Prediction and Classification Model on Functions of Cytochromes P450 Enzymes

Katrina Liu

Carnegie Mellon Univerisity

Introduction

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

Methodology

- Data preprocessing: Separate the P450 protein amino acid sequences based on two groups: Group 1 contains all P450 proteins with function hydroxylation; group 2 contains the other sequences.
- Smith-Waterman: Build a Swith-Waterman based pairwise local alignment program using the BLOSUM62 scoring matrix and different gap penalties. Variable(s): gap penalty in pairwise alignment.
- K-Nearest Neighbor: Based on the pairwise similarity scores, build a connectivity graph. Based on the connectivity graph, use leave-one-out cross-validation for each node to calculate the accuracy of the model with different k values. Variable(s): gap penalty in pairwise alignment, number of neighbors in k-Nearest Neighbors model.
- Depth-First Search: Perform DFS search from each node on the k connectivity graph differently based on different depths and different k values. Classify based on groups of proteins represented by visited nodes. Variable(s): gap penalty in pairwise alignment, number of neighbors in k-Nearest Neighbors model, depth restriction of DFS search of the neighbors of each protein in the connectivity graph.
- Domain similarity: From the Pfam-A protein family HMM, retrieve the domains of P450 proteins in each group and filter them based on different threshold values for e-values, and classify them based on their Jaccard index with group1 and group2. Variable(s): threshold value for the e-values of domains of each family.

Variable settings

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
HMM domain similarity	Threshold of e-values	0.1, 0.5, 1, 5

Table 1. Major variable settings in each step

Results

Pairwise local alignment with BLOSUM62[4]

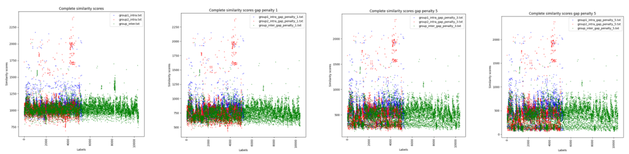

Figure 1. Pairwise alignment scores distribution based on different gap penalty values of 0, 1, 3, 5. X axis does not have any meaning except for indicating different pairs of proteins.

Simple k-nearest neighbors classification accuracy

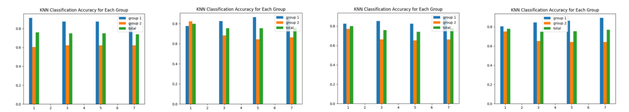

Figure 2. Accuracy of kNN model classification against different k values with different gap penalty 0, 1, 3, 5.

DFS incorporated classification

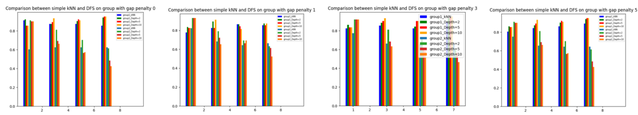

Figure 3. Comparison of results between kNN and DFS with different depths on pairwise similarity score with gap penalty 0, 1, 3, 5.

Jaccard indices frequency of protein domains

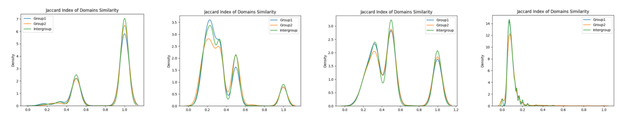

Figure 4. Jaccard index distribution of domain similarity based on e-value threshold values of 0.1, 0.5, 1, 5.

Analysis

1. Pairwise similarity scores distribution does not display a distinctive separation between scores in group 1, group 2, or intergroup for all different gap penalty values.
2. The simple kNN model classification performs well and achieves an accuracy rate above 85hydroxylation function (group 1) but only achieves an accuracy of around 65increases when the gap penalty gets either low or high. The accuracy increases as the value of k increases.
3. By adopting the strategy of DFS, the classification accuracies increase in general, especially when the depth has a low value. For group 1, the accuracy is improved for all sets of parameters. But for group 2, the accuracy is improved significantly when connectivity (value of k) and depth of DFS search are low.
4. Domains similarity does not display a distinctive difference between Jaccard indexes in group 1, group 2, or intergroup for different thresholds for e-value.

Conclusion

In conclusion, the simple kNN model classification performs well and achieves an accuracy rate of around 85% for proteins with hydroxylation function, but performs not as well for non-hydroxylation proteins (group 2) with an accuracy rate of around 65%. By employing DFS to investigate the neighbors with a higher depth of each protein in the connectivity graph, the accuracy of classification for hydroxylation proteins (group 1) increases for all cases. However, the accuracy of classifying non-hydroxylation proteins in group 2 is higher than that of k-NN when the connectivity (value of k) and depth of DFS search is low.

References

[1] Evelyn Fix and J. L. Hodges.
Discriminatory analysis. nonparametric discrimination: Consistency properties.
International Statistical Review / Revue Internationale De Statistique, 57(3):238–247, 1989.

[2] Ming Ma Jeffrey D. Rudolf, Chin-Yuan Chang and Ben Shen.
Cytochromes p450 for natural product biosynthesis in streptomyces: sequence, structure, and function.
Nat. Prod. Rep., 34:1141–1172, 2017.

[3] 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.

[4] J G Henikoff S Henikoff.
Amino acid substitution matrices from protein blocks.
Proceedings of the National Academy of Sciences, 89(22):10915–10919, 1992.

[5] M.S. Waterman T.F. Smith.
Identification of common molecular subsequences.
Journal of Molecular Biology, 147(1):195–197, 1981.