ABSTRACT:

Remote protein homology detection is a problem of detecting evolutionary relationship between proteins at low sequence similarity level. Among several problems in remote protein homology detection include the questions of determining which combination of multiple alignment and classification techniques is the best as well as the misalignment of protein sequences during the alignment process. Therefore, this paper deals with remote protein homology detection via assessing the impact of using structural information on protein multiple alignments over sequence information. This paper further presents the best combinations of multiple alignment and classification programs to be chosen. This paper also improves the quality of the multiple alignments via integration of a refinement algorithm. The framework of this paperbegan with datasets preparation on datasets from SCOP version 1.73, followed by multiple alignments of the protein sequences using CLUSTALW, MAFFT, ProbCons and T-Coffee for sequence-based multiple alignments and 3DCoffee, MAMMOTH-mult, MUSTANG and PROMALS3D for structural-based multiple alignments. Next, a refinement algorithm was applied on the protein sequences to reduce misalignments. Lastly, the aligned protein sequences were classified using the pHMMs generative classifier such as HMMER and SAM and also SVMs discriminative classifier such as SVM-Fold and SVM-Struct. The performances of assessed programs were evaluated using ROC, Precision and Recall tests. The result from this paper shows that the combination of refined SVM-Struct and PROMALS3D performs the best against other programs, which suggests that this combination is the best for RPHD. This paper also shows that the use of the refinement algorithm increases the performance of the multiple alignments programs by at least 4%.