Multiple alignment search for conserved motifs – python project

Background

In field of bioinformatics, the identification of conserved motifs within protein sequences plays a crucial role in understanding structure and function. Typically, sequence alignment tools like BLAST (Basic Local Alignment Search Tool) are employed for this purpose, however this method is somewhat limited as it only permits for the search of a singular motif. Within homologous proteins, amino acid substitutions, insertions and deletions frequently occur whilst preserving protein function, resulting in variable levels of motif conservation across different systems. A common method for depicting the conservation of a particular motif is by the use of histograms known as WebLogos (Fig 1). These provide a graphical representation of sequence consensus across multiple motifs. In a WebLogo, each letter in a stack represents the amino acids present in Nth position, with the size of each letter corresponding to its frequency in that position. The overall stack height, measured in an arbitrary unit ‘bits’, indicates the degree of conservation at each position When assessing the similarity of a particular protein to a family of homologous proteins, searching against multiple sequence motifs proves more insightful than using a single sequence alone. The following project allows for analysis of protein sequences against multiple motifs, providing a numerical score for similarity based on the relative occurrence of the amino acids at each position.

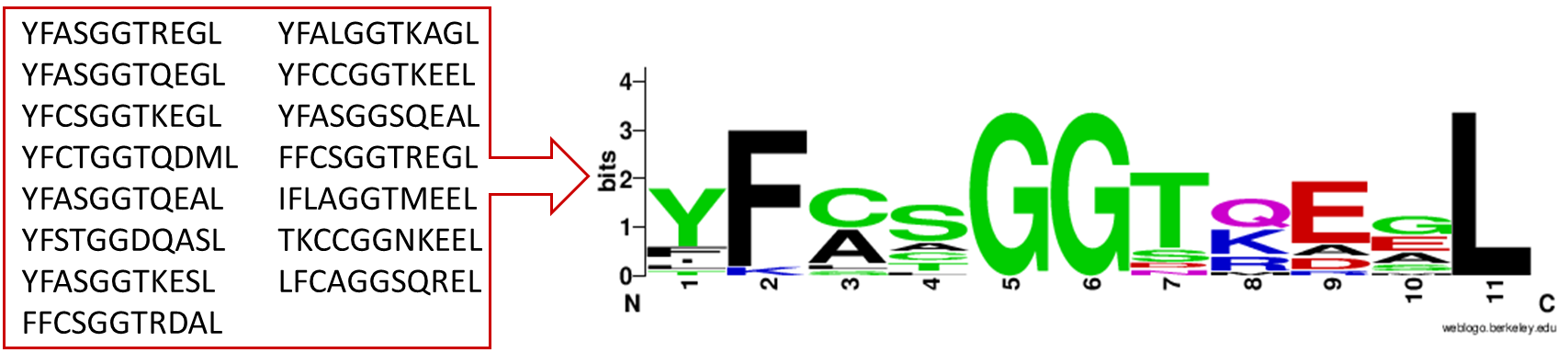


Figure 1 - Diagram showing the conversion of a list of motifs to a WebLogo

Discussion

The script was written in python and takes two input files: a CSV containing details of the proteins to be searched and a text file with a list of all the motifs to search against. The code is written such that it can allow for identification of not only the full motif but also for aligning sequences with 3 amino acids or greater (incomplete searches). This feature proves valuable in scenarios where an unexpected mutation occurs, as it enables the identification of incomplete motifs which can be checked against other incomplete motifs in close proximity. The script could be further developed to account for this issue by allowing for one misalignment in each motif, however such modification would require adjustments to other areas of the code and a different method for scoring may need to be considered. The outputted similarity scores for the identified motifs is determined by assessing alignment with the occurring amino acids in each position which are weighted relative to their frequency. Achieving a score of 1 corresponds to complete alignment with the most common amino acid in each position (in the case for fig 1 this would be the motif YFCSGGTQEGL). For the sequences found which are shorter than the full length of the motif a relative score is also given which accounts for the shorter length (the absolute score divides by the full length of the motif). A further beneficial addition to the code could involve checking each motif from the text file to see if an exact match occurs with any of the listed sequences.

Conclusion

The outcomes of the project have proven useful in preventing the oversight of potential conserved motifs in homologous proteins. Furthermore, the provision of a numerical similarity score allows for comparisons of conservation across multiple proteins. The script also exhibits potential for additional development to incorporate further functionalities as discussed above.