ISSN No:-2456-2165

Protein Remote Homology Detection-Methods and Evaluation Metrics

Gopinath K
Sona College of Arts and Science,
Salem , Tamilnadu, India.

Rajendran G
Thiruvalluvar Govt. Arts College, Rasipuram,
Namakkal, Tamilnadu, India

Abstract:- Protein Remote Homology Detection (PRHD) is a concept that aims to discover remote evolutionary links between proteins. PRHD research is currently vital for assessing protein structures and function. A variety of computational approaches have been developed in recent decades to overcome this challenge which requires constant-width characteristics to specify the Protein Sequences (PSs). However, with only a rudimentary knowledge of proteins, identifying their discrimination characteristics is not an easy task. Therefore, a brief comparative review and comparison of different computation methods is essential for PRHD. In this paper, a review of various PRHD methods with the help of different computational methods is presented. In addition, their benefits and drawbacks are discussed in a tabular form. Lastly, the whole survey is summarized and future directions are suggested to improve the efficiency of protein classification based on amino acid sequences, especially with low sequence identity between proteins.

Keywords:- *Protein Remote Homology Detection (PRHD), Protein Networks, Fold Recognition, Machine Learning, Deep Learning.*

I. INTRODUCTION

PRHD is important in bioinformatics because homologous proteins almost have similar patterns, so it helps for studying the 3D structure and function of proteins [1]. Unfortunately, due to low protein sequence similarity in datasets, predictors' performance is degraded for PRHD [2]. To deal with this challenge, numerous sophisticated computational solutions have been presented in recent decades. PRHD is the primary difficulty in establishing the unique Protein Sequence (PS) composition with the fewest similarities and identified proteins [3]. Some computing methods have been created, which may be classified into three broad varieties: discriminative feature approaches, and scoring strategies [4]. Discriminative approaches address PRHD as a categorization fault, with proteins represented as constant-width attribute vectors that are fed into classifiers to train the system. Finally, these systems can determine the homology link between unlabelled compounds. Scoring methods treat PRHD as a sorting activity or database recovery effort. PS alignment techniques support to extract more meaningful features from PSs. The Multiple Sequence Alignments (MSAs) methods rely on similarities between a pair of PSs revealed by the dynamic learning framework.

In contrast, these methods are unable to generate meaningful matches when the genome similarity is less than 35%. Some alignment strategies have been invented to improve the adaptability of MSAs, the other alignment approaches like Structure-based MSAs PSI-BLAST, IMPALA, COMA and COMPASS strategies were developed. Alignment strategies based on Hidden Markov Models (HMMs) transform an MSA into a location-specific that does not provide a precise maximum-ranking sequence but rather a collection of possible combinations. Some discriminative algorithms using data characteristics obtained from protein primary behaviours, such as motifs analysis, have been created to employ the positive and negative PS forms [5-6].

There are few independent pipelines that can efficiently build sensitive MSAs from a query input sequence in all of MSA generation, particularly when numerous massive sequence repositories are involved. Though the overall accuracy of these MSA tools is very good, when it comes to a given protein family, the efficiency of the various tools is typically inconsistent, resulting in poor accuracy findings. For PRHD, most all of the techniques depended on standard machine leaning techniques. As inputs, all machine learning algorithms require fixed length vectors. Despite this, protein sequence lengths vary greatly. The sequence-order knowledge and rank correlation effects are missing during the vectorization step, this is crucial for PS and nucleic acid analysis. Even though several methods have tried to include these details into predictors, in contrast, it is never a simple process due to the limited understanding of proteins.

Bioinformatics has extensively applied Deep Learning (DL) techniques to boost the discriminating power over other Machine Learning (ML) methods in recent times. Several successful DL approaches like Convolution Neural Network (CNN) and Recurrent Neural Network (RNN) utilizes chronological data of input elements. The transient relationships among the fundamental blocks like Long Short-Term Memory (LSTM) is used which identifies autonomously long-term and short-term relationships in PS based on a relevant data collected from the previous subsequent processes.

This article presents the detailed survey on different remote protein homology detection methods for providing efficient predictive results and reduces the problem of detecting homology in cases of low sequence similarity. Also, a comparative study is presented to address the advantages and disadvantages of those frameworks to suggest future scope. The rest of the sections are prepared as follows: Section II discusses different tree-based deep

ISSN No:-2456-2165

learning frameworks for various purposes. Section III provides the comparative analysis of those frameworks. Section IV summarizes the entire survey and recommends the upcoming scope.

II. LITERATURE SURVEY

A sequence-based technique dubbed Distance Pair Pseudo Amino Acid Composition (disPseAAC) was described for the efficient PRHD [7]. Using Chou's PseAAC methodology based on distance-pairs occurrences within a particular distance and different physicochemical property scores in the AA Index database, feature vectors were created by integrating the distribution of AA pairs inside the Chou's PseAAC approach. This technique would improve the predicting performance of the disPseAAC predictor by including sequence-order information and physicochemical features of proteins. However, this approach has a slow convergence rate.

A new approach called remote-3Dp system was introduced [8] for PRHD. The remote-3Dp approach relied on both anticipated 3D information and AA physiochemical characteristics. Initially, the remote-3DP approach was developed using 3D projected information and AA physiochemical characteristics. A NN was used to predict the 3D information (i.e., contact map) from the AA fragments obtained, and another NN was used to predict the beta-sheet connections. In SCOP families, this remote-3DP technique uses solely structural models to describe a protein and effectively distinguishes between remote and non-remote homologues. However, the computing cost of this method was intensive.

Using Markov Random Fields (MRF) and the Stochastic Search technique, the method called MRFy was created [9] for PRHD for Beta-Structural Proteins. The pair-wise relationships among amino acid residues combined together in a β-sheet were captured using the SMURF and SMURFLite MRF models in this technique. The initial guess was created initially in this MRFy approach, which comprises three separate criteria. Second, simulated annealing, the genetic algorithm, and the local search option were used to process the stochastic search. Finally, the output result of the best alignment list was assigned. However, this strategy is prone to mistakes and produces skewed predicted findings.

An ensemble classifier for PRHD called Support Vector Machine-Ensemble (SVM-E) was developed [10] with a weighted voting approach. This SVM-E integrates three fundamental classifiers on distinct feature spaces, such as Kmer, Auto-Cross Covariance (ACC), and Series Correlation Pseudo AA Composition (SC-PseAAC). These features incorporate the proteins properties from a variety of angles, including sequence composition, sequence-order information and PSs. To construct the discriminative weight vectors in the feature space, this approach used the PCA feature extraction method. Due to the enormous number of training sessions, this classifier may underperform at times.

A strategy was proposed using numerous cascading events with HMM (C-HMM) [11] to locate distant homologues from the PSs database. These C-HMM were divided into three categories: (i) C-HMM, which includes exhaustive sequence searches over many generations; and (ii) In C-HMM modules, Cd-hit was applied to aggregate all actual strikes gathered after every iteration, and only indicative patterns from the clusters were utilised to start the following generation. (iii) C-HMM, which was equivalent to Module-2 except that instead of using only typical sequences, the whole hits set within the cluster was evaluated to create an HMM profile that would be used as the origin for the following iteration. At the family and superfamily levels, as well as inside folds, this technique works better. However, this method came at a high computational expense.

Using a profile-based pseudo PSs (pPSs) and rank aggregation approach, a novel predictor named dRHP-PseRA - new predictor was created [12] to detect PRHD. The evolutionary information from the relevant profiles of pseudo proteins was extracted using this technique. At first, this protein representation method appeared to be capable of converting profile evolutionary information into Palindromes in Protein Sequences (pPSs). The pseudo proteins were then placed into predictors to determine PRHD. The pseudo proteins were then placed into predictors to determine PHR. Finally, these predictors used a linear weighted rank aggregation technique to construct the ranking lists. The performance of this was examined by combining four standard estimators: Position-Specific Iterative Basic Local Alignment Search Tool (PSI-BLAST), HHblits, Hmmer, and Coma. In constrast, this strategy was ineffective on bigger datasets.

A progressed technique referred to as PROTDEC-LTR2.0 was offered [13] for PRHD with the help of integrating pseudo protein and supervised ranking system. This system offer graphical interface which facilitates to discover possible sequences and possible structure of proteins. Initially, the query and all protein sequence information had been converted into PSs by using the PSI-BLAST, HH-BLITS. The sequences were represented as a characteristic matrix for each query. Later, this matrix was fed into the Learning to Rank (LTR) version to re-rank. In contrast, this technique has high computational time.

A computational estimator referred to as PRHD (ProDec) was suggested [14] based on Bidirectional Long Short-Term Memory (BI-LSTM) and Temporally Distributed Dense Layer (TDDL) to extract discriminative capabilities from pPSs. This approach includes input layer, Bi-LSTM layer, TDDL and output layer. This architecture captures both the long and short dependency facts of pPSs aggregating the results from every intermediate hidden value of Bi-LSTM. The TDDL assigns weights to several levels of structured links and fuses the facts with this network. However, the computational value of this technique was excessive.

A SVM (Support Vector Machine – hybrid) was presented [15] by combining SVM-ACC (Accuracy) and SVM- Physicochemical Distance Transformation (PDT) methods for PRHD. This SVM-ACC approach was employed to achieve appropriate phylogenetic proximity transformation for each PSs Position-Specific Scoring Matrix (PSSM) profiles. For each pair of Amino Acid, the SVM-PDT approach was used to derive the physicochemical proximity transformation from the Amino Acid - Index database. Finally, the discriminant weight of each feature from the hybrid model was calculated using the SVM technique. However, this method does not provide efficient results on larger datasets.

A distinctive, sequence-free method was developed [16] for PRHD with quasi-linear complexity processed for theoretical PHR searches. In this method, the PSs were converted into quantitative physio-chemical forms by using unique vector quantization approach which employs Discrete Cosine Transform reduction to transform PSs into a constant-width expression. Then, Dynamic Time Warping technique was used to generate the global similarity scores between proteins for each compressed representation which were subsequently fed into a Random Forest. However, the computation complexity of this method was high.

The Novel method named ProtDet-CCH was presented [17] by combining CNN – BLSTM – PSSM and a ranking method (HHblits,) for efficient PRHD. Initially, the characteristics of protein super-families were captured using a CNN layer, resulting in increased discriminative capability. To capture patterns fragments across proteins with minimal sequence similarity, the forward and backward LSTM algorithms use extra local and global sequence order information. After that, by taking PSSMs as protein representations, evolutionary data from MSAs may be simply included into the predictor. Finally, this method takes the benefits of CNN-BLSTM-PSSM and HHblits for the efficient PRHD. Although the accuracy results high, certain proteins cannot be correctly predicted.

A unique estimator Called ProtDec-LTR3.0 was presented [18] to include the attribute-profile into the Learning to Sorting techniques for the efficient PRHD. In this method, three attribute profiles like Top-1-gram, Top-2-gram, and ACC were utilized to distinguish the remote homology proteins for the MSA. This technique updates the value of each protein node in this network by dividing the sorting list scores of search input proteins into powerful and weaker PHR based on the input proteins scores. Finally, a web host and user manual were created so that the query PSs could be sent in FASTA format and their PHR could be discovered automatically. However, variation in the results was acquired throughout the performance.

A CNN-based network called ConvRes [19] was designed, which combines a variation Inception and Resnet block for the PRHD. This method captures a physiological property of Amino Acids to represent the PSs. This consecutive data was sent into a modified Inception block, which used different kernel sizes to extract conceptual characteristics from PSs. The properties of PSs can be increased after the Inception block because various kernel sizes might be viewed as different window widths depending on PSs. Then, using the preceding characteristics as input, the Resnet block was used as a predictor. Finally, this model would determine as if the input sequence is a member of a certain family. This method requires large amount to train, so overftting problem might be resulted.

A kernel approximation based method [20] was created to explore enormous large PSs for PRHD. The Nyström approximation was used to transmit the fundamental protein commonalities in a low-rank graph without defining each pair-wise resemblance. Employing flexible concurrent systems in dispersed memory using Apache-Hadoop/Spark, this method could explore protein set of connections for more proteins. Furthermore, the time required to form a protein network was somewhat longer than that required by the HMM. But, it has a high computational cost.

A new discriminative method called ReFold-MAP was presented [21] for efficient PRHD. This approach was utilised to extract complete aspects from three attribute profiles like Motif-PSSM, ACC-PSSM, and PDT-profile. To retrieve characteristics like protein pattern information and evolutionary information, the Motif-PSSM was applied. The link between any two amino acids is represented as ACC-PSSM. PDT-profile was used to investigate the impact of sequence information on prediction accuracy and to express protein structure. In constrast, the performance of this method has to be increased due to large computational error.

A novel approach known as CONVERT was introduced [22] to improve the efficiency of PRHD. The seq2seq model and scoring were used in this approach, which treats homology detection as a translation work. Moreover, this method introduces the concept of representative protein which contains characteristics of the protein family. The representative protein for each family was initially created using semi-global matching. The seq2seq model was then used to build a many-to-one link between proteins and representative proteins. Finally, an encoder portion was kept to build protein eigen vectors. Finally, the eigen vectors were used to generate a sorted list for the query sequence. However, this method requires large datasets for training.

A supervised and iterative BLAST was created [23] using PS (Position Specific Iterative) –BLAST for PRHD. This approach was successfully used to eliminate incorrectly detected homology based on error from PSSMs for boosting the performance of PSI BLAST for PRHD. These search non-homologous protein failures. Using the Sequence similarity matrix, homology scores were corrected and sorting quality search outcomes effectively. However, sequencing only takes place on single protein samples; this method lacks high-throughput capabilities.

ISSN No:-2456-2165

An integration of PCA and Non-dominated Sorting Genetic Algorithms (NSGA-II and NSGA-III) was devised [24] for PRHD. PCA starts by reducing the number of features in the original attribute set that were retrieved from an AA Index database. The SVM classifier was then used to classify the lower attribute set. Finally, to reduce the amount of patterns and classification errors over generations, optimization techniques such as NSGA II and III were employed to effectively explore the non-zero eigen space and return identifiable eigen vectors. In this strategy, a pareto optimal front was found with the smallest amount of features and the maximum average classification accuracy. However, this method was acquired with slow convergence rate.

The Sequence-Order Frequency Matrix Sampling and ML with Smith-Waterman (SOFM-SMSW) technique were developed [25] for PRHD. The SOFM-SMSW method used the Proportional Volume Sampling (PVS) approach to pick the best target sequences using the uniform distribution parmaeter. MSAs were utilised to generate a SOFM matrix, which was then used to predict a uniform distribution of each protein's SOFM. The intended sequence was generated from it. Then, classification was used to find the concatenation site of two PSs, which was processed using K-Nearest Neighbor (K-NN) to predict substitution ranks. The sequence matching was improved using the SMSW method, which was performed through MSA to get a refined matching rank. Finally, SVM was used to process the matching ranks for accurate prediction results. But, slow performance was resulted due to large computation steps.

III. COMPARATIVE ANALYSIS

In this section, a comparative study of different remote protein homology detection methods which are briefly studied in above section is presented in Table 1. The performance analysis in terms of Receiving Operating Characteristics curve (ROC) and / or Area Under the Curve (AUC) is also given in table1 along with the merits and Demerits of each method.

Algorithm	Merits	Demerits	Performance analysis
PCA [7]	Less computational cost and effective predictive performance	Slow convergence rate	Resulted Mean ROC = 0.92 and ROC50 scores = 0.721
Neural Network [8]	Low dimensionality of protein representation was observed	High computational cost	Roc Score = 0.963 Accuracy = 0.92
MRF(Markov Random Fields) and Stochastic Search algorithm [9]	Less computational time	Highly susceptible to errors and has biased predictive performance	Mean ROC curve = 0.95
SVM- ensemble classifier [10]	Due to the ensemble classifier, this method provides efficient predictive accuracy.	Lack of Performance Due To Larger Training Datsasets	Resulted Mean ROC = 0.94 and ROC50 scores = 0.744
Hidden Markov Models [11]	This approach effectively shortens the search time while preserving sequence information.	High time complexity	On applying diverse protein folds, C-HMM coverage for family = 94%; Super-family= 83% and fold levels = 40%
Rank Aggregation approach [12]	This method has less computational time	Not effective on larger datasets.	Resulted Mean ROC = 0.83 and ROC50 scores = 0.89
Supervised learning to rank (LTR) algorithm [13]	This method has better convergence rate.	High computational time	Resulted Mean ROC = 0.891 and ROC50 scores = 0.895
Bi-LSTM and Time distributed dense layer [14]	This method has higher discriminative power to employ more handcrafted protein features.	High computational cost	Resulted Mean ROC = 0.970 and ROC50 scores = 0.714
SVM-ACC and SVM-PDT [15]	This SVM-hybrid method relatively memory efficient	Less performance on larger datasets	Resulted Mean ROC = 0.95 and Matthews correlation coefficient (MCC) = 0.89
Discrete Cosine Transform compression and Dynamic Time Warping technique [16]	This approach improves detection speed significantly while less accuracy reduction.	High computational complexity	For the Gene Dataset (GD) Resulted Mean AUC = 0.94 and AUC1000 scores = 0.80 For PFAM dataset AUC = 0.96 and AUC1000 scores = 0.87 For SUPFAM dataset AUC = 0.93 and AUC1000

			scores = 0.81
CNN – BLSTM - PSSM and a HHblits[17]	This method has low computational complexity.	The resulted accuracy was high, but some proteins were not correctly predicted.	Resulted Mean ROC = 0.99 and ROC50 scores = 0.98
Page Rank algorithm and HITS algorithm [18]	This method wisely computes the rank score and takes less detection time	varied performance was resulted all through the performance	Resulted Mean ROC = 0.9117±0.0061 and ROC50 scores = 0.9121±0.0064
CNN-based network: ConvRes variant Inception and Resnet [19]	Less Computational time	Overfitting problem	Area under the receiver operating characteristic (AUROC) = 0.97
low-rank kernel approximation [20]	For exploring as vast protein network, this technique was efficient and effective	High computational cost	For remote homology detection. Mean AUC = 0.97 and AUC1000 scores = 0.86 In leave-one-out studies, the AUC of LP-LOKA (PSI) increased from 0.8 to 0.9 for distant homology and 0.64 to 0.72 for fold identification.
Support Vector Machine [21]	This method was efficient to use in any comprehensive evaluations	Large computational error was resulted	Resulted Mean ROC = 0.97 and ROC50 scores = 0.86
Semi-Global Alignment [22]	This approach is extremely rapid, yielding results in a matter of milliseconds.	Requires large number of datasets for training	Resulted Mean AUC = 0.931 and AUC1000 scores = 0.856
Sequence similarity matrix [23]	High gen generalizability was resulted	Because sequencing only takes place on single protein samples, this technology does not have high-throughput capabilities.	Resulted Mean ROC = 0.89 and ROC50 scores = 0.93
PCA SVM, NSGA-II and NSGA-III [24]	Minimum classification error and effective performance was resulted.	Slow convergence rate	Classification accuracy = 0.92 Classification error = 0.17
Machine Learning methods like KNN and SVM; PVS approach, MSA [25]	Less computational complexity	Slow performance due to high computational steps	Resulted Mean ROC = 0.94 and ROC50 scores = 0.96

Table 1: Comparison of PRHD techniques using different computational methods

IV. RESULT AND DISCUSSION

- *Receiver Operating Characteristic Curve (ROC) and ROC50*

It is not adequate to analyze the efficiency based on the accuracy of detecting PHR owing to the unbalanced samples in the benchmark. To combat this challenge, Receiver Operating Characteristic (ROC) and ROC50 are determined, which is broadly applied to analyze the classifying imbalanced databases. It is robust, if the distribution of positive and negative samples differs with period [26]. It is calculated using the True positive rate (TPR) and false positive rate (FPR).

TPR is the percentage of positive observations that were perfectly detected to the overall positive observations, i.e. $TPR = (TP/(TP + FN))$.

FPR is the percentage of positive observations that are imperfectly detected to the overall negative observations, i.e. $FPR = (FP/(TN + FP))$.

In the scenario of PRHD, the TPR defines the percentage at which structural and functional labels are properly detected as positive when the sequence relevance is less. The ROC50 score is the area under the ROC curve, up to the primary 50 FPs. The above-studied discriminative techniques serve PRHD as a sequence of binary categorization processes, the learned model allocates a chance for all test samples and their efficiency is determined depending on each sample in the test set.

ISSN No:-2456-2165

Further, the performance of the different PRHD method (which was taken from the above table 1) disPseAAC [7], dRHP-PseRA [12], ProtDec-LTR2.0 [13], ProtDet-CCH [17], ProtDec-LTR3.0 [18], PRHD-ReFold-MAP [21] and PRHD-SOFM-SMSW [25] are analyzed and compared in terms of graphical representation for ROC and ROC50 metric values from the existing methods to provide efficient PRHD detection and its significance results.

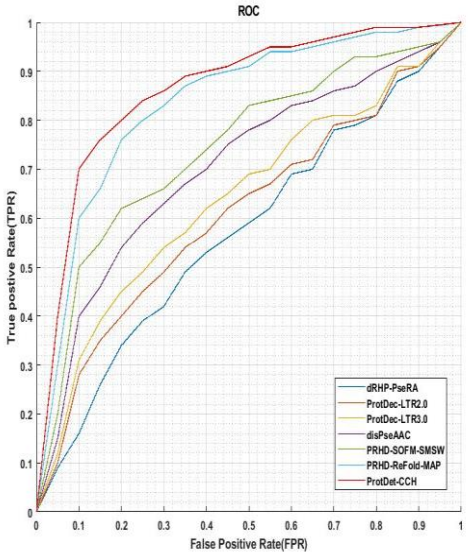

Fig. 1: Comparison on ROC values on different existing methods

In Figure 1, the ROC values for disPseAAC, dRHP-PseRA, ProtDec-LTR2.0, ProtDet-CCH, ProtDec-LTR3.0, PRHD-ReFold-MAP and PRHD-SOFM-SMSW are given. In case of ROC, ProtDet-CCH method is greater than disPseAAC, dRHP-PseRA, ProtDec-LTR2.0, ProtDet-CCH, ProtDec-LTR3.0, PRHD-ReFold-MAP and PRHD-SOFM-SMSW. From this analysis, ProtDet-CCH technique has the maximum detective performance for ROC values than the other existing techniques for PRHD function.

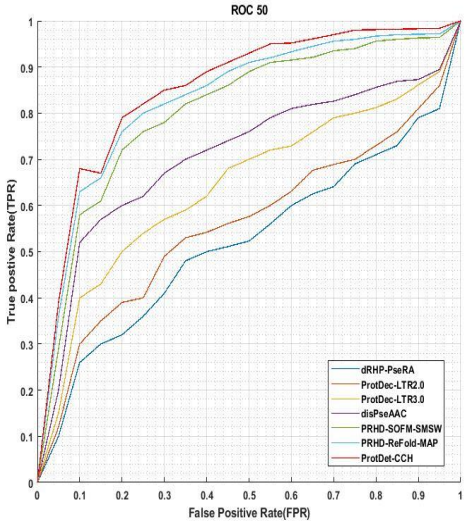

Fig. 2: Comparison onROC 50 values on different existing methods

In Figure 2, the ROC 50 values for disPseAAC, dRHP-PseRA, ProtDec-LTR2.0, ProtDet-CCH, ProtDec-LTR3.0, PRHD-ReFold-MAP and PRHD-SOFM-SMSW are given. In case of ROC 50, ProtDet-CCH method is greater than disPseAAC, dRHP-PseRA, ProtDec-LTR2.0, ProtDet-CCH, ProtDec-LTR3.0, PRHD-ReFold-MAP and PRHD-SOFM-SMSW. From this analysis, ProtDet-CCH technique has the maximum detective performance for ROC 50 values than the other existing techniques for PRHD function.

V. CONCLUSION

In this paper, a detailed comparative analysis on PRHD techniques based on different computational method have been presented. From this comparative analysis, it is obviously understood that the researchers who have practiced on remote homology analysis using different computation method finds some limitations like high computational time and cost, less detection value, low performance on large datasets etc., which lacks to provide efficient detection PRHD systems. As a result, future extensions of this work might address all of the above-mentioned concerns by concentrating on real-time applications that give effective and reliable analytical estimations while assuring PSs with considerable similarities.

REFERENCE

[1.] J. Chen, M. Guo, X. Wang and B. Liu, "A comprehensive review and comparison of different computational methods for protein remote homology detection", Briefings in Bioinformatics, vol. 19, No. 2, pp. 231-244, 2016.

[2.] B. Liu, J. Chen and X. Xialong Wang, "Application of learning to rank to protein remote homology detection", *Bioinformatics*, vol. *31, No.* 21, pp.3492-3498, 2015.

[3.] J. Echave, C. O. Wilke, "Biophysical models of protein evolution: understanding the patterns of evolutionary sequence divergence", Annual Review of Biophysics, 46, pp. 85-103, 2017.

[4.] S. Abeln, J. Heringa and K.A. Feenstra, "Introduction to protein structure prediction", arXiv:1712.00407, 2017.

[5.] Z. Yao, K. L Macquarrie, A. P. Fong, S. J. Tapscott, W. L Ruzzo and C. Robert Gentleman, "Discriminative motif analysis of high-throughput dataset", Bioinformatics, vol. 30, No. 6, pp. 775-783, 2014.

[6.] P. Stegmaier, A. Kel, E. Wingender and J. Borlak, "A discriminative approach for unsupervised clustering of DNA sequence motifs. *PLoS Computational Biology*, vol. *9, No.* 3, pp. 1-13, 2013.

[7.] B. Liu, J. Chen and X. Wang, "Protein remote homology detection by combining Chou's distance-pair pseudo amino acid composition and principal component analysis", *Molecular Genetics and Genomics*, vol. *290, No.* 5, pp. 1919-1931, 2015.

[8.] O. F. Bedoya and I. Tischer, "Remote homology detection of proteins using 3D models enriched with

physicochemical properties", *Ingeniería y competitividad*, vol. *17, No.*1, pp. 75-84, 2015.
[9.] N. M. Daniels, A. Gallant, N. Ramsey and L. J. Cowen, "MRFy: remote homology detection for beta-structural proteins using Markov random fields and stochastic search", *IEEE/ACM transactions on computational biology and bioinformatics*, vol. *12, No.*1, pp. 4-16, 2014.
[10.] J. Chen, B. Liu and D. Huang, "Protein remote homology detection based on an ensemble learning approach", *BioMed research international*, *2016*.
[11.] S. Kaushik, A. G Nair, E. Mutt, H. P. Subramanian and R. Sowdhamini, "Rapid and enhanced remote homology detection by cascading hidden Markov model searches in sequence space", *Bioinformatics*, vol. *32*, No. 3, pp. 338-344, 2016.
[12.] J. Chen, R. Long, X. L. wang, B. Liu, and K. c. Chou, "dRHP-PseRA: detecting remote homology proteins using profile-based pseudo protein sequence and rank aggregation", Scientific Reports, vol. 6, No. 1, pp. 1-7, 2016.
[13.] J. Chen, M. Guo, S. Li and B. Liu, "ProtDec-LTR2.0: an improved method for protein remote homology detection by combining pseudo protein and supervised Learning to Rank", *Bioinformatics*, vol. *33, No.* 21, pp. 3473-3476, 2017.
[14.] S. Li, J. Chen and B. Lin, "Protein remote homology detection based on bidirectional long short-term memory", *BMC bioinformatics*, vol. *18, No.* 1, pp. 1-8, 2017.
[15.] J. Xie, D. F. Lu, J. H. Shu, J. Wang, C. Meng and W. Zhang , "A hybrid support vector machine method for protein remote homology detection", In *Lecture Notes in Engineering and Computer Science: Proceedings of The International MultiConference of Engineers and Computer Scientists*.
[16.] D. Raimondi, G. Orlando, Y. Moreau and F. Vranken, "Ultra-fast global homology detection with discrete cosine transform and dynamic time warping", *Bioinformatics*, vol. *34, No.* 18, pp. 3118-3125, 2018.
[17.] 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, pp. 1203-1210, 2018.
[18.] B. Bin Liu, Y. Zhu, "ProtDec-LTR3. 0: protein remote homology detection by incorporating profile-based features into learning to rank", *IEEE Access*, *7*, pp. 102499-102507, 2019.
[19.] Y. Wang, J. Bao, F. Huang, J. Du and Y. Li, "Protein Remote Homology Detection Based on Deep Convolutional Neural Network", 2019.
[20.] R. Petegrosso, Z. Li, M. A. Srour, Y. Saad, W. Zhang and R. Kuang, "Scalable remote homology detection and fold recognition in massive protein networks", *Proteins: Structure, Function, and Bioinformatics*, vol. *87, No.* 6, 478-491, 2019.
[21.] 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*, *611*, pp. 114013, 2020.
[22.] S. Gao, S. Yu and S. Yao, "An efficient protein homology detection approach based on seq2seq model and ranking", *Biotechnology & Biotechnological Equipment*, *35*(1), pp. 633-640, 2021.
[23.] X. Jin, Q. Liao, H. Wei, J. Zhang and Bin Liu, "SMI-BLAST: a novel supervised search framework based on PSI-BLAST for protein remote homology detection", *Bioinformatics*, vol. *37, No.* 7, pp. 913-920, 2021.
[24.] M. Routray and S. Vipsita."Protein remote homology detection combining PCA and multiobjective optimization tools", *Evolutionary Intelligence*, pp. 1-10, 2021.
[25.] S. Nakshathram, R. Duraisamy and M. Pandurangam, Sequence-Order Frequency Matrix-Sampling and Machine learning with Smith-Waterman (SOFM-SMSW) for Protein Remote Homology Detection, 2021.
[26.] M. Kuhn and K. Johnson, "Applied predictive modelling", in Springer, New York, pp. 28-63, 2013.