HACBLALIGN: A HIERARCHICAL ATTENTION-BASED DEEP LEARNING FRAMEWORK FOR PROTEIN REMOTE HOMOLOGY DETECTION AND FOLD IDENTIFICATION

K. GOPINATH[1], G. RAJENDRAN[2]

[1]Research Scholar, Periyar University, Salem -11 and Assistant Professor of Computer Applications,

Sona College of Arts and Science, Salem-5, Tamilnadu, India

[2]Assistant Professor and Head, Department of Computer Science, Govt. Arts and Science College,

Modakkurichi, Erode, Tamilnadu, India

E-mail: [1]vengatgopinath@gmail.com

ABSTRACT

Protein **R**emote **H**omology Detection and Fold **I**dentification (PRHI) are the two most crucial steps in predicting protein structure. Even though many computational techniques like Multiple Sequence Alignments (MSAs) have been designed, those techniques were not able to create proper alignments due to the varying dimensions of a protein sequence. So, this paper presents a new progressive deep MSA technique to create a more suitable decision-making system for MSA of low similarity protein families. In this technique, a decision-making system is initially trained by the **H**ierarchical **A**ttention-based Convolutional Neural Network (**C**NN) with Bidirectional Long Short-Term Memory (**BL**STM) named HACBLalign to progressively align the given protein sequences by determining various posterior probability matrices. This model progressively builds a global alignment by aggregating essential subsequences alignment into sequence alignment. The attention level allows the model to choose qualitatively informative subsequences and sequences. As a result, high-quality MSA is obtained. Then, the top-N-gram and Auto-Cross-Covariance (ACC) features are extracted based on the Position-Specific Scoring Matrix (PSSM) from aligned protein sequences. Further, such features are fed into the CNN with a Softmax classifier to recognize protein homologies and folds. At last, the experimental results illustrate that the HACBLalign accomplishes a 92.4%, 92.5% and 92.1% accuracy on SCOP 1.53, SCOP 1.67 and superfamily databases respectively in recognizing protein homologies and folds compared to the conventional MSA techniques.

Keywords: *Protein Homology, Multiple Sequence Alignment, CNN, PSSM, Softmax*

1. INTRODUCTION

The categorization of proteins into structural and functional groups based on their amino acid sequences, particularly with low sequence identities, is known as protein remote homology detection in bioinformatics. For both fundamental studies and clinical practice, protein remote homology detection is a crucial step that may be used to anticipate the 3D structure and function of proteins [1, 2]. Since protein structures are more conserved than protein sequences, distant homology proteins have comparable structures and activities but lack readily observable sequence similarities. The alignment score often enters a twilight zone when the amino acid level protein sequence similarity is less than 35% [3, 4]. As a result, computational methods that solely rely on the properties of the protein sequence frequently fail to discover protein-distant homology. Also, protein sequence analysis has difficulty in detecting remote homologies and folds of proteins due to the low similarity of protein sequences. The classical predictors are majorly split into sequence-based, ranking-based, and discriminative-based techniques.

Protein similarity may be measured using sequence-based alignment techniques. But, these techniques are unable to produce appropriate alignment calculations when the sequence similarity is smaller [3]. Besides, certain profile-based alignment techniques combine the evolutionary data for proteins using MSAs. In contrast to PSI-BLAST (Position-Specific Iterative Basic Local Alignment Search Tool), Hidden Markov Model (HMM)-based techniques like HMMER, Sequence Alignment and Modeling (SAM), HHblits (HMM-HMM-based lightning-fast iterative sequence search), etc., may increase the efficiency of similarity prediction. High-speed protein sequence similarity prediction using a probabilistic approach is possible with HMMER [5].

The homologous affinity between the sequences is measured by ranking algorithms using similarity scores [6]. In comparison to features based on sequence, features based on profiles can more accurately express protein properties [7]. The popular ranking techniques are ProtEmbed [8], RankProp [9], and so on. To measure the similarity scores, motifs are crucial to the protein structures and processes [10]. The protein fragments known as motifs have specific spatial and functional conformations. In protein structures, structural motifs are patterns. Classical motif-based feature mining methods are MotifCNN and MotifDCNN [11].

The supervised learning framework used by the discriminative techniques may be converted into a binary classification based on the label data of the protein families. Protein primary sequences are used in several discriminative approaches to extract features. Profile-based features, such as profile kernel, All Fixed-width subsequences (AF)-PSSM, and Smith-Waterman (SW)-PSSM, increase detection sensitivity since they include evolutionary information. Amongst, SW-PSSM based on profiles contains two kernel functions made up of profile-profile similarity and scores for sequence-sequence similarity. Nowadays, discriminative techniques deliver cutting-edge results compared to the other techniques [12-15].

Discriminative approaches, as opposed to paired algorithms and generative methods, may quickly incorporate different protein sequence properties and learn the information from both positive and negative samples in a given benchmark dataset. The requirement for feature vectors with constant lengths as input is a critical characteristic of discriminative techniques. From these perspectives, a novel discriminative technique named ReFold-MAP [16] has been developed, which obtains comprehensive characteristics according to the three distinct profile-based characteristics: motif-PSSM [12], ACC-PSSM [17] and PDT-profile [18] for MSAs. Those characteristics termed MAP characteristics include the structural motif kernel data, the evolutionary data, and the sequence data. Then, this characteristic vector was learned by the Support Vector Machine (SVM) classifier to recognize the protein remote homologies and folds. On the other hand, the classical MSA-based models cannot be sufficiently obtained accurate alignments because the quality of the sequence alignment is varied. So, completely automated methods are still needed to get precise protein sequence alignments.

In this article, a novel progressive deep MSA technique is proposed to create a more proper decision-making model for MSA of low-similarity protein families and handle a huge database. In this technique, a decision-making model is trained using the Hierarchical Attention-based CNN-BLSTM called HACBLalign to progressively align the given protein sequences by determining various posterior probability matrices. This model progressively builds a global alignment by aggregating essential subsequences alignment into sequence alignment. Attention level supports the model for choosing qualitatively informative subsequences and sequences. The high quality of MSA is obtained due to the greater coverage and alignment depth resulting from the combination of a diverse source of sequence databases. After completing this new MSA process, the top-N-gram and ACC features are extracted based on the PSSM from aligned protein sequences. Moreover, those features are learned by the CNN with Softmax classifier to recognize protein homologies and folds. So, it can get better accuracy on the alignment of protein families, particularly on low similarity families.

The remaining sections of this article are written as follows: Section II provides an overview of current studies related to the PRHI. The HACBLalign model is described in Section 3, and its effectiveness is analyzed in Section 4. The conclusion of this work and potential improvements are presented in Section 5.

2. LITERATURE SURVEY

An improved artificial neural network was proposed by Sudha et al. (2018) [19] for identifying protein folds and forecasting the structural label. Conversely, as the number of neurons increased, its complexity also increased.

In addition, its accuracy was not effective as it requires more characteristics. By incorporating three contour-based variables into the training model, Liu and Li (2018) [20] developed a novel technique named ProtDet-CCH, which integrates CNN-BLSTM-PSSM and a ranking scheme HHblits for PRHI. However, its computational efficiency was not high.

Mensi et al. (2010) [21] investigated the issue of PRHI, which examines the functional similarity of proteins and modeled it as a binary Multiple-Instance Learning (MIL) dilemma to discern similar and non-similar proteins. This MIL method depends on the dissimilarity interpretation that involves the mixture of N-gram interpretations. However, it needs more features to further increase recognition performance. Adhikari et al. (2020) [22] presented the protein contact prediction with the help of dilated CNNs with a dropout called the DEEPCON scheme. This scheme learns two distinct types of characteristics, such as covariance characteristics from the MSAs and sequence-based characteristics, to predict the protein contacts. But, it needs other characteristics and large-scale databases to improve the predictive performance.

Fukuda and Tomii (2020) [23] designed a new model using a Deep Neural Network (DNN) with Evolutionary Coupling Analysis (ECA) called DeepECA to predict protein contact according to the data obtained from either deep or shallow MSAs. In this model, the noisy sequences were removed by using the weightage of a particular character in the sequences in MSA. But, it analyzes the accuracy of every domain, not a complete protein sequence. An end-to-end DNN called the CopulaNet model [24] was developed to predict residue co-evolution from the MSA. The major units of this model are i) an encoder to perform context-related mutation for all residues and ii) an aggregator to define the residue co-evolution. This residue co-evolution was considered to train the 2D residual network and predict the inter-residue distances for any residue pairs. But, the efficiency was degraded while the receptive field dimension was high.

Gao et al. (2021) [25] designed a CONVERT method to recognize the protein homology by discovering the multiple-to-single correlation between proteins and agent proteins using the seq2seq model. Additionally, scoring was performed to discover the sorted list and align the protein sequence. On the other hand, the runtime was high if the number of proteins was high. Rashed et al. (2021) [26] developed FPGA and a modified CNN to accelerate DNA pairwise sequence alignment. It was based on the creation of a truth table of a look-up table of each possible mixture of the DNA strings after transforming the DNA string from alphabets to binary interpretations. But, the CNN performance was degraded due to the more labels and needed to adjust their hyper-parameters.

Jin et al. (2021) [27] developed a Supervised-Manner-based Iterative BLAST (SMI-BLAST) depending on PSI-BLAST for PRHI. But, its complexity was high while increasing the number of protein sequences. Routray and Vipsita (2021) [28] developed the Principal Component Analysis (PCA) and multi-objective optimization tools for PRHI. First, various physicochemical properties from the AAIndex corpus were obtained and considered to obtain a group of representative characteristics by the PCA. Then, NSGA-II and NSGA-III optimization algorithms were utilized to search the non-zero Eigen space and retrieve differentiable eigenvectors. But, its complexity was high for a large-scale protein sequence.

From the literature, it is seen that the earlier studies focused on developing machine learning and deep learning techniques for PRHI. Even though these techniques provide good outcomes for PRHI, the research gap exists in PRHI since those techniques are most applicable to the limited number of protein sequences. Also, they need more significant features for achieving better prediction accuracy. In contrast with those techniques, the HACBLalign is a novel technique, which is suitable for aligning a large number of protein sequences and choosing the most significant features to predict protein homologies with a less computational burden.

3. PROPOSED METHODOLOGY

3.1. Objectives of this work

3.1.1 General objective

The objective of this work is to propose a new progressive deep Multiple Sequence Alignment (MSA) technique, HACBLalign, for Protein Remote Homology Detection and Fold Identification (PRHI).

3. 1.2 Specific objectives

The specific objective of this work is to develop a highly accurate protein homology and fold identification technique using a Hierarchical Attention-based Convolutional Neural Network (CNN) with Bidirectional Long Short-Term Memory (BLSTM). The goal is to achieve a minimum accuracy of 90% in protein homology and fold identification.

3.2. Steps in HACBLALIGN Techniques

The HACBLalign technique is described in this section. An outline of the presented PRHI system is given in Figure 1.

3.2.1 Protein Sequence Acquisition

In this work, three benchmark databases, including the SCOP v1.53, SCOP v1.67, and the superfamily database, are considered to analyze the efficiency of various MSA techniques. The SCOP v1.53 database has 4532 sequences from 54 families, whereas the SCOP v1.67 has 11037 sequences from 102 families. The superfamily database has 1195 folds of 1962 superfamilies. A database of structural and functional labels for every protein sequence is called Superfamily. It is constructed based on a set of HMMs that reflect structural protein domains at the level of the SCOP superfamily. By comparing protein sequences from over 2478 completely sequenced genomes to HMMs, the labels are generated.

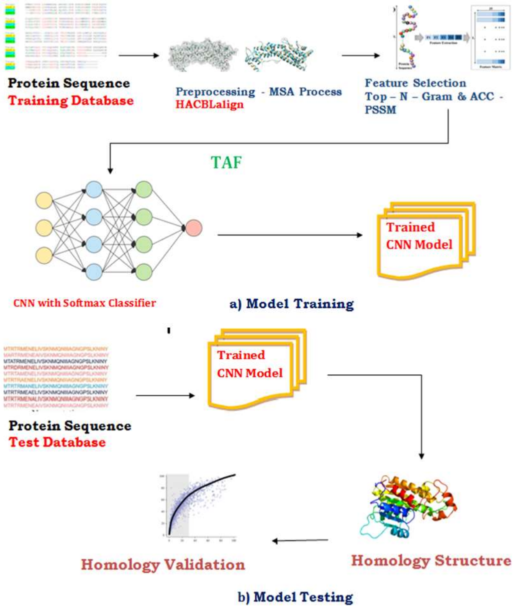

Figure 1: Framework of PRHI System

3.2.2 Generation of MSA to create a Decision-Making System

In this HACBLalign technique, a new automated alignment system is developed for a large-scale protein sequence. For a protein sequence$Q$, this HACBL model (as demonstrated in Figure 2) constructs many MSAs precisely. This HACBL comprises of five major levels: word embedding, convolution, BLSTM, attention and Fully Connected (FC) levels.

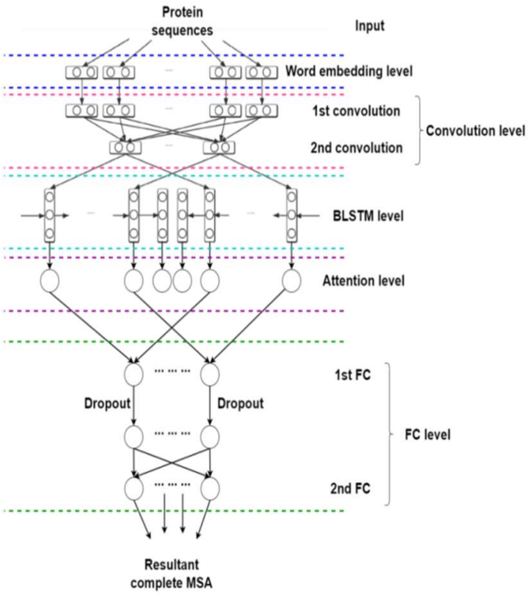

Figure 2: Structure of HACBL-based Decision-Making System

(i) Architecture of HACBL Model

Primarily, a given Q is transformed into the 512×8 matrix at the word embedding level. The obtained matrix is fed to the two convolution levels having kernel sizes of 6 and 3, respectively (both convolution levels are followed by the max-pooling of size 2).

Then, the outcome of the second convolution level is given to the BLSTM level having a hidden size of 64 for determining the correlation among all characters in a protein sequence pair. After that, the attention level is applied to capture more relevant alignments of protein sequences, which are passed to the two FC levels, whereas a dropout rate of 0.5 is added to the primary FC level. Finally, the outcome of the second FC level represents the accurate alignments for a given protein sequence.

(ii) HACBLalign: Decision-making system-based novel progressive alignment technique

The decision-making system, i.e., HA+CNN+BLSTM model, is incorporated with the standard MSA based on the progressive mechanism to create a novel alignment technique for MSAs called HACBLalign. For a protein family F , HACBLalign generates the MSA according to the below processes:

- Decision-making for determining the Posterior Probability Matrix (PPM)

All pairs a, b from F is added to this technique to provide a tag label$\left(tag_{a,b}\right)$. Those tags define the certain determination technique of PPM applied for F based on the dominant fraction of tags, which every of its pair acquires after being fed into the decision-making system. As the fraction of all accurate tags of categorizing Fvia MSAs, the dominant fraction of estimated tags is computed as:

$$\text{dominate fraction} = \underset{x}{\operatorname{argmax}}\left(\frac{\mathrm{FTE_x}}{\mathrm{RTF_x}}\right) \qquad (1)$$

In Eq. (1), FTE_xstands for the fraction of x^{th} estimated tag and RTF_xstands for the fraction of x^{th}real tag where x indicates the positive integer not greater than $n$tags. This task is portrayed in Fig 3, which defines the partitioning F into pairs by the decision-making system to predict the tag for all pairs and determine the dominant fraction of tags for finding the tag of F.

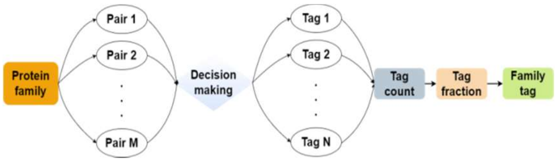

Figure 3: Task of Separating F into Pairs by Decision-Making System

According to the resulting tag of F, the Pair HMM (P-HMM), the division operation, the Root Mean Square (RMS) of P-HMM, and the division operation or the RMS of P-HMM, the division operation and arbitrary HMM are performed to determine the PPM.

- Computation of distance matrix and formation of guidance tree

A pairwise alignment is calculated on each pair a, b in F by obtaining the highest weight route via the PPM and the highest sum is denoted by $\mathcal{P}(a, b)$. The distance between a and b is calculated as:

$$\mathrm{Distance[a][b]} = 1 - \frac{\mathcal{P}(\mathrm{a,b})}{\min(\mathrm{L_a,L_b})} \qquad (2)$$

In Eq. (2), L_a and L_bIndicate a and b's length, correspondingly.

A guidance tree calculates the association (a) sequence and sequence, (b) sequence and profile and (c) profile and profile. Characterizing two sets $I$and J, the distance (D)Between their union and the other set K is described as:

$$\mathrm{D[I \cup J] \cup [K]} = \frac{\mathrm{|I| \times D[I][J] + |J| \times D[I][K]}}{\mathrm{|I| + |J|}} \qquad (3)$$

In Eq. (3), $|I|$, $|J|$ and $|K|$Represent the weights of sets I, J and K . Depending on this distance matrix, the process is started from the sequences of the least distance and a binary tree called the guidance tree is regularly constructed.

- Alteration of uniformity

In this stage, another sequence is considered for relaxing the PPM of all pairs a and $b\left(P_{a,b}\right)$ to calculate the substitution scores. This reduction task is defined as:

$$P'_{a,b} = \frac{1}{|S|}\left(2 \times P_{a,b} + \sum_{z \in S} P_{a,z} \times P_{z,b}\right) \qquad (4)$$

In Eq. (4), S is the sequences collection in F and $P'_{a,b}$ denotes the new converted PPM of pair$\langle a, b\rangle$.

- Alignment improvement

Two child nodes, or sequences, are merged from the real node to generate a profile and these are fused to the root node of the tree to obtain a complete MSA incorporating all sequences based on the guidance tree and the relaxed PPM.

- Standardization

It aims to minimize any potential errors in the placement of earlier sequences. Every aligned sequence is separated into two sets at random intervals using the iterative fine-tuning procedure and then they are realigned using a profile-profile arrangement. Only when the maximum total is bigger than it was previously are any fine-tunings valid.

3.2.3 Feature Selection Process

Once the MSAs are obtained, two feature selection techniques are applied: top-N-gram and ACC-PSSM.

- Top-N-Gram features:

A novel feature selection technique was developed by Liu et al. (2014) [29], which determines the frequency distribution of 20 general amino acids in the considered protein sequence, sorts them in descending order, and chooses N amino acids occurring most frequently based on the mixture of frequency values. The mixture of N amino acid characters is top-N-gram and characteristics of the protein sequence are acquired based on the frequency of occurrence of all top-N-grams. Also, this fundamental component of the protein includes evolutionary data. The feature mining technique is described by considering the distance d between top-N-grams as:

$$Distance_{d=0}(S') = \left\{T_{i_1}^{0}(S'), T_{i_2}^{0}(S'), \dots, T_{i_{20}}^{0}(S')\right\} \quad (5)$$

$$Distance_{1\le d\le d_{max}}(S') = \left\{T_{i_1 i_1}^{0}(S'), T_{i_2 i_2}^{0}(S'), \dots, T_{i_{20} i_{20}}^{0}(S')\right\} \quad (6)$$

In Equations (5) and (6), $i_1, i_2, \dots, i_{20}$ are 20 amino acids, their frequencies of occurrence are in descending order, S' refers to the sequence having top-1-grams and $T_{i_1}^{0}$ denotes the frequency of occurrence of top-1-gram having i_1 with distance is 0 in $T_{i_1}^{0}$. To avoid additional dimensionalities from impacting efficiency, N is assigned to 1. But, feature mining techniques are similar if the distance is higher than 1. So, the highest d must be assigned to the value higher than 2 to precisely find the impact of the feature mining technique. Accordingly, d is assigned to 3 so that $20 + 20 \times 20 \times 3$ dimensional characteristics are determined.

- ACC-PSSM features:

The ACC-PSSM depends on the PSSM and defines the correlation between 2 amino acids. ACC features have Auto-Covariance (ACov) and Cross-Covariance (CCov), which are determined as:

$$ACov(i,d) = \sum_{j=1}^{L-d} \frac{(s_{i,j} - \bar{s_i})(s_{i,j+d} - \bar{s_i})}{L-d} \quad (7)$$

$$CCov(i_1, i_2, d) = \sum_{j=1}^{L-d} \frac{(s_{i_1,j} - \overline{s_{i_1}})(s_{i_2,j+d} - \overline{s_{i_2}})}{L-d} \quad (8)$$

In Eqns. (7) and (8), i denotes the residue ($i \in [1,20]$), d is the distance between 2 distinct residues, $s_{i,j}$ is the score of i at location$j$ in the PSSM, $\bar{s_i}$ is the mean scoreand L is length of the protein sequence. The sizes of ACov-PSSM and CCov-PSSM are $20 \times \alpha$ and $380 \times \alpha$, correspondingly, whereas α is assigned to 7 for determining the impact of the significance between 2 residues. After obtaining both feature vectors, those are linearly merged depending on PSSMs to get a complete feature set called TAF.

3.2.4 CNN-based Protein Homology Recognition

In the final stage, CNN with softmax function is performed, which involves training and test processes. During the training process, the feature vectors and tags from the training sequences to train the CNN with softmax model. During the test process, the test sequences are transformed into the TAF vectors following a similar procedure as the training sequences and then they are recognized by the training model. The structure of CNN with the softmax function is depicted in Figure 4.

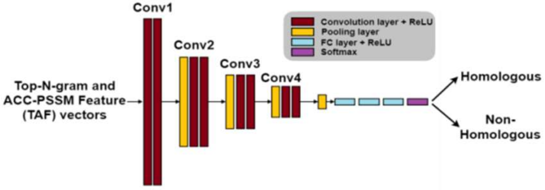

Figure 4: Structure of CNN with Softmax for PRHI

4. EXPERIMENTAL RESULTS

In this section, the effectiveness of this HACBLalign-TAF technique is analyzed by implementing it in MATLAB 2019b using three benchmark databases discussed in Section 3.1. From these databases, 70% of the sequences are considered for the training process and 30% of the sequences are considered for the testing process. Also, the observed efficiencies are evaluated with the existing techniques viz., ReFold-MAP [16], motif-PSSM [12], ACC-PSSM [17], PDT-profile [18], ProtDet-CCH [20], DeepECA [23], CopulaNet [24], SMI-BLAST [27] and PCA-NSGA-III [28] in terms of precision, recall, accuracy, Receiver Operating Characteristics (ROC) and ROC50.

- Accuracy: It defines the fraction of properly recognized protein homologies to the sum number of protein sequences tested.

$$Accuracy = \frac{\text{True Positive (TP)} + \text{True Negative (TN)}}{\text{TP} + \text{TN} + \text{False Positive (FP)} + \text{False Negative (FN)}} \quad (9)$$

- Precision: It defines the fraction of aligned positions, which are accurately aligned.

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (10)$$

- Recall: It defines the fraction of aligned residues that are accurately aligned.

$$\text{Recall} = \frac{\text{TP}}{\text{TP+TN}} \quad (11)$$

- F-Measure: It defines the f-measure of proposed and existing PRHI techniques

$$\text{F measure} = \frac{\text{2 X precision X recall}}{\text{precision+r}} \quad (12)$$

- ROC and ROC50 curve: The ROC value determines the balance between specificity and sensitivity. It plots TPs against FPs in the normalized Area Under the Curve (AUC). Similarly, ROC50 value is the Area under the ROC curve up to the 50 false positives. The ROC curve is drawn by measuring TP rate and FP rate as:

$$TP\ rate = \frac{TP}{TP+FN} \quad (13)$$

$$FP\ rate = \frac{FP}{FP+TN} \quad (14)$$

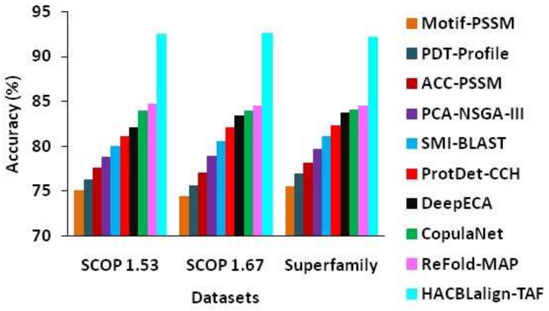

Figure 5: Comparison of Accuracy for Proposed and Existing Frameworks

Figure 5displays the accuracy of proposed and existing PRHI techniques executed on three distinct benchmark databases. For SCOP 1.53 database, the accuracy of HACBLalign-TAF is 23.2% higher than the Motif-PSSM, 21.3% higher than the PDT-Profile, 19.2% higher than the ACC-PSSM, 17.4% higher than the PCA-NSGA-III, 15.5% higher than the SMI-BLAST, 13.9% higher than the ProtDet-CCH, 12.7% higher than the DeepECA, 10.1% higher than the CopulaNet and 9.1% higher than the ReFold-MAP techniques. For SCOP 1.67 database, the accuracy of HACBLalign-TAF is 24.3% higher than the Motif-PSSM, 22.4% higher than the PDT-Profile, 20.1% higher than the ACC-PSSM, 17.2% higher than the PCA-NSGA-III, 14.9% higher than the SMI-BLAST, 12.8% higher than the ProtDet-CCH, 10.9% higher than the DeepECA, 10.3% higher than the CopulaNet and 9.6% higher than the ReFold-MAP techniques.

Additionally, the accuracy of HACBLalign-TAF for the superfamily database is 22% higher than the Motif-PSSM, 19.8% higher than the PDT-Profile, 17.9% higher than the ACC-PSSM, 15.7% higher than the PCA-NSGA-III, 13.7% higher than the SMI-BLAST, 11.9% higher than the ProtDet-CCH, 10% higher than the DeepECA, 9.6% higher than the CopulaNet and 9.1% higher than the ReFold-MAP techniques.

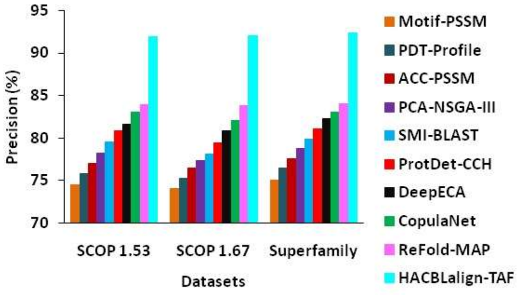

Figure 6: Comparison of Precision for Proposed and Existing Frameworks

Figure 6 portrays the precision of proposed and existing PRHI techniques applied to the 3 distinct benchmark databases. For SCOP 1.53 database, the precision of HACBLalign-TAF is 23.4% greater than the Motif-PSSM, 21.2% greater than the PDT-Profile, 19.4% greater than the ACC-PSSM, 17.5% greater than the PCA-NSGA-III, 15.6% greater than the SMI-BLAST, 13.7% greater than the ProtDet-CCH, 12.6% greater than the DeepECA, 10.7% greater than the CopulaNet and 9.5% greater than the ReFold-MAP techniques. For SCOP 1.67 database, the precision of HACBLalign-TAF is 24.3% greater than the Motif-PSSM, 22.3% greater than the PDT-Profile, 20.3% greater than the ACC-PSSM, 19% greater than the PCA-NSGA-III, 17.8% greater than the SMI-BLAST, 15.9% greater than the ProtDet-CCH, 13.9% greater than the DeepECA, 12.2% greater than the CopulaNet and 9.8% greater than the ReFold-MAP techniques.

Additionally, the precision of HACBLalign-TAF for the superfamily database is 23.1% greater than the Motif-PSSM, 20.8% greater than the PDT-Profile, 18.9% greater than the ACC-PSSM, 17.3% greater than the PCA-NSGA-III, 15.5% greater than the SMI-BLAST, 14% greater than the ProtDet-CCH, 12.2% greater than the DeepECA, 11.2% greater than the CopulaNet and 9.9% greater than the ReFold-MAP techniques.

Figure 7 depicts the recall of proposed and existing PRHI techniques applied to the three distinct benchmark databases. For SCOP 1.53 database, the recall of HACBLalign-TAF is 22.2% greater than the Motif-PSSM, 20.5% greater than the PDT-Profile, 18% greater than the ACC-PSSM, 15.5% greater than the PCA-NSGA-III, 14.2% greater than the SMI-BLAST, 12.9% greater than the ProtDet-CCH, 11.3% greater than the DeepECA, 10.6% greater than the CopulaNet and 9.4% greater than the ReFold-MAP techniques.

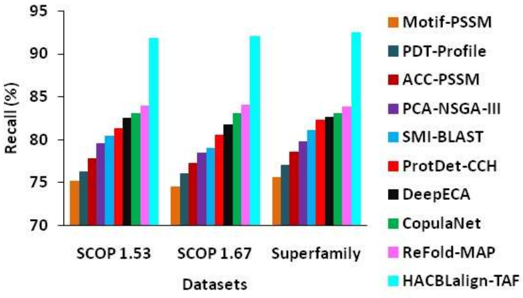

Figure 7: Comparison of Recall for Proposed and Existing Frameworks

For SCOP 1.67 database, the recall of HACBLalign-TAF is 23.5% greater than the Motif-PSSM, 21.1% greater than the PDT-Profile, 19.2% greater than the ACC-PSSM, 17.3% greater than the PCA-NSGA-III, 16.5% greater than the SMI-BLAST, 14.3% greater than the ProtDet-CCH, 12.6% greater than the DeepECA, 10.8% greater than the CopulaNet and 9.5% greater than the ReFold-MAP techniques. Additionally, the recall of HACBLalign-TAF for the superfamily database is 22.2% greater than the Motif-PSSM, 20% greater than the PDT-Profile, 17.7% greater than the ACC-PSSM, 15.9% greater than the PCA-NSGA-III, 14.1% greater than the SMI-BLAST, 12.3% greater than the ProtDet-CCH, 11.9% greater than the DeepECA, 11.3% greater than the CopulaNet and 10.3% greater than the ReFold-MAP techniques.

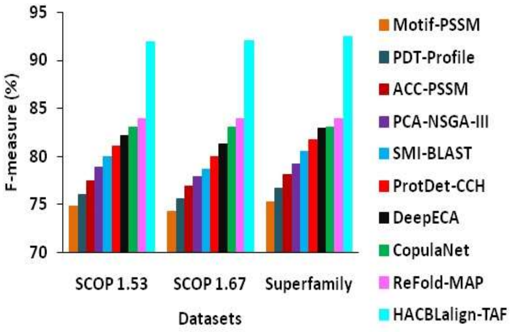

Figure 8: Comparison of F-measure for Proposed and Existing Frameworks

Figure 8illustrates the f-measure of proposed and existing PRHI techniques applied to the three distinct benchmark databases. For SCOP 1.53 database, the f-measure of HACBLalign-TAF is 22.9% larger than the Motif-PSSM, 20.9% larger than the PDT-Profile, 18.7% larger than the ACC-PSSM, 16.5% larger than the PCA-NSGA-III, 14.9% larger than the SMI-BLAST, 13.3% larger than the ProtDet-CCH, 11.9% larger than the DeepECA, 10.7% larger than the CopulaNet and 9.5% larger than the ReFold-MAP techniques. For SCOP 1.67 database, the f-measure of HACBLalign-TAF is 23.8% larger than the Motif-PSSM, 21.7% larger than the PDT-Profile, 19.6% larger than the ACC-PSSM, 18.1% larger than the PCA-NSGA-III, 17% larger than the SMI-BLAST, 15% larger than the ProtDet-CCH, 13.2% larger than the DeepECA, 10.8% larger than the CopulaNet and 9.7% larger than the ReFold-MAP techniques. Additionally, the f-measure of HACBLalign-TAF for superfamily database is 22.7% larger than the Motif-PSSM, 20.5% larger than the PDT-Profile, 18.3% larger than the ACC-PSSM, 16.7% larger than the PCA-NSGA-III, 14.8% larger than the SMI-BLAST, 13.1% larger than the ProtDet-CCH, 11.5% larger than the DeepECA, 11.3% larger than the CopulaNet and 10.1% larger than the ReFold-MAP techniques.

Figure 9 portrays the ROC and ROC50 values obtained for the proposed and existing PRHI techniques applied to the SCOP 1.53 database.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

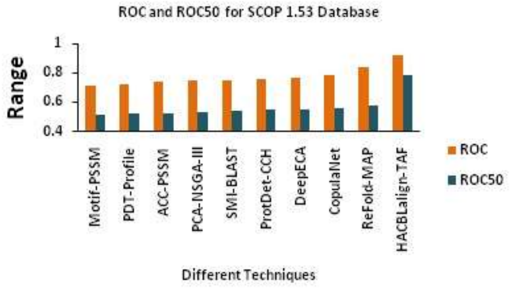

Figure 9: Comparison of ROC & ROC50 for Proposed and Existing Techniques on SCOP 1.53 Database

It observes that the ROC of HACBLalign-TAF is 29% larger than the Motif-PSSM, 26.7% larger than the PDT-Profile, 24.8% larger than the ACC-PSSM, 23.5% larger than the PCA-NSGA-III, 22.6% larger than the SMI-BLAST, 21.1% larger than the ProtDet-CCH, 19.8% larger than the DeepECA, 17.2% larger than the CopulaNet and 9.7% larger than the ReFold-MAP techniques. Similarly, the ROC50 of HACBLalign-TAF is 54.7% larger than the Motif-PSSM, 51.8% larger than the PDT-Profile, 49.9% larger than the ACC-PSSM, 48.2% larger than the PCA-NSGA-III, 46.4% larger than the SMI-BLAST, 44.5% larger than the ProtDet-CCH, 43.2% larger than the DeepECA, 41% larger than the CopulaNet and 36.8% larger than the ReFold-MAP techniques.

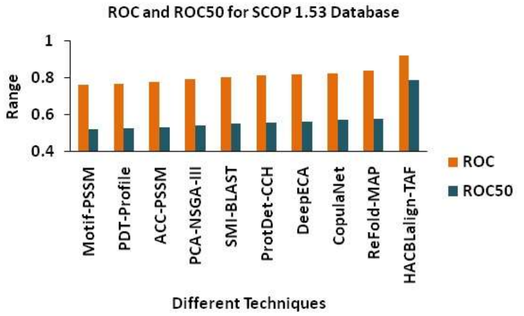

Figure 10: Comparison of ROC and ROC50 for Proposed and Existing Techniques on SCOP 1.67 Database

Figure 10 depicts the ROC and ROC50 values obtained for the proposed and existing PRHI techniques applied to the SCOP 1.67 database. It observes that the ROC of HACBLalign-TAF is 21.1% larger than the Motif-PSSM, 19.7% larger than the PDT-Profile, 18% larger than the ACC-PSSM, 16.2% larger than the PCA-NSGA-III, 14.3% larger than the SMI-BLAST, 13.4% larger than the ProtDet-CCH, 12.5% larger than the DeepECA, 12% larger than the CopulaNet and 9.7% larger than the ReFold-MAP techniques. Similarly, the ROC50 of HACBLalign-TAF is 52.6% larger than the Motif-PSSM, 42.4% larger than the PDT-Profile, 48.7% larger than the ACC-PSSM, 46.1% larger than the PCA-NSGA-III, 43.7% larger than the SMI-BLAST, 42.4% larger than the ProtDet-CCH, 40.7% larger than the DeepECA, 38.5% larger than the CopulaNet and 36.6% larger than the ReFold-MAP techniques.

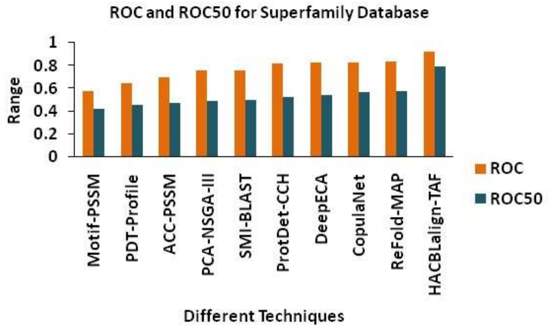

Figure 11: Comparison of ROC and ROC50 for Proposed and Existing Techniques on Superfamily Database

Figure 11 demonstrates the ROC and ROC50 values obtained for the proposed and existing PRHI techniques applied to the superfamily database. It observes that the ROC of HACBLalign-TAF is 59.8% larger than the Motif-PSSM, 42.6% larger than the PDT-Profile, 30.9% larger than the ACC-PSSM, 21.4% larger than the PCA-NSGA-III, 20.7% larger than the SMI-BLAST, 12.4% larger than the ProtDet-CCH, 11.3% larger than the DeepECA, 10.7% larger than the CopulaNet and 9.5% larger than the ReFold-MAP techniques. Similarly, the ROC50 of HACBLalign-TAF is 85.2% larger than the Motif-PSSM, 73.5% larger than the PDT-Profile, 67.7% larger than the ACC-PSSM, 59.9% larger than the PCA-NSGA-III, 57.5% larger than the SMI-BLAST, 51.1% larger than the ProtDet-CCH, 44.3% larger than the DeepECA, 38.6% larger than the CopulaNet and 36.5% larger than the ReFold-MAP techniques.

In summary, the accuracy is utilized to analyze how many correct predictions are achieved by the new HACBLalign-TAF technique on three distinct datasets. Guo et al. [16] achieved 84.7%, 84.4%, and 84.4% accuracy in their ReFold-MAP using SCOP 1.53, SCOP1.67, and superfamily datasets respectively; the difference in accuracy value from the proposed technique is 9.1%, 9.6%, and 9.1% respectively. Jin et al. [27] achieved 80%, 80.5%, and 81% accuracy in their SMI-BLAST technique using SCOP 1.53, SCOP1.67, and superfamily datasets respectively; the difference in accuracy value from the proposed technique is 15.5%, 14.9%, and 13.7% respectively. In addition to that, the proposed technique is compared with the existing techniques in terms of precision, recall, f-measure, and ROC, as shown in Figure 6 – Figure 11. The precision, recall, and f-measure values of the proposed technique on SCOP 1.53 are 91.9%, 91.8%, and 91.9% which are greater than the existing techniques such as ReFold-MAP, SMI-BLAST, etc. Similarly, the precision, recall, and f-measure values of the proposed technique on SCOP 1.67 and superfamily datasets are higher than the existing techniques listed above.

5. CONCLUSION AND FUTURE WORK

Open research issues in the field of protein predictions are protein - protein interaction, functional characterization of uncharacterized proteins, post transitional modification and Protein remote homology detection and fold identification. In this work, a novel technique named HACBLalign was developed for progressively aligning a huge number of protein sequences according to their distinct PPMs.

The HACBLalign played a vital role in aligning protein sequences and selecting the most relevant features (i.e., TAF), so the CNN with a softmax classifier achieved the best accuracy and sensitivity values when compared to the other techniques. By doing effective alignment on the protein sequence datasets, the computational complexity has been reduced, and the technique has not faced any complexity in aligning the protein sequences, particularly on low similarity families.

The objective of this work has been met and it can be concluded that the HACBLalign technique achieved good performance compared to the other existing techniques.

In future work, the proposed technique can be integrated with the advanced pattern search algorithms to increase the efficiency of the decision-making model while using a vast amount of sequences.

REFERENCES

[1] C.W. Ko, J. Huh, and J.W. Park, “Deep Learning Program to Predict Protein Functions Based on Sequence Information”, *MethodsX*, vol. 9, 2022, pp. 1-11.

[2] A. Mehta, and H. Mazumdar, “Recent Trends in Machine Learning-Based Protein Fold Recognition Methods”, *Biointerface Research in Applied Chemistry*, vol.11, no. 4, 2021, pp. 11233-11243.

[3] B.Y. Khor, G.J. Tye, and T.S. Lim, “General Overview on Structure Prediction of Twilight-Zone Proteins”, *Theoretical Biology and Medical Modelling*, vol. 12, no. 1, 2015, pp. 1-11.

[4] S. Xie, P. Li, Y. Jiang, and Y. Zhao, “A Discriminative Method for Protein Remote Homology Detection Based on N-Gram”, *Genetics and Molecular Research*, vol. 14, no. 1, 2015, pp. 69-78.

[5] M. Remmert, A. Biegert, A. Hauser, and J. Soding, “HHblits: Lightning-Fast Iterative Protein Sequence Searching by HMM-HMM Alignment”, *Nature Methods*, vol. 9, no. 2, 2012, pp. 173-175.

[6] B. Liu, Y. Zhu, and K. Yan, “Fold-LTR-TCP: Protein Folds Recognition Based on Triadic Closure Principle", *Briefings in Bioinformatics*, vol. 21, no. 6, 2020, pp. 2185-2193.

[7] B. Liu, J. Chen, M. Guo, and X. Wang, “Protein Remote Homology Detection and Fold Recognition Based on Sequence-Order Frequency Matrix”, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 16, no. 1, 2017, pp. 292-300.

[8] I. Melvin, J. Weston, and W.S. Noble, “Detecting Remote Evolutionary Relationships among Proteins by Large-Scale Semantic Embedding”, *PLoS Computational Biology*, Vol. 7, No. 1, 2011, pp. 1-9.

[9] I. Melvin, J. Weston, C. Leslie, and W.S. Noble, “RANKPROP: a Web Server for Protein Remote Homology Detection”, *Bioinformatics*, vol. 25, no. 1, 2009, pp.121-122.

[10] B. Liu, C.C. Li, and K. Yan, “DeepSVM-fold: Protein Fold Recognition by Combining Support Vector Machines and Pairwise Sequence Similarity Scores Generated by Deep Learning Networks”, *Briefings in Bioinformatics*, vol. 21, no. 5, 2020, pp.1733-1741.

[11] C.C. Li, and B. Liu, “MotifCNN-fold: Proteins Fold Recognition Based on Fold-Specific Features Extracted by Motif-Based Convolutional Neural Networks”, *Briefings in Bioinformatics*, vol. 21, no. 6, 2020, pp. 2133-2141.

[12] X. Gao, D. Wang, J. Zhang, Q. Liao, and B. Liu, “IRBP-motif-PSSM: Identification of RNA-Binding Proteins Based on Collaborative Learning”, *IEEE Access*, vol. 7, 2019, pp. 168956-168962.

[13] M.S. Refahi, A. Mir, and J.A. Nasiri, "A Novel Fusion Based on the Evolutionary Features for Protein Fold Recognition Using Support Vector Machines", *Scientific Reports*, vol. 10, no. 1, 2020, pp. 1-13.

[14] Q. Zou, G. Liu, X. Jiang, X. Liu, and X. Zeng, "Sequence Clustering in Bioinformatics: An Empirical Study", *Briefings in Bioinformatics*, vol. 21, no. 1, 2020, pp. 1-10.

[15] M.R. Bouadjenek, K. Verspoor, and J. Zobel, "Automated Detection of Records in Biological Sequence Databases that are Inconsistent with the Literature", *Journal of Biomedical Informatics*, vol. 71, 2017, pp. 229-240.

[16] Y. Guo, K. Yan, H. Wu, and B. Liu, "ReFold-MAP: Protein Remote Homology Detection and Fold Recognition Based on Features Extracted from Profiles", *Analytical Biochemistry*, vol. 611, 2020, pp. 1-8.

[17] Q. Dong, S. Zhou, and J. Guan, "A New Taxonomy-Based Protein Fold Recognition Approach Based on Autocross-Covariance Transformation", *Bioinformatics*, vol. 25, no. 20, 2009, pp. 2655-2662.

[18] B. Liu, X. Wang, Q. Chen, Q. Dong, and X. Lan, "Using Amino Acid Physicochemical Distance Transformation for Fast Protein Remote Homology Detection", *PLoS One*, vol. 7, 2012, pp. 1-10.

[19] P. Sudha, D. Ramyachitra, and P. Manikadan, "Enhanced Artificial Neural Network for Protein Fold Recognition and Structural Class Prediction", *Gene Reports*, vol. 12, 2018, pp. 261-275.

[20] B. Liu, and S. Li, "ProtDet-CCH: Protein Remote Homology Detection by Combining Long Short-Term Memory and Ranking Methods", *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 16, no. 4, 2019, pp. 1203-1210.

[21] M. Mensi, A. Toto, A. Donati and G. Mauri, "Protein sequence analysis and prediction methods", *Current Bioinformatics,* vol. 5, no. 1, 2010, pp. 21-34.

[22] B. Adhikari, "DEEPCON: Protein Contact Prediction Using Dilated Convolutional Neural Networks with Dropout", *Bioinformatics*, vol. 36, no. 2, 2020, pp. 470-477.

[23] H. Fukuda, and K. Tomii, "DeepECA: An End-To-End Learning Framework for Protein Contact Prediction from a Multiple Sequence Alignment", *BMC Bioinformatics*, vol. 21, no. 1, 2020, pp. 1-15.

[24] F. Ju, J. Zhu, B. Shao, L. Kong, T. Liu, W. Zheng, and D. Bu, "CopulaNet: Learning Residue Co-Evolution Directly from Multiple Sequence Alignment for Protein Structure Prediction", *Nature Communications*, vol. 12, no. 1, 2021, pp. 1-9.

[25] S. Gao, S. Yu, and S. Yao, "An Efficient Protein Homology Detection Approach Based on Seq2seq Model and Ranking", *Biotechnology and Biotechnological Equipment*, vol. 35, no.1, 2021, pp. 633-640.

[26] A.E.E.D. Rashed, M. Obaya, and H.El.D. Moustafa, "Accelerating DNA Pairwise Sequence Alignment Using FPGA and a Customized Convolutional Neural Network", *Computers and Electrical Engineering*, vol. 92, 2021, pp. 1-21.

[27] X. Jin, Q. Liao, H. Wei, J. Zhang, and B. Liu, "SMI-BLAST: A Novel Supervised Search Framework Based on PSI-BLAST for Protein Remote Homology Detection", *Bioinformatics*, vol. 37, no. 7, 2021, pp. 913-920.

[28] M. Routray, and S. Vipsita, "Protein Remote Homology Detection Combining PCA and Multi objective Optimization Tools", *Evolutionary Intelligence*, Vol. 2021, pp.1-10.

[29] B. Liu, J. Xu, Q. Zou, R. Xu, X. Wang, and Q. Chen, "Using Distances Between Top-N-Gram and Residue Pairs for Protein Remote Homology Detection", *BMC Bioinformatics*, vol. 15, no. 2, 2014, pp. 1-10.