Additional Paper for “Protein Classification using Machine Learning”

Paper reference:

Abstract:
The Authors present an algorithm to apply likelihood ratio approximants (LRA) in protein classification and ranking. Classification (and ranking) is done based on the minimum distance between the highest ranking members of the two classes (of which one is the class in ‘contention’ and the other class is the complementary class). The first step towards such a classification is to have a set of predefined classes and apply sequence similarity detection (Alignment Based or Alignment Free) methods.
A similarity function is then obtained— identical proteins have a distance of zero, while proteins from different families have a higher distance measure, and correspondingly a lower similarity. The sequence similarity scores are obtained from some alignment algorithm like BLAST, Smith-Waterman, etc.

Similarity $s$ is mapped to Distance $d$ using some monotonically decreasing function $f$, $d \sim f(s)$. $f$ is usually an exponential based function, with a user-controllable parameter $\beta$ thrown in.

In this study, three different datasets have been used to obtain different classification scenarios. The first case study attempts to classify a protein based on other families from the same super group. In the second sequence, there exists a strong similarity between members of the same group, and also between members of different groups. In the third dataset, there is relatively high similarity between members of the same group, but members of different groups are dissimilar.

It was observed that using LRA scoring as a post processing step after Alignment based or Alignment free algorithms substantially improves the ranking efficiency.

Discussion:
This application is based on the Neyman-Pearson Formulation that for simple point hypothesis testing, the LR method is most optimal. Protein classification is more complex than a simple binary decision; hence the authors have decided to base the decision on a minimal distance approach between the highest ranking members of the given classes. Also, the equation is used as a scoring function; so there is no user defined threshold as a classifier, which is good because arbitrary inputs are to be minimized.

Another advantage of the scoring approach is that it provides the ability to be able to create a new sub-class based on the similarity measures obtained, which would not be possible based on original scores without LRA.

The Machine learning techniques discussed in class are more efficient in placing a protein under a new sub-group, but the AUC levels would definitely show improvement if the LRA is used as a post-processing step.

LRA does provide a consistent performance increase in protein sequence classification tasks, so as a standalone algorithm, LRA works, but more experiments would be needed to determine if it can be integrated with other algorithms leading to an optimal classifier.

It would be good to test the algorithm over other sequences to gauge the extent to which the efficiency depends on the kind of data set, and the accuracy with which sub-groups can be recognized.