Remarks

A) Approximating class conditional probabilities using sequence similarity/dissimilarity functions

Let denote *A* a finite set of the proteins in the database that consist of two classes called "+" and "-". We partition the set *A* into two subsets called a *train set* and a *test set* and call the positive and negative elements in the train set *train+* and *train-*, respectively. Furthermore we define a *(sequence) similarity function* over set *A* and denote it by $h(x, y)$, where *x* and *y* are elements of *A*. This function gives a higher value on "similar" sequences, and a lower value on "different" ones. We can define in analogous way a *(sequence) dissimilarity function* $l(x, y)$ which gives a zero or minimal value on equal or "similar" sequences, and a high value on "different" sequences. Since any dissimilarity function can be transformed into a similarity function by a monotone decreasing function, we can restrict our investigations only similarity functions.

For any protein $x \in A$ and sequence similarity function h we define the similarity between x and class "+" by $s(x,+) = \max_{y \in train+} h(x, y)$, and similarly, between x and class "-" by $s(x,-) = \max_{y \in train-} h(x, y)$, respectively. The class-conditional probability functions $p(x|+)$ and $p(x|-)$ can be approximated by

$$p(x|+) \approx \frac{s(x,+)}{M_+}, \tag{1}$$

$$p(x|-) \approx \frac{s(x,-)}{M_-}, \tag{2}$$

where $x \in A$, $M_+ = \sum_{a \in A} s(a,+)$ and $M_- = \sum_{a \in A} s(a,-)$. These are valid probability functions, because their range is [0,1] and their sum over all proteins is equal to 1.

B) The Log Likelihood Ratio Approximant

Now, based on Eqs. (1-2) we have the following estimation for the log likelihood ratio:

$$\begin{aligned} \log LR(x) &= \log\left(\frac{p(x|+)}{p(x|-)}\right) \approx \log\left(\frac{s(x,+)/M_+}{s(x,-)/M_-}\right) \\ &= \log\left(\frac{s(x,+)}{s(x,-)}\right) + \log\left(\frac{M_-}{M_+}\right) \end{aligned} \tag{3}$$

The *bias* term in the last expression is independent of the element *x*, so it does not count in the ranking of the elements of *A*, i.e it does not affect the ROC analysis and the AUC value, as well. This is why we can define the log likelihood ratio approximant by

$$LRA(x) \sim \left(\frac{s(x,+)}{s(x,-)}\right)$$

in the paper. Basically, ROC curves and AUC calculations are used to investigate the performance of learning algorithms under changing conditions such as misclassification costs or class distributions. For practical purposes one can build classifiers using threshold values, which imply the bias term appeared in Eq. (3).