Texila International Journal of Public Health
ISSN: 2520-3134
DOI: 10.21522/TIJPH.2013.12.04.Art054

Enhanced HACBLalign Method using Transitional Pattern Search and Pre-Trained Classification Model for Protein Remote Homology Detection and Fold Recognition

Gopinath Krishnaraj[1,2*], Rajendran Gurusamy[3]
[1]*Department of Computer Science, Periyar University, Salem– 636 011, Tamil Nadu, India*
[2]*Department of Computer Applications, Sona College of Arts and Science, Salem – 636 005, Tamil Nadu, India*
[3]*Department of Computer Science, Govt. Arts and Science College, Modakkurichi, Erode – 638104, Tamil Nadu, India*

Abstract

One of the most important tasks to predict the structure of proteins is Protein Remote Homology Detection and Fold Recognition. To do this, a Hierarchical Attention-based Convolutional Neural Network with Bidirectional Long Short-Term Memory called the HACBLalign algorithm was proposed by the authors, which performs Multiple Sequence Alignments (MSAs), extracts features, and recognizes protein homologies. But, when the quantity of Protein Sequences (PSs) increases, the number of times the decision-making system runs also increases. To avoid this issue, this article proposes an Enhanced HACBLalign (EHACBLalign) method using Transitional Pattern Search (TPS) and pre-trained classification for Protein Remote Homology Detection and Fold Recognition. During the alignment stage, the intermediate sequences such as Hit Regions (HRs) are identified by the TPS. Then, the HRs are extended in middle layers and utilized as a query in all TPS iterations. Besides, the HACBLalign algorithm is applied in all intermediate layers for generating pairwise alignments. Moreover, each pairwise alignment between intermediate sequences is merged to get the final alignment. Further, various characteristics are obtained from the chosen alignment and learned by the pre-trained Convolutional Neural Network (CNN) with a softmax function for recognizing protein remote homologies precisely. This enhances the performance of the decision-making system for large-scale PS databases. Finally, the test outcomes exhibit that the EHACBLalign realizes a 94.6%, 94.1%, and 93.4% accuracy on SCOP 1.53, SCOP 1.67, and superfamily corpora, respectively in Protein Remote Homology Detection and Fold Recognition.

Keywords: *Convolutional Neural Network (CNN), Fold Recognition, HACBLalign, MSA, Protein Remote Homology, Transitional Pattern Search.*

Introduction

Protein Remote Homology Detection and Fold Recognition is a term used in bioinformatics to describe the classification of proteins into morphological and chemical categories. It may be utilized to discover the 3D shape and activity of molecules in both basic investigations and therapeutic applications [1]. Because of the small homogeneity of PSs, amino genome sequencing has trouble discovering remote homologies of proteins. Sequence-based, ranking-based, and discriminative-based approaches are the majority of the standard predictors. The homology of amino acids can be assessed by sequence-based alignment methods [2, 3]. The effectiveness of correlation estimation could be improved by

Received: 16.08.2024 ***Accepted: 13.10.2024*** ***Published on: 27.12.2024***
****Corresponding Author: vengatgopinath@gmail.com***

Hidden Markov Model (HMM)-based strategies [4].

It is possible to estimate protein homologies using a probabilistic method with the help of the HMMER software package [5]. Ranking methods/models that use correlation coefficients to determine rankings calculate the homology proximity between the nucleotides. ProtEmbed, RankProp, and other well-known ranking methods are also available [6]. Motifs are essential to protein morphologies to calculate correlation coefficients. MotifCNN and MotifDCNN are two traditional motifs-based feature extraction schemes [7]. According to the ground truth labels of the protein families, the supervised training paradigm utilized by the classification approaches could be transformed into a binary classification. Profile-based traits like profile kernel, All Fixed-width subsequences with PSSM, and the Smith-Waterman algorithm PSSM improve detectability [8]. Two kernels for SW-PSSM are values for profile-profile correlation and values for sequence similarity [9]. Compared to existing approaches, classification methods produce cutting-edge outcomes.

Classification methods/models may swiftly combine various nucleotide sequence features and understand the relevant information from both true and false examples in a particular dataset, in contrast to pairing strategies and generation strategies. A crucial prerequisite of classification approaches is the need for features extracted with fixed lengths as input. ReFold-MAP, a unique classification approach that yields complete traits by the three separate profile-based properties of motif-PSSM, ACC-PSSM, and Physicochemical Distance Transformation (PDT)-profile for MSAs, has been created in light of such viewpoints [10]. The morphological motif kernel data, the genetic data, and the nucleotide sequences are together referred to as MAP traits. Moreover, the Support Vector Machine (SVM) algorithm used this characteristic vector to identify remote protein homologies. Nonetheless, the accuracy of the sequence alignment is inconsistent, making it difficult for the traditional MSA-based algorithms to produce correct alignments. Therefore, integrated methods/models are still required to get perfect PS alignments.

To combat these problems, a novel progressive deep MSA algorithm was developed, which creates highly appropriate decision-making model MSA of low similarity protein families [11]. This algorithm uses a decision-making system that has been trained by the HACBLalign to gradually align the PSs by calculating multiple posterior probability matrices. By combining the alignment of crucial subsequences into a sequence alignment, this algorithm gradually creates a global alignment. Additionally, the attention layers enable this algorithm to pick sequences and subsequences that are qualitatively essential. As a consequence, an enhanced MSA was produced. The top-N-gram and Auto-Cross-Covariance (ACC) characteristics were retrieved from aligned PSs using the Position-Specific Scoring Matrix (PSSM). Moreover, those characteristics were classified by the CNN with a softmax function to identify protein homologies. However, the alignment efficiency degraded while increasing the quantity of PSs. If many PSs were acquired, then multiple sequence pairs were also acquired; as a result, the sum of epochs the decision-making system trains was increased.

Therefore, an Enhanced HACBLalign (EHACBLalign) method is proposed in this paper to decrease the sum of epochs the decision-making system trains for a large number of PSs. In this algorithm, the TPS is adopted for achieving appropriate alignment generation by detecting the intermediate sequences. During the alignment generation stage, a method is proposed that extends the HR recognized using the TPS and is considered as a query in every iteration of the

TPS. In addition, the HACBL align algorithm is applied in all intermediate layers for generating pairwise alignments. Moreover, each pairwise alignment between intermediate sequences is merged to get the final alignment, which is used to extract various characteristics. Further, the CNN with a softmax function pre-trained by the BERT model is applied for learning the extracted characteristics and recognizing protein remote homologies precisely. Thus, this algorithm enhances the efficiency of Protein Remote Homology Detection and Fold Recognition by appropriately generating alignments of many PSs and enhancing the decision-making system.

A novel sequence alignment creation method using the k-nearest Neighbor (kNN) algorithm for predicting protein structures. The substitution scores at all residue pairs were predicted rather than the predetermined substitution matrix. Besides, a scheme was adopted to transform pairwise arrangements to statistical vectors of latent space to predict the PS pattern. Moreover, the estimated values were utilized to create arrangements for template-based modelling. However, it needs a long execution period due to the kNN algorithm and database dimension [12].

A novel DeepMSA that comprises homologous PSs and arrangements generated from multiple sources of full and meta-genome corpora via complementary HMM algorithms. The DeepMSA was initially used to create MSAs for residue-level connection estimation using multiple coevolution and deep learners. Then, many threading methods were executed for homologous structure recognition and the DeepMSA has been utilized for secondary pattern estimation. However, it did not often provide an efficient connection estimation because the absolute impact of MSAs was a balance of arrangement exposure and precision [13].

A neural net to obtain precise estimations of the gaps among a couple of residues that take additional data regarding the pattern compared to the connection estimation. Based on this data, a latent average force was built, which precisely defines the protein structure, and the resultant potential was optimized by gradient descent to create patterns with no multifaceted sampling processes. However, it needs further enhancements to properly predict unknown sequence structures [14].

A Novel Deep-learning Threader (NDThreader) method using Deep Convolutional Residual Neural Fields (DRNF) to arrange a query PS to prototypes without considering distance data. After that, the Alternating Direction Method of Multipliers (ADMM) was used to enhance pattern-prototype arrangements using the estimated gap latent [15]. Moreover, 3D frameworks were created from a pattern-prototype arrangement by providing it and pattern coevolution data into the ResNet for estimating inter-atom gap distribution that was passed to PyRosetta for 3D framework creation. However, the problem was that many basic units were executed separately, which may have impacted the modelling accuracy.

A multi-task Deep learning Distance (DeepDist) estimator depending on novel ResNet models to concurrently estimate true inter-residue gaps and categorize them into several gap periods. However, it was not effective according to the MSE of the estimated gap [16]. An ensemble model to solve the function prediction by automatically allocating Gene Ontology (GO) expressions to the PS. In this model, the GO predictions created by random forest and neural network categorizers were combined. However, training a neural network was difficult and it needed to choose the relevant features for increasing efficiency [17].

An enhanced CASP14 MULTICOM protein structure prediction framework by integrating different novel modules: (i) a novel deep learner-based PS inter-residue gap estimator to enhance prototype-free tertiary

pattern estimation, (ii) an improved prototype-based tertiary pattern estimation method, and (iii) gap-related efficiency analysis schemes enabled by the deep learner. However, its efficiency relies on the efficiency of deep leaning-based residue-residue gap estimation that in turn relies on the efficiency of MSA [18]. A unified method called US-align2, executes sequential alignment, semi-non-sequential alignment, and fully non-sequential arrangement for PSs by a single rating factor. However, it did not take conformational dissimilarities among a couple of correlated patterns [19]. Profile-based direct kernels for remote homology detection and fold recognition [20]. The effectiveness of motif kernels generated by genetic programming for improving remote homology and fold detection [21]. The classification of protein structures in the SCOP database [22]. BERT, a deep bidirectional transformer model for enhanced language understanding [23].

Proposed Methodology

This section explains the EHACBLalign algorithm briefly. Fig.1 illustrates the architecture of the proposed Remote Homology Detection and Fold Recognition.

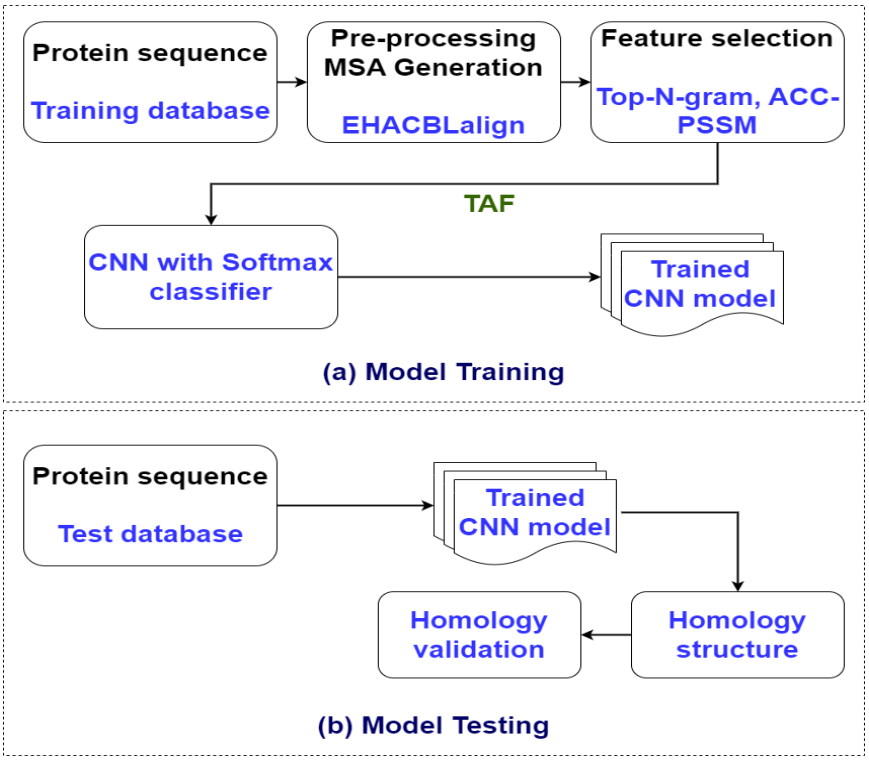

Fig. 1. Architecture of Proposed EHACBLalign Method for Protein Remote Homology Detection and Fold Recognition

Dataset Description

In this work, three benchmark datasets, namely the SCOP v1.53, SCOP v1.67, and the superfamily corpus are acquired to assess the effectiveness of selective MSA algorithms. The SCOP v1.53 corpus [20] possesses 4532 PSs from 54 groups, whilst the SCOP v1.67 [21] possesses 11037 PSs from 102 groups. The superfamily corpus [22] possesses 1195 folds of 1962 superfamilies. A superfamily is a corpus that contains labels for each PS's morphological properties. Depending on a collection of HM`Ms that represent structural protein motifs at the tier of the SCOP superfamily, it was built. The labels are produced by matching PSs from approximately 2478 fully sequenced genomes to HMMs.

Improved MSA Generation

In this work, the TPS method is used, which explores homologs of the query PS, and simultaneously utilizes the outcomes as fresh queries to identify protein remote homologs. The MSA generation is improved by two processes. Initially, a sub-area of the given PS is utilized as intermediary outcomes, which are considered as the successive query rather than considering the entire PS of the given homolog. Several PSs in the corpora have many domains within a single sequence; so, False Positives (FPs) are acquired since the fields that are not associated with the query PS are taken as successive queries in transitional explorations. By reducing the exploration space to a given homology area, it is projected that the sum of FPs can be minimized. Then, rankings to the outcomes are allocated via the calculation of relationships among intermediary PSs. The relationships are determined in the TPS. In this method, the relationship values on the route from the query to the absolute hits are summed and ranked to create absolute exploration outcomes. When many routes occur between the query and the absolute hit, the route having the optimal value is chosen. The value is utilized as a relationship value and the route of the minimum value is chosen as the optimal.

An outline of the TPS method is depicted in Fig. 2, wherein violet, yellow, and green circles symbolize the query, intermediary, and hit PSs, correspondingly. Initially, homologs of the query PS in the intermediary corpus are searched and the resultant hits define the initial intermediaries. Then, such hits are utilized as queries and the consecutive sequence of hits defines the next intermediaries. Finally, such intermediaries are utilized as queries and the absolute corpus is explored. Here, the hit value is determined by the sum of all hit values.

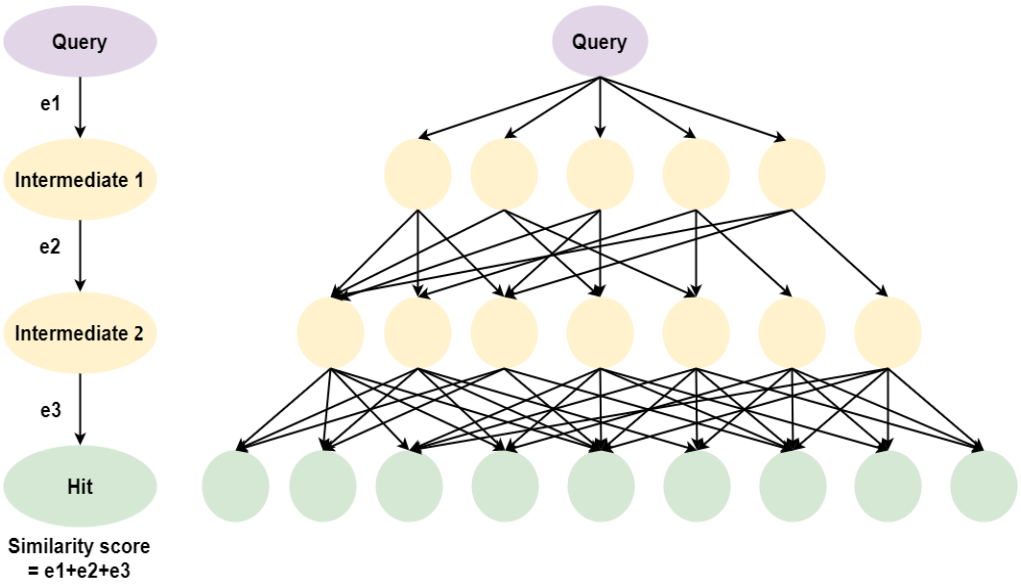

Fig. 2. Outline of Transitional Sequence Search

However, it is complex to create arrangements among remote homologs since the pattern uniqueness among them is always less. Though pairwise alignments are created, the dimension of the arrangement area is extremely tiny for proper alignment, resulting in improper recognition of protein homology structure. This issue is resolved by using intermediate sequences identified by the TPS. During the TPS stage, it utilizes only arranged pattern areas, whereas other fields are not utilized for the exploration. The HR's are extended to the intermediary layers to extend the aligned region to the degree to be likely that they do not involve other fields excessively. The dimension of the extension is one of the hyperparameters. The extended sub-sequence is taken as a query in all transitional

searches. In all intermediary layers, pairwise arrangements are created by the HACBLalign. At the last stage of arrangement formation, this method combines pairwise arrangements among intermediary PSs. All sub-pairwise alignments are conserved that define the locations of residues in a pairwise arrangement are conserved. A pairwise arrangement between the query PS and one of the absolute hits is partitioned from the combined arrangement.

An outline of the proposed alignment formation method is illustrated in Fig. 3, wherein a pairwise alignment is created in all intermediate layers. Also, each pairwise alignment between intermediate sequences is combined and conserved.

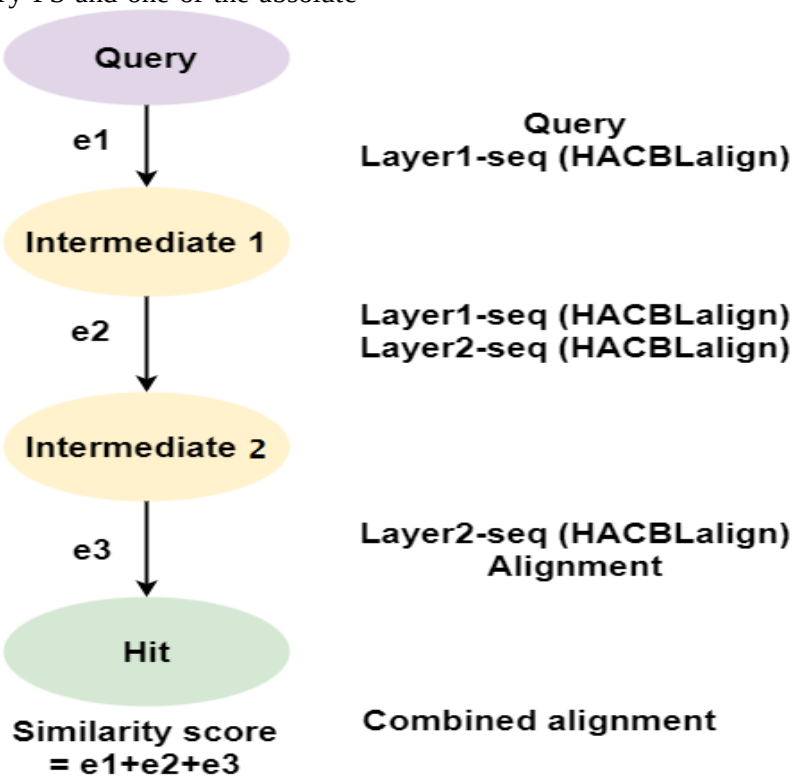

Fig. 3. Alignment Formation with Intermediate Sequences

Once the final alignment is acquired, top-N-gram and ACC-PSSM features are extracted [11]. Those extracted features are further learned by the pre-trained CNN with a softmax function for creating a trained model.

Pre-learned CNN Model for Protein Remote Homology Detection and Fold Recognition

For recognizing protein homologies, pre-learned CNN with softmax classifier is applied, which encompasses both learning and testing procedures. In the learning procedure, CNN is trained by randomly initialized weights. If the CNN is learned in an unsupervised manner like in the BERT model [23] could aid deep supervised training and enhance the efficacy of Remote Homology Detection and Fold Recognition. While utilizing the pre-learned CNN for Protein Remote Homology Detection and Fold Recognition, merely the variables of the last layer are set arbitrarily and each other variable is set to the pre-learned weights (as shown in Fig. 4). So, the feature vectors and tags from the training sequences are learned by the pre-learned CNN model. During the test process, the test sequences are transformed into the TAF vectors by a similar procedure to the learning sequences and recognized via the learned model.

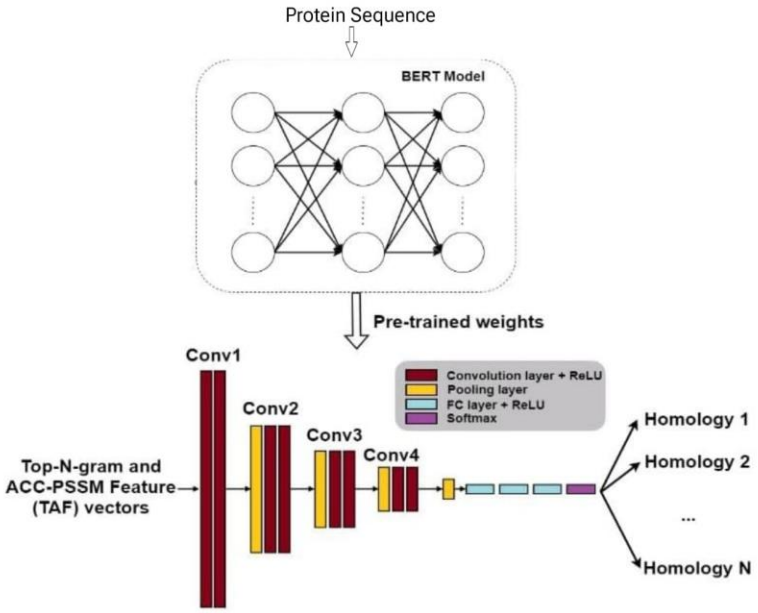

Fig. 4. Pre-trained CNN Model for Remote Homology Detection and Fold Recognition

Results

This part assesses the success of this EHACBLalign-TAF algorithm applied to the three distinct datasets using MATLAB 2019b. In this analysis, 70% of the PSs are used in the learning phase and 30% of the PSs are used in the test phase. The measured values are compared to that of previous algorithms: ReFold-MAP [10], HACBLalign-TAF [11], DeepMSA [13], NDThreader [15], DeepDist [16], and US-align2 [19] in terms of the following metrics.

It specifies the proportion of correctly identified protein homologies to all PSs examined.

$$Accuracy = \frac{True\ Positive\ (TP) + True\ Negative\ (TN)}{TP + TN + FP + False\ Negative\ (FN)}$$

For SCOP 1.53, SCOP 1.67 and Superfamily datasets, the accuracy of EHACBLalign – TAF is 94.6%, 94.1% and 93.4% receptively which is higher when compared with well-known models/methods as shown in Fig. 5.

Accuracy

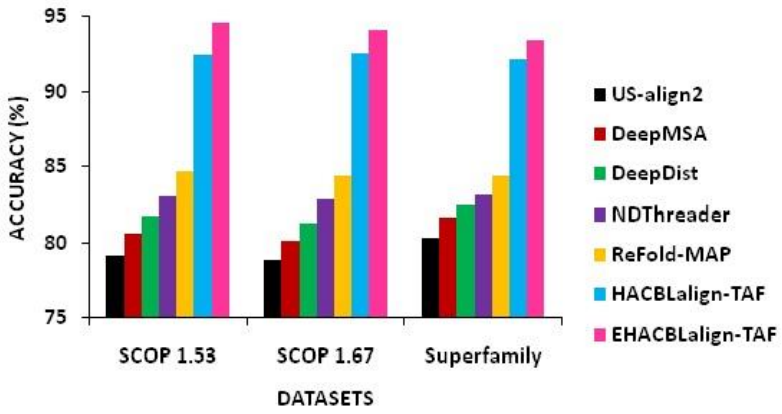

Fig. 5. Comparison of accuracy for proposed and existing Protein Remote Homology Detection and Fold Recognition Models /Methods over SCOP 1.53, SCOP 1.67 and Superfamily Data Sets

Precision

It specifies the percentage of perfectly aligned locations.

$$Precision = \frac{TP}{TP+FP}$$

For SCOP 1.53, SCOP 1.67 and Superfamily datasets, the precision of EHACBLalign – TAF is 94.3%, 93.9% and 94.2% respectively which is higher when compared with well-known models/methods as shown in Fig. 6.

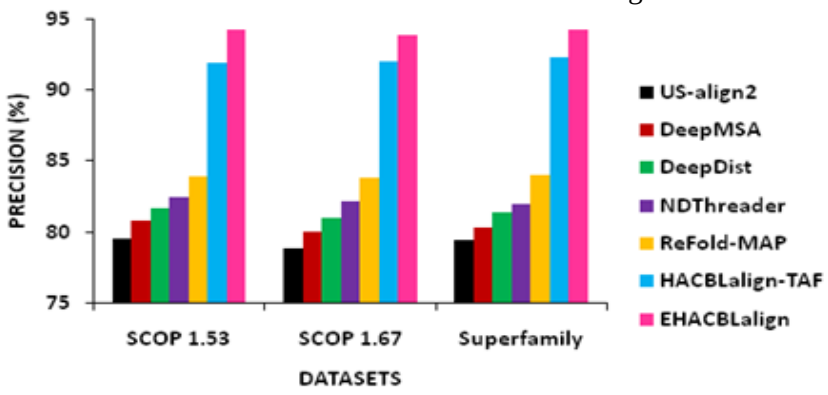

Fig. 6. Comparison of Precision for proposed and Existing Protein Remote Homology Detection and Fold Recognition Models /Methods over SCOP 1.53, SCOP 1.67 and Superfamily Data Sets

Recall

It specifies the proportion of precisely aligned residues among those that are aligned.

$$Recall = \frac{TP}{TP + FN}.$$

For SCOP 1.53, SCOP 1.67 and Superfamily datasets, the recall of EHACBLalign – TAF is 94.47%, 94.5% and 94.4% respectively which is higher when compared with well-known models/methods as shown in Fig. 7.

F – Measure

It defines the f-measure of proposed and existing Protein Remote Homology Detection and Fold Recognition techniques.

$$\text{F measure} = \frac{2 \text{ x Precision x Recall}}{\text{precision} + \text{recall}}$$

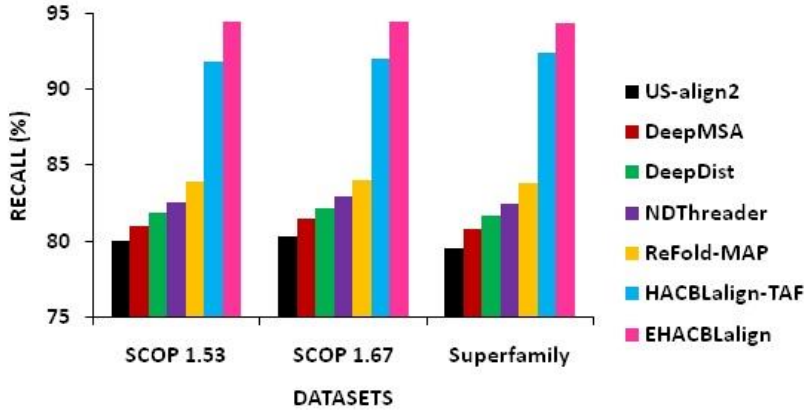

Fig. 7. Comparison of Recall for Proposed and Existing Protein Remote Homology Detection and Fold Recognition Models /Methods over SCOP 1.53, SCOP 1.67 and Superfamily Data Sets

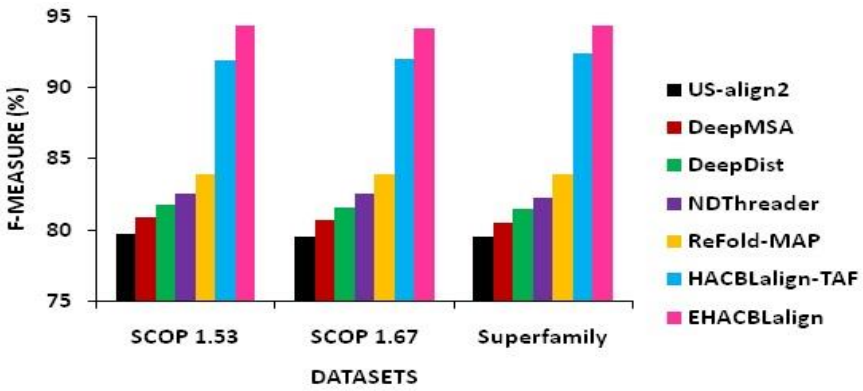

Fig. 8. Comparison of F-Measure for Proposed and Existing Protein Remote Homology Detection and Fold Recognition Models /Methods over SCOP 1.53, SCOP 1.67 and Superfamily Data Sets

For SCOP 1.53, SCOP 1.67 and Superfamily datasets, the F-measure of EHACBLalign – TAF is 94.39%, 94.2% and 94.4% which is higher when compared with well-known models/methods as shown in Fig. 8.

Receiver Operating Characteristics (ROC) and ROC50 Curve

The ratio between specificity and sensitivity is set by the ROC value. Within the normalized Area Under the Curve (AUC), it compares TP Rates (TPRs) and FP Rates (FPRs). The ROC50 value, in a similar vein, is the region of the ROC curve up to 50 FPs. By calculating the TPR and FPR as follows:

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN}$$

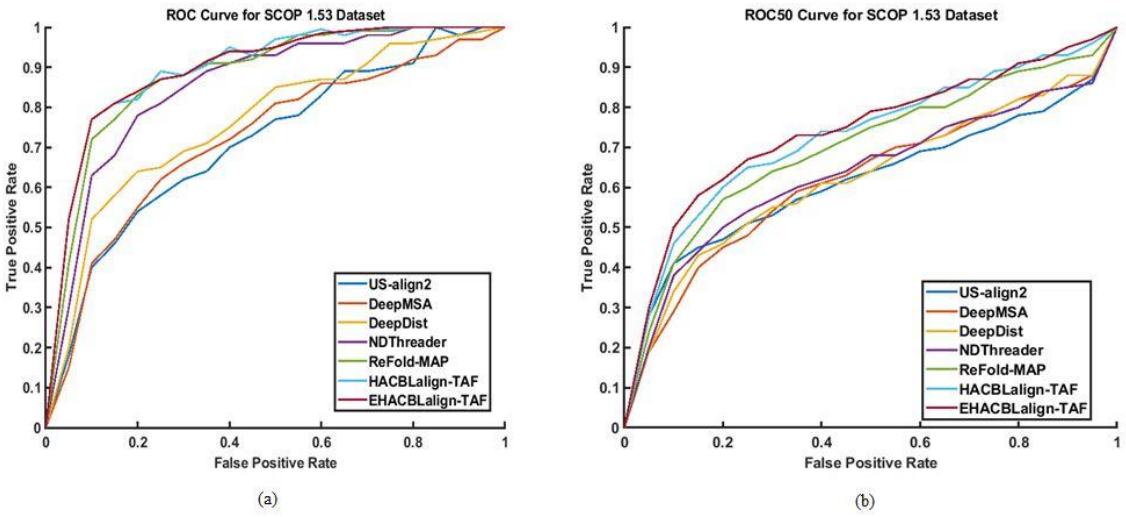

Fig. 9. Comparison of ROC and ROC 50 Curves for Proposed and Existing Protein Remote Homology Detection and Fold Recognition Models / Methods over SCOP 1.53 Dataset

Fig. 9 (a) and (b) depict the ROC and ROC50 values of various Remote Homology Detection and Fold Recognition methods tested on the SCOP 1.53 datasets. It indicates that the ROC of EHACBLalign-TAF is significantly larger than those of US-align2, DeepMSA, DeepDist, NDThreader, ReFold-MAP, and HACBLalign-TAF. Similarly, the ROC50 of EHACBLalign-TAF is markedly superior to those of US-align2, DeepMSA, DeepDist, NDThreader, ReFold-MAP, and HACBLalign-TAF.

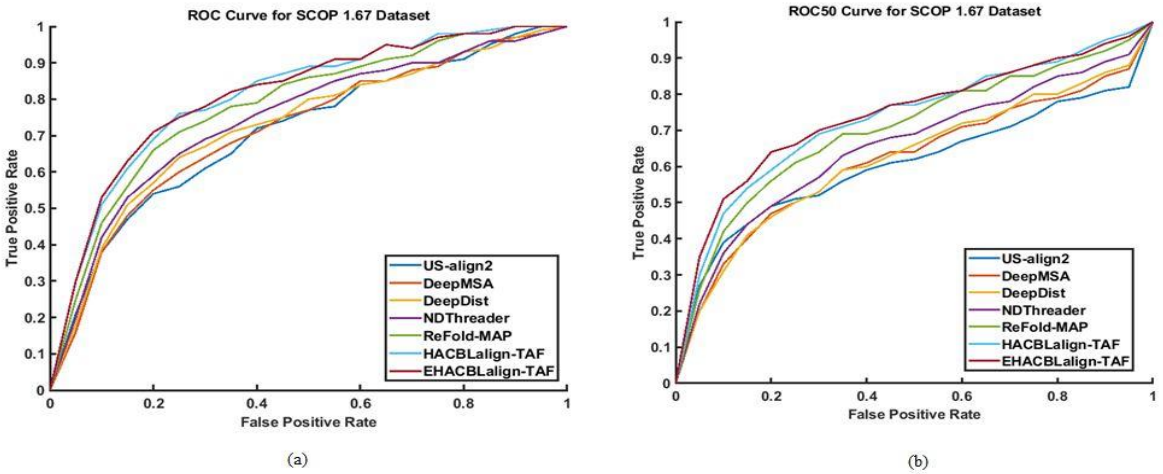

Fig. 10. Comparison of ROC ROC 50 curves for Proposed and Existing Protein Remote Homology Detection and Fold Recognition models / Methods over SCOP 1.67 Dataset

Fig. 10 (a) and (b) present the ROC and ROC50 values of various Protein Remote Homology Detection and Fold Recognition methods tested on SCOP 1.67 datasets. The results indicate that EHACBLalign-TAF outperforms US-align2, DeepMSA, DeepDist, NDThreader, ReFold-MAP, and HACBLalign-TAF in both ROC and ROC50 metrics. EHACBLalign-TAF achieves higher ROC and ROC50 values compared to these methods, demonstrating its superior effectiveness in remote homology detection and fold recognition.

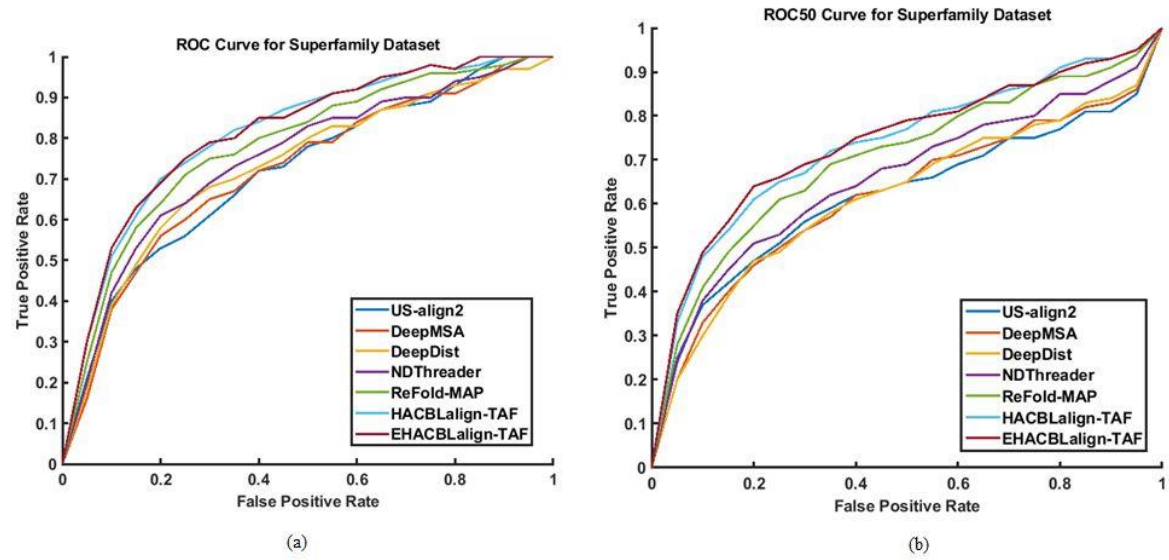

Fig. 11. Comparison of ROC and ROC Curves for Proposed and Existing Protein Remote Homology Detection and Fold Recognition Models / Methods over Superfamily Dataset

Fig. 11 (a) and (b) demonstrate the ROC and ROC50 values of various Protein Remote Homology Detection and Fold Recognition methods tested on the superfamily corpus. The results show that EHACBLalign-TAF achieves higher ROC and ROC50 values compared to US-align2, DeepMSA, DeepDist, NDThreader, ReFold-MAP, and HACBLalign-TAF. Specifically, EHACBLalign-TAF surpasses these methods in ROC and ROC50 metrics, highlighting its enhanced performance in remote homology

detection and fold recognition. These improvements are attributed to advancements in multiple sequence alignment (MSA) and the decision-making system for Protein Remote Homology Detection and Fold Recognition using large-scale protein sequences.

Conclusion

In this work, the EHACBLalign method was designed to reduce the complexity of the decision-making system by improving Multiple Sequence Alignment and Protein Remote Homology Detection and Fold Recognition. Finally, the experimental results illustrate that the EHACBLalign outperforms the benchmark methods in terms of the evaluation metrics taken into consideration viz, accuracy, precision, recall and F-measure. However, it is a challenge that EHACBLalign has high complexity because CNN may process more uninformative features. Therefore, future work is to develop advanced CNN models that process more uninformative features to enhance Protein Remote Homology Detection and Fold Recognition.

Conflict of Interest

There is no conflict of interest.

Acknowledgement

The authors declare that no specific acknowledgements are necessary for this work.

References

[1]. Lv, Z., Ao, C., and Zou, Q., 2019, Protein function prediction: from traditional classifier to deep learning, *Proteomics*, 19(14), 1-5.

[2]. Jing, X., Dong, Q., Hong, D., and Lu, R., 2019, Amino acid encoding methods for protein sequences: A comprehensive review and assessment, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 17(6), 1918-1931.

[3]. Rajapaksa, S, Sumanaweera, D, Lesk, A, M, Allison, L, Stuckey, P, J, Garcia de la Banda, M, and Konagurthu, A, S., 2022, On the reliability and the limits of inference of amino acid sequence alignments, *Bioinformatics*, 38(Supplement_1) , i255-i263.

[4]. Peyravi, F, Latif, A, and Moshtaghioun, S. M., 2019, Protein tertiary structure prediction using hidden Markov model based on lattice, *Journal of Bioinformatics and Computational Biology*, 17(02), 1-18.

[5]. Wilburn, G. W., and Eddy, S. R., 2020. Remote homology search with hidden Potts model, *PLOS Computational Biology*, 16(11), 1-22.

[6]. Chen, J., Guo, M., Wang, X., and Liu, B., 2018, A comprehensive review and comparison of different computational methods for protein remote homology detection, *Briefings in Bioinformatics*, 19(2), 231-244.

[7]. Li, C, C., and Liu, B., 2020, MotifCNN-fold: protein fold recognition based on fold-specific features extracted by motif-based convolutional neural networks, *Briefings in Bioinformatics*, 21(6), 2133-2141.

[8]. Wu, Z., Liao, Q., and Liu, B., 2020, A comprehensive review and evaluation of computational methods for identifying protein complexes from protein–protein interaction networks, *Briefings in Bioinformatics*, 21(5), 1531-1548.

[9]. Liu, B., Chen, J., Guo, M., and Wang, X., 2017, Protein remote homology detection and fold recognition based on sequence-order frequency matrix, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 16(1), 292-300.

[10]. Guo, Y., Yan, K., Wu, H., and Liu, B., 2020, ReFold-MAP: Protein remote homology detection and fold recognition based on features extracted from profiles, *Analytical Biochemistry*, 611, 1-8.

[11]. Gopinath, K., and Rajendran, G., 2023, HACBLalign: A Hierarchical Attention-based deep learning for protein remote homology and fold identification, *Journal of Theoretical and Applied Information Technology*, 14(101), 5578 – 5588.

[12]. Makigaki, S, and Ishida, T., 2020, Sequence alignment using machine learning for accurate template-based protein structure prediction, *Bioinformatics*, 36(1), 104-111.

[13]. Zhang, C., Zheng, W., Mortuza, S. M., Li, Y., and Zhang, Y., 2020, Deep MSA: Constructing deep multiple sequence alignment to improve contact prediction and fold-recognition for distant-homology proteins, *Bioinformatics*, 36(7), 2105-2112.

[14]. Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., and Hassabis, D., 2020, Improved protein structure prediction using potentials from deep learning, *Nature*, 577(7792), 706-710.

[15]. Wu, F., and Xu, J., 2021, Deep template-based protein structure prediction, *PLoS Computational Biology*, 17(5), 1-18.

[16]. Wu, T., Guo, Z., Hou, J., and Cheng, J., 2021, DeepDist: Real-value inter-residue distance prediction with deep residual convolutional network, *BMC Bioinformatics*, 22(1), 1-17.

[17]. Hakala, K., Kaewphan, S., Bjorne, J., Mehryary, F., Moen, H., Tolvanen, M., and Ginter, F., 2022, Neural network and random forest models in protein function prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 19(3), 1772-1781.

[18]. Liu, J., Wu, T., Guo, Z., Hou, J., and Cheng, J., 2022, Improving protein tertiary structure prediction by deep learning and distance prediction in CASP14. *Proteins: Structure, Function, and Bioinformatics*, 90(1), 58-72.

[19]. Zhang, C., and Pyle, A, M., 2022, A unified approach to sequential and non-sequential structure alignment of proteins, RNAs and DNAs, *Iscience*, 25(10), 1-13.

[20]. Rangwala, H, and Karypis, G., 2005, Profile-based direct kernels for remote homology detection and fold recognition, *Bioinformatics*, 21(23), 4239-4247.

[21]. Håndstad, T., Hestnes, A. J., and Sætrom, P., 2007, Motif kernel generated by genetic programming improves remote homology and fold detection, *BMC Bioinformatics*, 8(1), 1-16.

[22]. Andreeva, A., Kulesha, E., Gough, J., and Murzin, A, G., 2020, The SCOP database in 2020: expanded classification of representative family and superfamily domains of known protein structures. *Nucleic Acids Research*, 48(D1), D376-D382.

[23]. Devlin, J., Chang, M. W., Lee, K., and Toutanova, K., 2018, Bert: Pre-training of deep bidirectional transformers for language understanding, *arXiv preprint arXiv:1810.04805*.