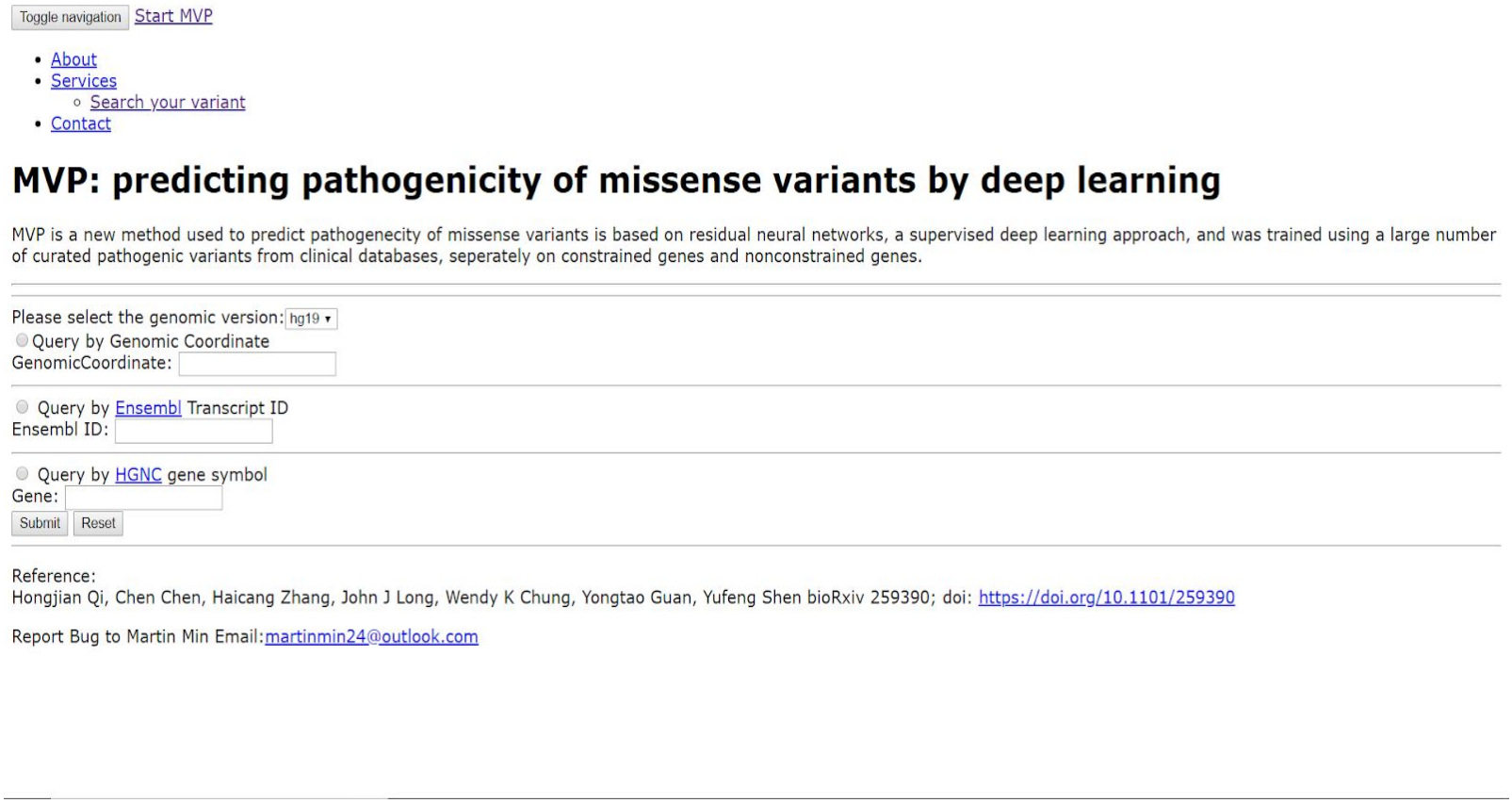

Figure 3. Finalized website and query selections

Results

Chromosome	Position	Ref	alt	aaref	aaalt	GeneSymbol	Ensembl_transcriptid	MVP_score
1	69091	A	G	M	V	OR4F5	ENST00000335137	0.197625483188
1	69095	T	C	V	A	OR4F5	ENST00000335137	0.329540904979
1	69095	T	G	V	G	OR4F5	ENST00000335137	0.465721554213
1	69097	A	C	T	P	OR4F5	ENST00000335137	0.378148810121
1	69097	A	G	T	A	OR4F5	ENST00000335137	0.314417295294
1	69095	T	A	V	E	OR4F5	ENST00000335137	0.564761482634
1	69097	A	T	T	S	OR4F5	ENST00000335137	0.277730125212
1	69098	C	A	T	N	OR4F5	ENST00000335137	0.344710718752
1	69098	C	G	T	S	OR4F5	ENST00000335137	0.292423486923
1	69098	C	T	T	I	OR4F5	ENST00000335137	0.416454006429
1	69100	G	A	E	K	OR4F5	ENST00000335137	0.480497669815
1	69101	A	C	E	A	OR4F5	ENST00000335137	0.518038692251
1	69101	A	G	E	G	OR4F5	ENST00000335137	0.533325340949
1	69101	A	T	E	V	OR4F5	ENST00000335137	0.525613252392
1	69102	A	C	E	D	OR4F5	ENST00000335137	0.494366844524
1	69102	A	T	E	D	OR4F5	ENST00000335137	0.494366844524
1	69091	A	T	M	L	OR4F5	ENST00000335137	0.243398259712
1	69093	G	A	M	I	OR4F5	ENST00000335137	0.240491677333
1	69103	T	A	F	I	OR4F5	ENST00000335137	0.392547445146
1	69092	T	G	M	R	OR4F5	ENST00000335137	0.27855597813
1	69092	T	A	M	K	OR4F5	ENST00000335137	0.278143212241
1	69100	G	C	E	Q	OR4F5	ENST00000335137	0.546780063783
1	69093	G	C	M	I	OR4F5	ENST00000335137	0.240491677333
1	69093	G	T	M	I	OR4F5	ENST00000335137	0.240491677333
1	69094	G	C	V	L	OR4F5	ENST00000335137	0.28722502521
1	69103	T	G	F	V	OR4F5	ENST00000335137	0.500931478032
1	69104	T	A	F	Y	OR4F5	ENST00000335137	0.57809870819
1	69104	T	C	F	S	OR4F5	ENST00000335137	0.679297102143
1	69105	C	A	F	L	OR4F5	ENST00000335137	0.236890367714
1	69105	C	G	F	L	OR4F5	ENST00000335137	0.236890367714

Figure 4. Queried data organized into a table

Discussion

The ability to query the finalized data is important because the predicted pathogenicity of each missense variant can be easily given to researchers who are working on a similar project along with more accurate predictions and interpretations in clinical genetic testing. Thus, future methods can build upon the findings of MVP in order to address similar issues such as false positives, inflation of benchmark performance, and low positive predictive value. In addition to scientific use, the website can be a commercial service that caters to clinical geneticists. Furthermore, data from other deep learning methods similar to MVP can also be used with the website to query data if they were to have fields similar to that of the MVP data.

The most challenging part of creating an efficient query was greatly decreasing the time required to make a query based on four fields. The four fields —chromosome, position, ref, and alt— were concatenated into a separate field because previously, attempting to query by the four fields alone wasn't even possible due to the shortage of memory needed to handle the immense amount of information. After creating the new field, querying by the four fields only took a second, which proves a great success in the creation of 'genomic coordinate.'

Conclusion

In summary, a website with the ability to query data from a relatively large database was created with multiple fields through the use of MongoDB and node.js. The website serves as a median to receive data on specific missense variants' pathogenicity for the purpose of researchers being able to compare results when working on similar projects or for commercial use in clinical genetic testing. MongoDB was the most ideal platform for database usage because of its convenience, extensive capabilities for efficiency, and cross-platform usage. Finally, queries were able to be made within a second, which is an indicator of the success of the usage of a concatenated field to optimize query execution.

Supplementary Materials

Final data can be found here:
https://www.dropbox.com/s/bueatvqnkvqcb54/MVP_scores_hg19.txt.bz2?dl=0
MongoDB can be found here: https://www.mongodb.com/
Node.js can be found here: https://nodejs.org/en/
GitHub for the project described: https://github.com/mmin4906/website

Acknowledgements

Thanks to Professor Yufeng Shen of the Department of Systems Biology, Columbia University for giving me the opportunity to work on this project to learn more about the application of deep learning to bioinformatics and thanks to Dr. Haicang Zhang and Jingwei Ren, who are with the same department, for helpful discussions regarding the use of MongoDB and the creation of the website.

References

1. Qi, H. et al. (2018). MVP: predicting pathogenicity of missense variants by deep learning.
2. Homsy, J. et al. De novo mutations in congenital heart disease with neurodevelopmental and other congenital anomalies. Science (New York, N.Y.) 350, 1262-1266 (2015).
3. Jin, S.C. et al. Contribution of rare inherited and de novo variants in 2,871 congenital heart disease probands. Nature genetics 49, ng. 3970 (2017).
4. Iossifov, I. et al. The contribution of de novo coding mutations to autism spectrum disorder. Nature 515, 216-221 (2014).
5. De Rubeis, S. et al. Synaptic, transcriptional and chromatin genes disrupted in autism. Nature 515, 209-215 (2014).
6. Zuk, O. et al. Searching for missing heritability: designing rare variant association studies. Proceedings of the National Academy of Sciences of the United States of America 111, E455-464 (2014).
7. Goodfellow, I., Bengio, Y. & Courville, A. Deep Learning. (MIT Press, 2016).
8. Dorschner, M.O. et al. Actionable, pathogenic incidental findings in 1,000 participants' exomes. The American Journal of Human Genetics 93, 631-640 (2013).
9. Landrum, M.J. et al. ClinVar: public archive of interpretations of clinically relevant variants. Nucleic acids research 44, D862-D868 (2015).
10. Stenson, P.D. et al. The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. Human Genetics, 1-13 (2017).
11. UniProt Consortium, Ongoing and future developments at the Universal Protein Resource. Nucleic acids research 39, D214-D219 (2011).
12. Sestoft, P. (2011). Organizing research data. Acta Veterinaria Scandinavica, 53(Suppl 1), p.S2.
13. Tilkov, S. and Vinoski, S. (2010). Node.js: Using JavaScript to Build High-Performance Network Programs. IEEE Internet Computing, 14(6), pp.80-83.
14. Bangare, Sunil & Gupta, S & Dalal, M & Inamdar, A. (2016). Using Node.Js to Build High Speed and Scalable Backend Database Server. International Journal of Research in Advent Technology (E-ISSN: 2321-9637). 4. 19.