Prediction and Classification Model on Functions of Cytochromes P450 Enzymes

15-400, Spring 2021

Katrina Liu

Carnegie Mellon University

May 16, 2021

1 Abstract

Cytochrome P450 enzymes are very important in steroid biosynthesis and drug metabolism in human and natural product biosynthesis pathways.

By building a classification/prediction model on the functions of P450 enzymes, we can better understand their behavior and classify them based on their functions for newly discovered P450 enzymes. This project builds classification models of the hydroxylation function of proteins in P450 family using k nearest model and depth first search for their neighbors in connectivity graphs based on their pairwise alignment scores. We also examined the prediction accuracy of models and we try to add additional features that boost prediction accuracy.

2 Introduction

2.1 Background

Cytochrome P450 enzymes are very important in steroid biosynthesis and drug metabolism in human and natural product biosynthesis pathways[10]. Cytochrome P450 pathways have been classified based on the gene sequence similarity: a number is assigned for each family, a letter is assigned for each subfamily, and a final number is assigned for specific isoforms[8]. Cytochrome P450 proteins play a key role in the metabolic process of many drugs. The key cytochrome P450 drugs are involved in approximately 80 percent of oxidative drug metabolism[13]. Six of the proteins, CYP1A2, CYP2C9, CYP2C19, CYP2D6, CYP3A4, and CYP3A5, in the family metabolize 90% of the drugs[7].

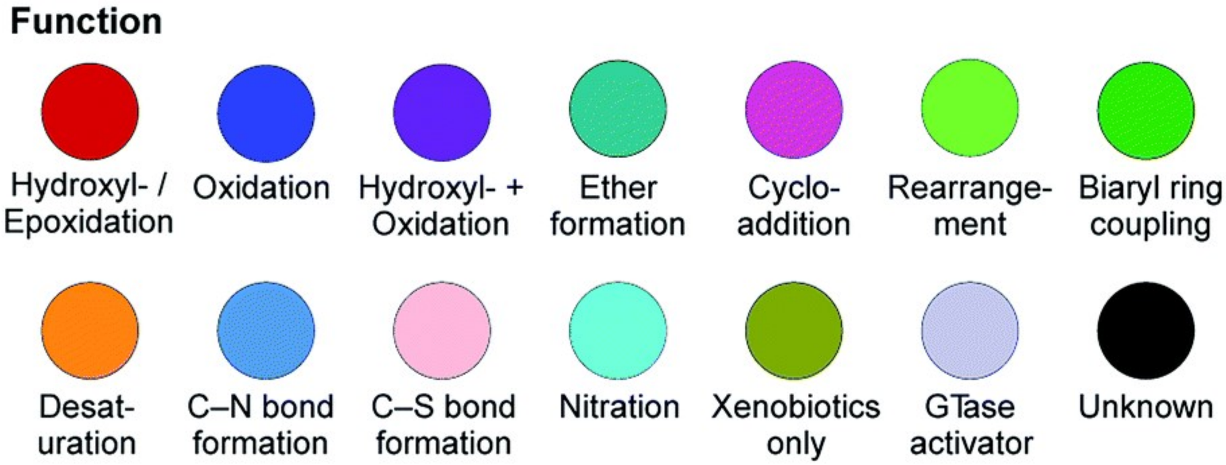

Figure 1: Different classes of functions of cytochrome P450 proteins[10]

P450s are commonly associated with the hydroxylation, epoxidation, and dealkylation of xenobiotics found in human drug metabolism.The common functions of cytochrome P450 family are mentioned in Figure 1 and their specific metabolic functionality is demon-

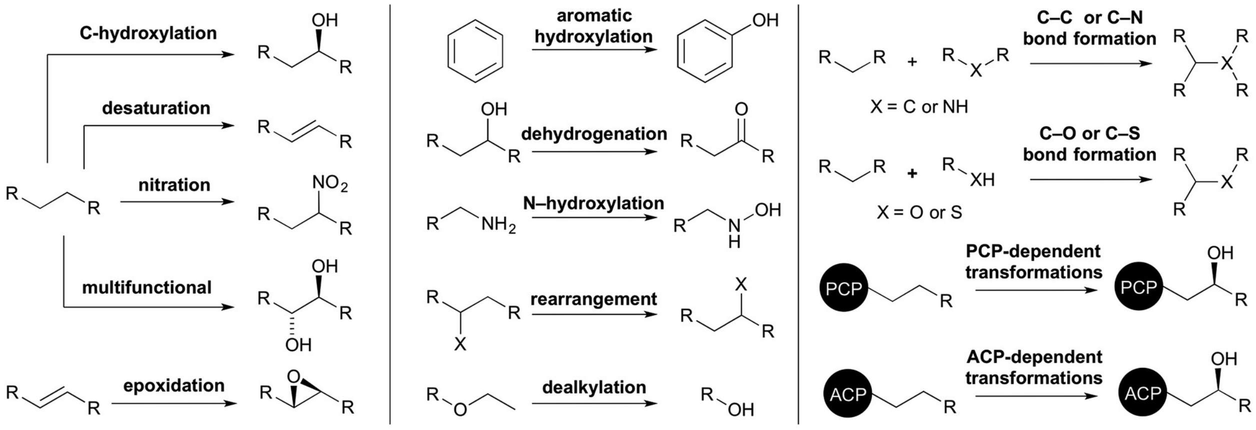

Figure 2: Metabolic reactions each class of functions of cytochrome P450 proteins catalyzes[10]

strated in Figure 2. According to the experimental charaterized streptomycete origin table, More than two-thirds of the characterized P450 [10].

By building a classification/prediction model on the functions of P450 enzymes, we can better understand their behavior and classify them based on their functions for newly discovered P450 enzymes.

2.2 Research Goal

To build a well-performed classification model on the functions of P450 proteins based on their amino acid sequences.

2.3 Related Work

Many classification methods have been used for the classification on Cytochrome P450 family using different metrics. As mentioned in Section 2.1, the naming convention is based on the gene sequence similarity. Other classification metrics have been used as well. For example, machine learning methods have been used to examine the structural features of P450 enzymes and classify based on three major isoforms of P450 protein family: CYP1A2, CYP2D6, and CYP3A4[4]. Their examination adopted many commonly used machine learning classification methods, which include k-nearest neighbors, decision tree induction using the CHAID and CRT algorithms, random forests, artificial neural networks, and support vector machines using the radial basis function (RBF) and homogeneous polynomials as kernel functions.

2.4 Contribution

Despite the fact that there are many classification metrics that have been used in past research, this project focus on the functions of cytochrome P450 enzymes based on the amino acid sequences, which associates with the gene sequences and structural features mentioned in Section 2.3. However, the key difference is that this project directly connects the metabolic functions and amino sequences, leaving out the classification metrics involving structural features.

3 Methodology

3.1 Data processing

The P450 protein amino acid sequences are separated based on their functions into two groups: Group 1 contains all P450 proteins with function hydroxylation; group 2 contains the sequences with the other function. We build a Swith-Waterman[11] based pairwise local alignment program using the BLOSUM62[5] scoring matrix and different gap penalties. It process the data from Fasta[3] format into pairwise alignment scores for future processing. The gap penalty in pairwise alignment is the variable in this step. We hope to conclude the best gap penalty value for this purpose.

3.2 Classification

3.2.1 K-Nearest Neighbors Model[2]

Based on the pairwise similarity scores produced by the Smith-Waterman based pairwise local alignment program, we build connectivity graphs with different number of neighbors (values of k). The connectivity graph is constructed by connecting each protein node to the protein nodes which it has the k highest pairwise alignment scores with. The connectivity graphs indicate the similarity of the amino acid sequences of P450 protein. The resulting group from the classification of the protein is determined by which group proteins in the neighbors dominates over the other. In other words, it is determined by the group that has more proteins in the neighbors of the node. Using the connectivity graphs, we apply leave-one-out cross-validation for each protein node to calculate the classification accuracy of the k-nearest neighbors model. The variables in this step involve the gap penalty in pairwise alignment and the number of neighbors in k-nearest neighbors model.

3.2.2 K-Nearest Neighbors with Higher Depth

For this model, we explored the neighbors with higher depth and want to see how it will affect the classification accuracy. We construct the connectivity graphs in the same way in Section 3.2.1. We perform depth-first search[12] from each protein node on the k connectivity graph with different depth restriction values. Then, the model classifies the proteins based on the dominant group of proteins represented by visited nodes during the depth-first search. Similar to the process in Section 3.2.1, we used leave-one-out cross-validation to measure the classification accuracy of the model. The variables involved in this model is the gap penalty in pairwise alignment, the number of neighbors in k-nearest neighbors model, and the depth restriction of visited nodes during depth-first search in the connectivity graph.

3.3 Additional Features

3.3.1 Protein Domain Similarity

Protein domains are distinctive functional and structural units in proteins. By comparing the domain similarity of each pair of P450 proteins, we compare their similarity in functions in another way.

From the Pfam-A protein family Hidden Markov Model[9], we retrieve the domains of P450 proteins in each group by using HMMer software[1] searching the sequences on the hidden markov model and filter the domain sets of each protein based on different threshold values for e-values for each domain. If e-value is low, the higher chance the protein is associated with this domain. Then, we try to classify them based on the Jaccard similarity coefficient[6] of domain sets of the proteins in group1 and group2. The variable used in this feature is the threshold value for the e-values of domains of each protein family.

4 Results

4.1 Experimental Setup

In Section 3, we introduced the variables used in each step. Table 1 contains the values of these variables we used in our project.

Procedure Name	Variable Name	Values Used
Alignment	Gap penalty	0,1,3,5
K-nearest neighbor	number of neighbors (value of k)	1,3,5,7
Depth-first search	depth restriction	2,5,10
Domain similarity	Threshold of e-values	0.1, 0.5, 1, 5

Table 1: Major variable settings in each procedure

4.2 Experimental Evaluations

4.2.1 Pairwise Similarity Score Distribution

Figure 3 displays the distribution of the resulting similarity scores of the two groups. From the figure, pairwise alignment score distributions with different gap penalty do not display a distinctive separation between scores in group 1, group 2, or inter group alignment scores for all different gap penalty values. Therefore, from the raw alignment scores, there is no noticeable classification method that can be applied.

4.2.2 Accuracy of K-Nearest Neighbors

The simple kNN model classification performs well and achieves an accuracy rate above 85% for proteins with hydroxylation function (group 1) but only achieves an accuracy

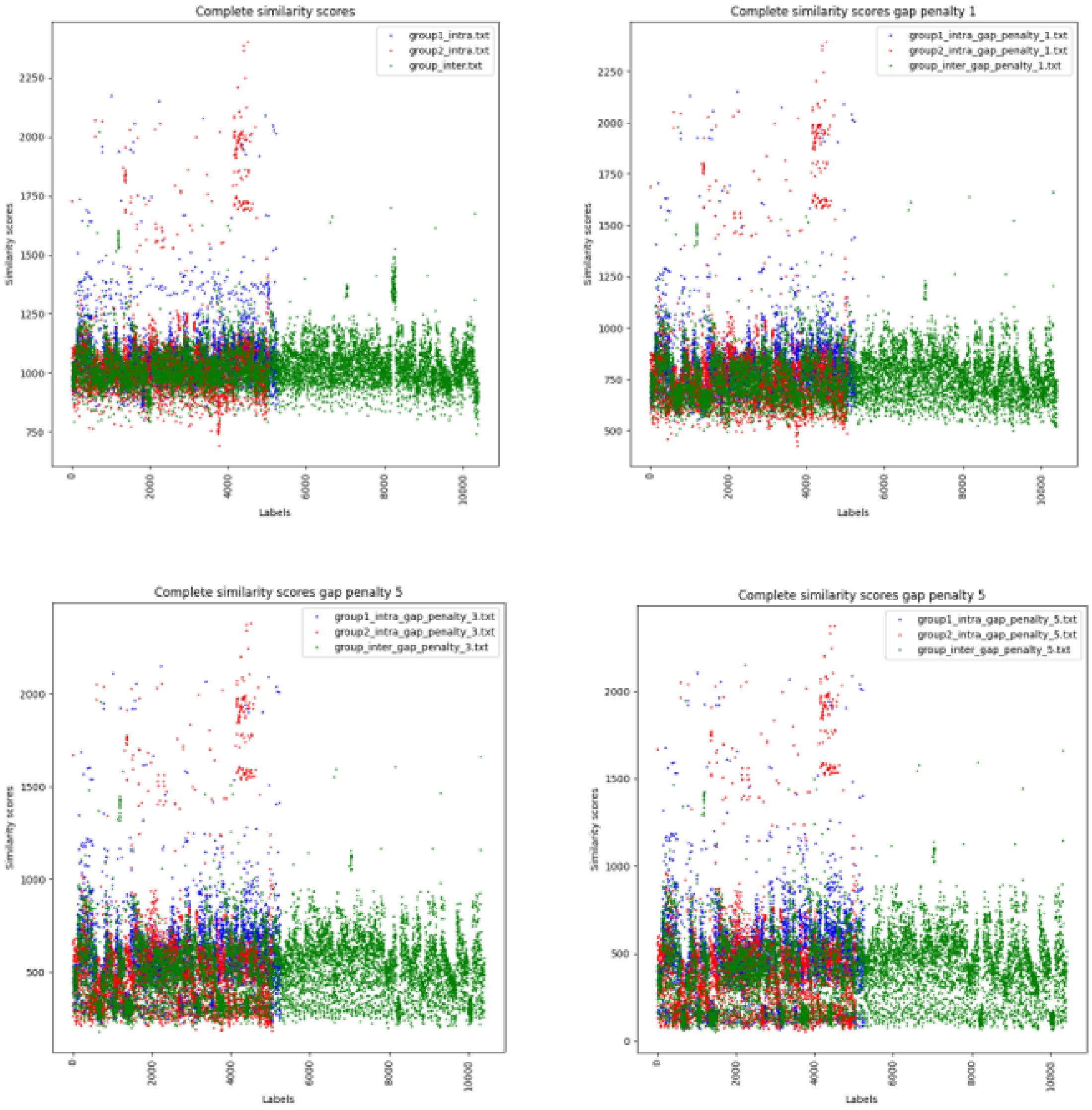

Figure 3: Pairwise alignment scores distribution based on different gap penalty values of 0, 1, 3, 5. X axis does not have any meaning except for indicating different pairs of proteins

of around 65% for proteins in the other group. However, the accuracy increases when the gap penalty gets either lower or higher. The accuracy increases as the number of neighbors (value of k) increases. Figure 4 and Table 4.2.2 demonstrates the accuracy of k-nearest neighbors model for different set of parameters. We see that the best classification accuracy is 91.3% for group 1 and 82.2% for group 2. The best total accuracy is 80.0%.

Gap penalty	Connectivity	Group 1 Accuracy	Group 2 Accuracy	Total Accuracy
0	1	0.912621	0.603960	0.759804
0	3	0.873786	0.623762	0.750000
0	5	0.873786	0.623762	0.750000
0	7	0.854369	0.623762	0.740196
1	1	0.776699	0.821782	0.799020
1	3	0.825243	0.683168	0.754902
1	5	0.864078	0.643564	0.754902
1	7	0.854369	0.663366	0.759804
3	1	0.825243	0.772277	0.799020
3	3	0.854369	0.663366	0.759804
3	5	0.825243	0.653465	0.740196
3	7	0.893204	0.663366	0.779412
5	1	0.805825	0.752475	0.779412
5	3	0.844660	0.653465	0.750000
5	5	0.864078	0.643564	0.754902
5	7	0.893204	0.643564	0.769608

Table 2: Accuracy of the k-nearest neighbors model of different parameter settings

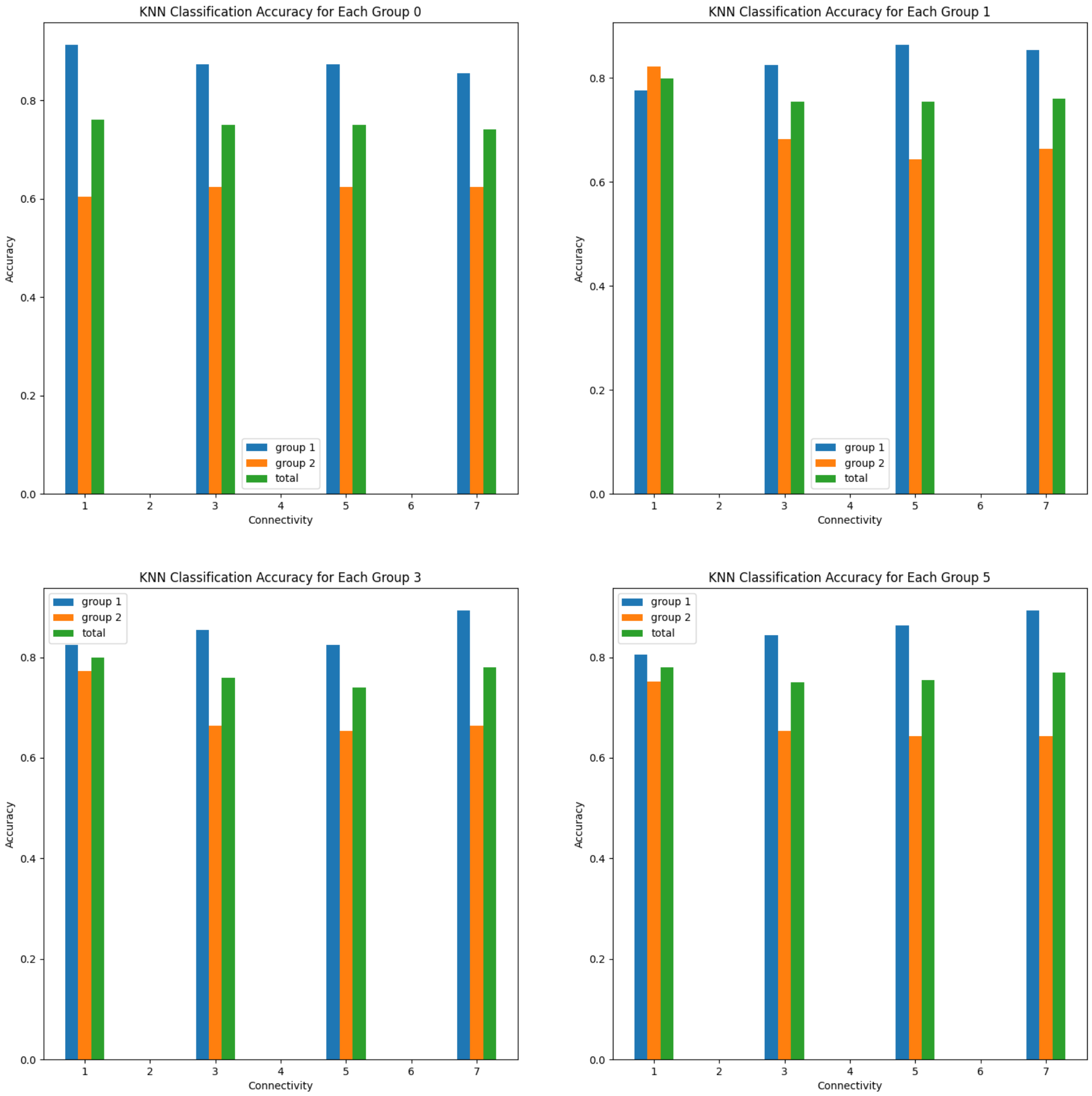

Figure 4: Accuracy of kNN model classification against different k values with different gap penalty 0, 1, 3, 5.

4.2.3 K-Nearest Neighbors Model with Higher Depth

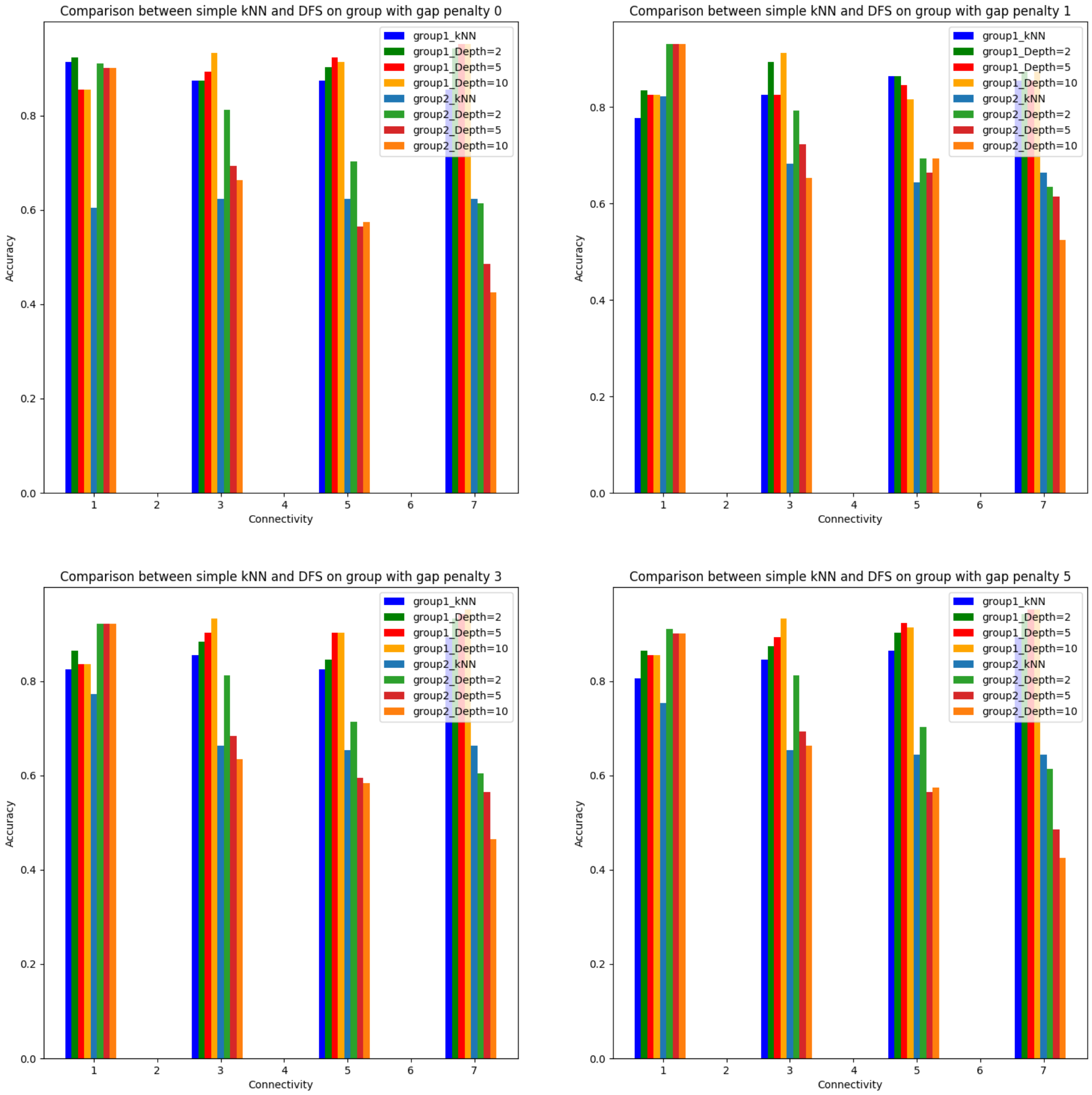

Figure 5: Comparison of results between kNN and DFS with different depths on pairwise similarity score with gap penalty 0, 1, 3, 5.

By examining neighbors with a higher depth, the classification accuracy increases in general, especially when the depth has a low value. Table 3 contains the accuracy for different sets of parameters. Figure 5 demonstrates the comparison of accuracy between simple k-nearest neighbors model and k-nearest neighbors with high depth model. For group 1, the best accuracy is improved to a value of 95.14% when the gap penalty, connectivity, and depth restriction are all high. For group 2, the best classification accuracy is improved to a value of 93.0% when connectivity is low. This is a significant improvement for classification in group 2.

Gap Penalty	Connectivity	Depth Restriction	Group1 Accuracy	Group2 Accuracy
0	1	2	0.922330	0.910891
0	1	5	0.854369	0.900990
0	1	10	0.854369	0.900990
0	3	2	0.873786	0.811881
0	3	5	0.893204	0.693069
0	3	10	0.932039	0.663366
0	5	2	0.902913	0.702970
0	5	5	0.922330	0.564356
0	5	10	0.912621	0.574257
0	7	2	0.941748	0.613861
0	7	5	0.951456	0.485149
0	7	10	0.951456	0.425743
1	1	2	0.834951	0.930693
1	1	5	0.825243	0.930693
1	1	10	0.825243	0.930693
1	3	2	0.893204	0.792079

1	3	5	0.825243	0.722772
1	3	10	0.912621	0.653465
1	5	2	0.864078	0.693069
1	5	5	0.844660	0.663366
1	5	10	0.815534	0.693069
1	7	2	0.873786	0.633663
1	7	5	0.854369	0.613861
1	7	10	0.873786	0.524752
3	1	2	0.864078	0.920792
3	1	5	0.834951	0.920792
3	1	10	0.834951	0.920792
3	3	2	0.883495	0.811881
3	3	5	0.902913	0.683168
3	3	10	0.932039	0.633663
3	5	2	0.844660	0.712871
3	5	5	0.902913	0.594059
3	5	10	0.902913	0.584158
3	7	2	0.932039	0.603960
3	7	5	0.941748	0.564356
3	7	10	0.951456	0.465347
5	1	2	0.864078	0.910891
5	1	5	0.854369	0.900990

5	1	10	0.854369	0.900990
5	3	2	0.873786	0.811881
5	3	5	0.893204	0.693069
5	3	10	0.932039	0.663366
5	5	2	0.902913	0.702970
5	5	5	0.922330	0.564356
5	5	10	0.912621	0.574257
5	7	2	0.941748	0.613861
5	7	5	0.951456	0.485149
5	7	10	0.951456	0.425743

Table 3: Accuracy of k-nearest neighbors with high depth model of different parameter settings

4.2.4 Protein Domain Similarity

Despite what we have discussed in Section 3.3.1, we do not see a distinction between the groups based on the collected domain similarity data. Figure 6 displays the frequency of the Jaccard similarity coefficient for different groups with different threshold values for the e-value. However, domains similarity does not display a distinctive difference based on Jaccard indexes between group 1, group 2, or inter group for different thresholds for e-value.

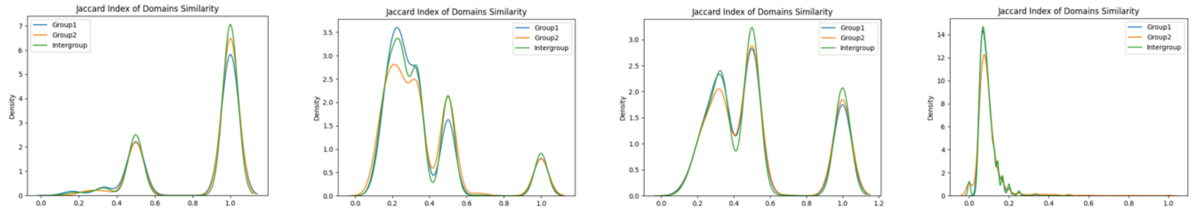

Figure 6: Jaccard index distribution of domain similarity based on e-value threshold values of 0.1, 0.5, 1, 5.

5 Surprises

For the positive side, we did not expect that by extending k-nearest neighbors with higher depths can boost the classification accuracy greatly. This model chooses to include different neighbor nodes which are of higher depths instead of the most similar neighbors. However, the reason why there is a such great improvement in classification accuracy especially for group 2 remains undiscovered.

On the other hand, the unexpected factor that does not boost the accuracy is that the protein domain similarity does not behave in distinctive ways for different groups. According to the implication behind domains, they represent the functional units in proteins, which should be closely related to our classification metric. One possible explanation is that, the thresholds for the parameter e-value we choose are too high to be effective. That is, the threshold values are too high so that they are no longer a deterministic measure for filtering related protein domains for each protein.

6 Conclusions and Future work

In conclusion, the simple kNN model classification performs well and achieves an accuracy rate of 91.3% for proteins with hydroxylation function when gap penalty is 0 and connectivity is 1, but performs not as well for non-hydroxylation proteins (group 2) with an accuracy rate of 82.2% when gap penalty is 1 and connectivity is 1. The total accuracy reaches 80% at best behavior when gap penalty is 1 and connectivity is 1. By investigating the neighbors of a higher depth of each protein node in the connectivity graph, the accuracy of classification for hydroxylation proteins (group 1) reaches 95.14% when gap penalty is 0, connectivity is 7, and depth restriction is 5 or 10. The accuracy of classifying non-hydroxylation proteins in group 2 increases greatly than that of simple k-nearest neighbors as it reaches 93.06% when gap penalty is 1, connectivity is 1, and any depth restriction values.

In summary, the k-nearest neighbors with high depth model performs better than simple k-nearest neighbor especially for non-hydroxylation proteins when connectivity is low. For the parameter settings that simple k-nearest neighbors model performs well with, the k-nearest neighbors with high depth model performs better with higher accuracy.

To improve the model with even better accuracy, different parameter settings can be tried and combined together to see if the accuracy is affected. Also, the reason of why the k-nearest neighbors with high depth model outperforms the simple k-nearest neighbors model should be investigated as well. Lastly, building classifiers on functions not limited to hydroxylation/non-hydroxylation of P450 proteins is another to extend this project.

References

[1] Hmmer. http://hmmer.org. Accessed: 2021-05-07.

[2] Evelyn Fix and Joseph L. Hodges. Discriminatory analysis. nonparametric discrimination: Consistency properties. *USAF School of Aviation Medicine, Randolph Field, Texas.*, 1951.

[3] Tom Goldstein, Christoph Studer, and Richard Baraniuk. A field guide to forward-backward splitting with a fasta implementation. *arXiv eprint*, abs/1411.3406, 2014.

[4] Baumann U Helma C Drewe J. Hammann F, Gutmann H. Classification of cytochrome p(450) activities using machine learning methods. *Mol Pharm.*, 6(6):1920–6, 2009.

[5] S. Henikoff and J.G. Henikoff. Amino acid substitution matrices from protein blocks. *PNAS*, 89(22):10915–10919, 1992.

[6] Paul Jaccard. The distribution of the flora in the alpine zone.1. *New Phytologist.*, 11(2):37–50, 1912.

[7] T Lynch and A Price. The effect of cytochrome p450 metabolism on drug response, interactions, and adverse effects. *Am Fam Physician*, 76(3):391–396, 2007.

[8] AM McDonnell and CH Dang. Basic review of the cytochrome p450 system. *J Adv Pract Oncol*, 4(4):263–268, 2013.

[9] Jaina Mistry, Sara Chuguransky, Lowri Williams, Matloob Qureshi, Gustavo A Salazar, Erik L L Sonnhammer, Silvio C E Tosatto, Lisanna Paladin, Shriya Raj,

Lorna J Richardson, Robert D Finn, and Alex Bateman. Pfam: The protein families database in 2021. *Nucleic Acids Research*, 49(D1):D412–D419, 10 2020.

[10] Ma M Shen B Rudolf JD, Chang CY. Cytochromes p450 for natural product biosynthesis in streptomyces: sequence, structure, and function. *Nat Prod Rep*, 34(9):1141–1172, 2017.

[11] Michael S. Smith, Temple F. Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology*, 141(1):195–197, 1981.

[12] R TARJAN R. Depth- first search and linear graph algorithms. pages 114–121, January 1971. IEEE Conf Rec 1971 12th annu symp on switching amp;amp; automata theory ; Conference date: 13-10-1971 Through 15-10-1971.

[13] GR Wilkinson. Drug metabolism and variability among patients in drug response. *N Engl J Med*, 352(21):2211–21, 2005.