Guessing Possible Harmful Variants of DCAF17 gene

The focus of this research is on the gene DCAF17 which encodes the same named protein DCAF17. This protein has full name DDB1 and CUL4 associated factor 17 and is also known as C2orf37. It is one of the many DCAF proteins that associates with the bounded DDB1 and CUL4 proteins. DCAF17 has only 1 domain, which makes up about one third of its nucleic acid sequence (uniprot n.d.). The reason this gene is selected is the fact that, if some certain mutation happens in this gene, a rare regressive autosomal disease, Woodhouse–Sakati syndrome, occurs in patients. Woodhouse–Sakati syndrome is an endocrine system related disease that is first diagnosed in 1983 (National Center for Biotechnology Information n.d.). The effects of this disease on patients include Hypogonadism, diabetes mellitus, alopecia, mental retardation and electrocardiographic abnormalities. To this date, it has been reported in Eastern Europe, Middle East and India (National Center for Biotechnology Information n.d.).

The main aim of this research is creating a model which can predict whether a mutation in this gene can be harmful, by looking at mutations on closely related versions of this gene in other species. Because the used data was limited quantity wise, only mutations between different groups of amino acids were considered when predicting. The predictions are evaluated using pathologic gene samples obtained from ClinVar database of NCBI (National Center for Biotechnology Information) and non-pathologic samples obtained from gnomAD.

Firstly, a set of amino acid sequences were obtained from blastp using the reference amino acid sequence (National Center for Biotechnology Information n.d.). After picking only one isoform for each species, the sequences were aligned and using the resulting alignment a phylogenetic tree were constructed using MEGA X (MEGA X n.d.). The obtained phylogenetic tree can be seen in figure 1, where the clade that includes primates are colored blue and humans are colored green. A close up of this clade can also be seen on figure 2. The trees are visualized using FigTree (Rambaut n.d.).

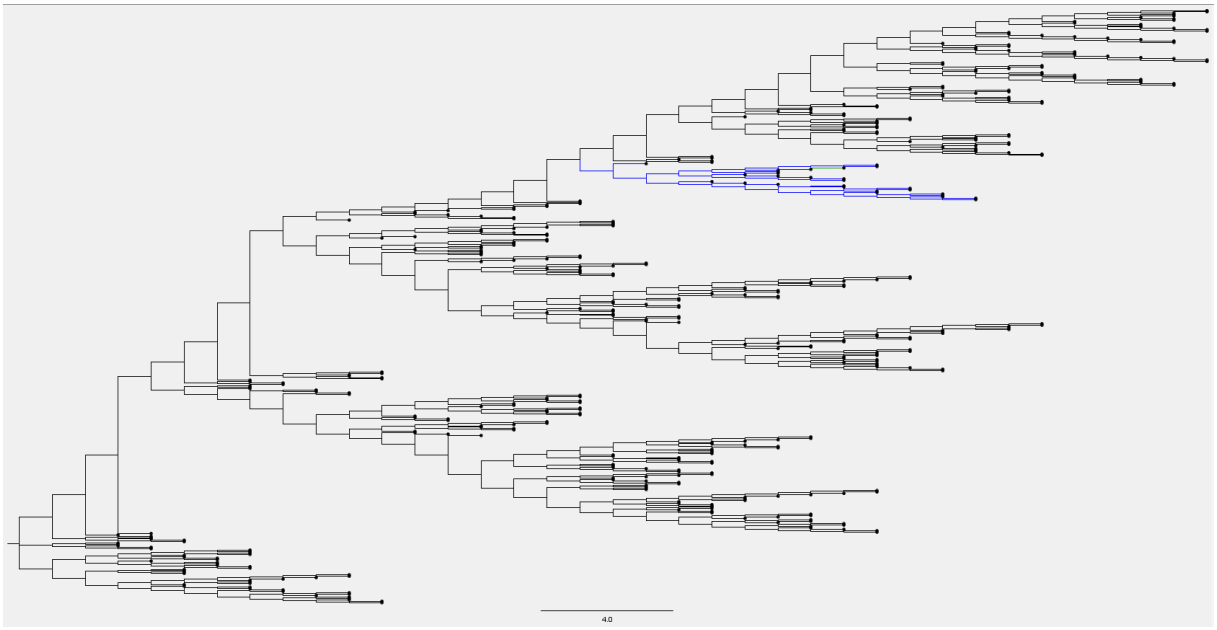

Figure 1: Phylogenetic tree without name of the species

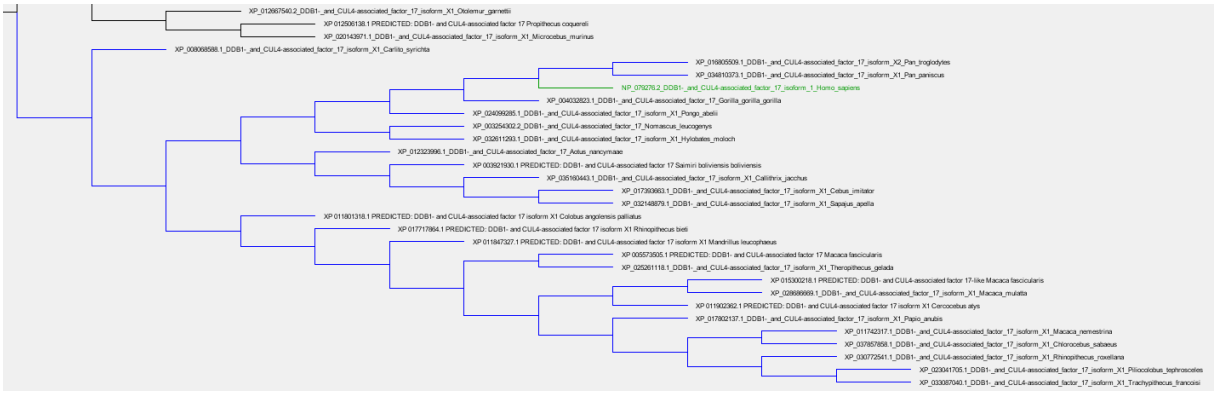

Figure 2: Homo Sapiens on the phylogenetic tree

After creating the phylogenetic tree, a model was constructed using python. This model classifies each index of the amino acid sequence of the alignment by looking at all of the sequences. If there is a consensus on this index, this location is considered important and any mutation that changes the consensus' amino acid group would be predicted as harmful. If there is not a consensus, or the group is not changed, then the mutation is predicted as non-harmful. Whether there is a consensus is determined by a variable (i.e., if threshold is 0.5, if more than half of the species have the same amino acid in that position, the model says there is a consensus there).

Then, to evaluate the model, data were obtained from ClinVar and gnomAD databases. From ClinVar, mutations with reported symptoms that fit Woodhouse–Sakati syndrome are obtained and from gnomAD, missense nonharmful mutations where resulting gene's allele count is reported more than 5 is taken. From the databases 21 nonharmful mutations, 2 likely pathogenic mutations and 9 pathogenic mutations were obtained. For varying thresholds, the results are given in figure3. From these results, 0.85 as a threshold value seems plausible, as between the threshold values where there are no skipped harmful mutations, it has the best accuracy.

```
threshold = 0.5  | Accuracy = 0.69, Sensitivity = 1.0, Specificity = 0.52
threshold = 0.55 | Accuracy = 0.69, Sensitivity = 1.0, Specificity = 0.52
threshold = 0.6  | Accuracy = 0.69, Sensitivity = 1.0, Specificity = 0.52
threshold = 0.65 | Accuracy = 0.69, Sensitivity = 1.0, Specificity = 0.52
threshold = 0.7  | Accuracy = 0.72, Sensitivity = 1.0, Specificity = 0.57
threshold = 0.75 | Accuracy = 0.75, Sensitivity = 1.0, Specificity = 0.62
threshold = 0.8  | Accuracy = 0.75, Sensitivity = 1.0, Specificity = 0.62
threshold = 0.85 | Accuracy = 0.78, Sensitivity = 1.0, Specificity = 0.67
threshold = 0.9  | Accuracy = 0.78, Sensitivity = 0.91, Specificity = 0.71
threshold = 0.95 | Accuracy = 0.81, Sensitivity = 0.82, Specificity = 0.81
threshold = 1.0  | Accuracy = 0.84, Sensitivity = 0.55, Specificity = 1.0
```

Figure 3: results for various thresholds

Lastly, how much of the phylogenetic tree should be used is investigated. To do this a metric *clade height* is introduced. This, for a clade including Homo sapiens, measures the number of unique clades within it that includes Homo sapiens minus 1(minus 1 comes from the fact that only taking the homo sapiens' reference sequence is meaningless for making predictions). For example, using figure2, the clade with height 0 is consisting of Pan troglodytes, Pan paniscus, Homo sapiens while the clade with clade height 1 is consisting of the elements of clade with height 0 and gorilla. Lastly, the clade with height 9 is the tree itself. Using only elements of each clade including Homo sapiens, the model is used with thresholds 0.85 and 0.9. This approach is used to detect whether some difference in the function of this protein for some distant species (albeit highly unlikely) would affect the model negatively. The results can be seen on figure 4.

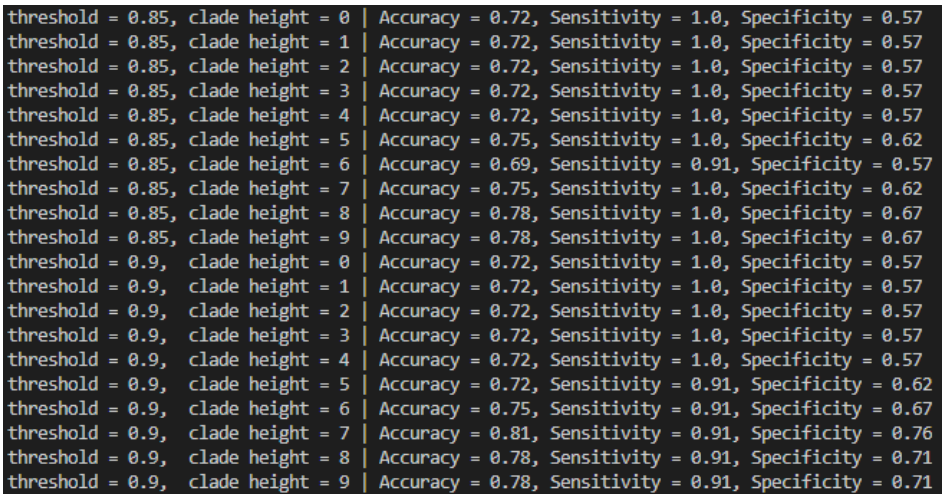

Figure 4: Results for threshold 0.85 and 0.9 and various clade heights

These results are mostly in line with the expected result of having more evolutionary information about a gene making the prediction better with a few exceptions. First is the jump in accuracy at threshold 0.9 and clade height 7. After comparing the results for clade heights, I concluded that it is most likely an anomaly because of the amount of data being too low. The other exception is the results of clade height 8 and 9 being the same, which is more or less the same for every threshold which would mean excluding the most distant clade to the Homo sapiens would not change the model that much.

While this approach gives results with more than 70 percent accuracy and 100 percent sensitivity, because there is too low number of genes found on the databases these results are not too reliable. The results might be heavily affected by the current data, which would make them change drastically with more data about both harmful and non-harmful mutations.

This approach does not consider the mutations that are happening outside of the ORFs. ClinVar database have some Pathogenic entries where the mutation happened outside of the ORFs. Focusing on nucleic acid sequence instead of amino acid sequence might have made the model more comprehensive.

References

MEGA X. n.d. https://www.megasoftware.net/ (accessed Janaury 13, 2021).

National Center for Biotechnology Information. *Standard Protein BLAST.* n.d. https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins (accessed January 13, 2021).

—. *Woodhouse-Sakati Syndrome.* n.d. https://www.ncbi.nlm.nih.gov/books/NBK378974/#:~:text=Woodhouse%2DSakati%20syndrome%20(WSS)%20is%20characterized%20by%20the%20endocrine,hearing%20loss%2C%20and%20intellectual%20disability. (accessed January 13, 2021).

Rambaut, Andrew. *FigTree.* n.d. http://tree.bio.ed.ac.uk/software/figtree/ (accessed January 13, 2021).

uniprot. *Q5H9S7 (DCA17_HUMAN).* n.d. https://www.uniprot.org/uniprot/Q5H9S7 (accessed Janauary 13, 2021).

Firstly, an amino acid sample of DCAF17 was obtained from NCBI. Then, this similar amino acid sequences were found using blastp on ncbi's site. 794 number of samples were found, while the aim was finding 1000 sequences. From ncbi's site, some additional information was gathered, such as the fact that DCAF17 have only one protein domain, and for humans it is ranged from 31st to 505th amino acid, while having 520 amino acids. The existence of only one protein domain were further tested by using hmmer and cdvist. By using the obtained amino acid sequences, using MUSCLE on MEGA application, a multiple sequence assignment was created. After observing some big gaps, a brief investigation was done on the assignment, from which some sequences were found out to be responsible to a big number of gaps (the threshold was established such that if only one or two sequences were present in one column, they would be responsible for this gap, and if a sequence were responsible for more than 40 such gaps, it would be taken out from the data). The main reason for doing this was decreasing workload for computer as much as possible. After doing this, the same phenomena were observed, so the steps are repeated. After the second iteration only one sequence above the threshold was discovered, and because its affect would not be much, it was left. By doing these, 9 sequences were left out and the length of the alignment went down from 1673 to 1400.

After creating the multiple sequence assignment, a phylogenic tree was constructed. Maximum likelihood method was used while creating the tree. The clade where homo sapiens were in was selected. The sequences in this clade were taken separately for future analysis.

For further information on this gene, NCBI site's orf finder was used on this gene's reference nucleotide sequence. The reason for that is to be able to learn what a nucleotide mutation would change amino acid sequence. Information about whether they are on positive or negative strand, starting position, ending position were noted to use in future analysis.

In order to gain data on humans, ClinVar database of NCBI was used and a total of 155 gene entries where a mutation happened on DCAF17 gene were collected. From their HGVS code, the location of their mutation was determined. These locations were compared with the ORFs depending on the disease's severity. For genes that are marked with pathogenic and disease's symptoms, 8 out of 8 had their mutation in ORFs. For genes that are marked with only disease's symptoms, (regardless of being pathogenic) this ratio was 125 out of 139. And finally, for all genes 133 out of 152 had a mutation in ORFs.

• Figures (min 4) with legends

• Discussion (½ page)

• Materials and Methods (½-1 page)

• References