Sequence-Order Frequency Matrix - Sampling and Machine learning with Smith-Waterman (SOFM-SMSW) for Protein Remote Homology Detection

Sajithra Nakshathram
Bharathiar University

Ramyachitra Duraisamy (jaichitra1@yahoo.co.in)
Bharathiar University https://orcid.org/0000-0002-7060-6206

Manikandan Pandurangan
Loyola College School of Computational Sciences

Research Article

Keywords: Protein remote homology detection, Proportional Volume Sampling, Sequence-Order Frequency Matrix, k-Nearest Neighbor, Smith-Waterman.

Posted Date: August 2nd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-729077/v1

Sequence-Order Frequency Matrix - Sampling and Machine learning with Smith-Waterman (SOFM-SMSW) for Protein Remote Homology Detection

N.Sajithra[1], , D. Ramyachitra[2], P.Manikandan[3]
[1]Research Scholar, [2,3]Assistant Professor,
[1,2]Department of Computer Science, Bharathiar University, Coimbatore-641 046, India.
[3]Department of Data Science, Loyola College, Chennai – 600 094
sajithramidhun@gmail.com, jaichitra1@yahoo.co.in, manimkn89@gmail.com.

ABSTRACT

***Background*:** Protein Remote Homology Detection (PRHD) is used to find the homologous proteins which are similar in function and structure but sharing low sequence identity. In general, the Sequence-Order Frequency Matrix (SOFM) was used for protein remote homology detection. In the SOFM Top-n-gram (SOFM-Top) algorithm, the probability of substrings was calculated based on the highest probability value of substrings. Moreover, SOFM-Smith Waterman (SOFM-SW) algorithm combines the SOFM with local alignment for protein remote homology detection. However, the computation complexity of SOFM based PRHD is high since it processes all protein sequences in SOFM.

***Objective*:** Sequence-Order Frequency Matrix - Sampling and Machine learning with Smith-Waterman (SOFM-SMSW) algorithm is proposed for predicting the protein remote homology. The SOFM-SMSW algorithm used the PVS method to select the optimum target sequences based on the uniform distribution measure.

***Method*:** This research work considers the most important sequences for PRHD by introducing Proportional Volume Sampling (PVS). After sampling the protein sequences, a feature vector is constructed and labeling is performed based on the concatenation between two protein sequences. Then, a substitution score which represents the structural alignment is learned using k-Nearest Neighbor (k-NN). Based on the learned substitution score and alignment score, the protein homology is detected using Smith-Waterman algorithm and Support Vector Machine (SVM). By selecting the most important sequences, the accuracy of PRHD is improved and the computational complexity for PRHD is reduced by using structural alignment along with the local alignment.

***Results*:** The performance of the proposed SOFM-SMSW algorithm is tested with SCOP database and it has been compared with various existing algorithms such as SVM Top-N-gram, SVM pairwise, GPkernal, Long Short-Term Memory (LSTM), SOFM Top-N-gram and SOFM-SW.

***Conclusion*:** The experimental results illustrate that the proposed SOFM-SMSW algorithm has better accuracy, precision, recall, ROC and ROC 50 for PRHD than the other existing algorithms.

Keywords: Protein remote homology detection, Proportional Volume Sampling, Sequence-Order Frequency Matrix, k-Nearest Neighbor, Smith-Waterman.

1. INTRODUCTION

In living organisms, the proteins are considered as an important functional unit and those are involved in various biological processes [1]. Proteins had similar functions and structures in the same family. Further details on an obscure protein can be gained based on the protein family [2-6]. Protein Remote Homology Detection (PRHD) methods are processed with the aim of finding the families of a protein. Development of new drugs for a specific disease is achieved using protein remote homology detection. Generally, PRHD methods are categorized as discriminative methods, ranking methods and sequence-based alignment methods. According to the similarities among a couple of protein sequences, protein homology detection is achieved in sequence-based alignment methods [7]. Discriminative methods [8] extract features from initial protein sequences and differentiate protein families based on the extracted features. Ranking methods calculate the proteins homology relationship by depicting all the proteins into a feature space according to the distance in the feature space. From these methods, the alignment-based methods achieve the state-of-the-art performance for PRHD.

A Sequence-Order Frequency Matrix (SOFM) [9] was used for Protein Remote Homology Detection – Fold Recognition (PRHD-FD) that combined the sequence-order effects of amino acids with the Multiple Sequence Alignment (MSA). After the construction of SOFM, Top-n-gram was performed on that matrix to transform it into fixed length vector. Then, a SOFM-Top was processed for PRHD-FD. In order to find the similarity between any two SOFMs, Smith-Waterman local alignment algorithm was used [10, 20]. The local alignment similarity was given as input Support Vector Machine (SVM) for PRHD-FD. In this research work, the Proportional Volume Sampling (PVS) method is introduced to consider only the target proteins for PRHD-FD which reduces the computation time for SOFM based PRHD. Furthermore, the error rate of the SVM is further reduced by considering protein structural alignment along with the protein local alignment for PRHD. A substitution score is predicted using kNN that is used in Smith-Waterman algorithm for refining the sequence alignment. After that, MSA is applied on the sequence alignment to obtain refined SOFM matrix and alignment score. The alignment score is trained in SVM for PRHD. The remaining sections of this research work are organized as follows: Section 2 elaborates the literature study of existing techniques in protein remote homology detection, Section 3 illuminates the methodology of proposed SOFM-

SMSW method, Section 4 emphasizes the results and discussion for the SCOP database and finally Section 5 describes the conclusion.

2. LITERATURE SURVEY

Multi-layer Support Vector Machine (SVM) classifier is used for homology detection and fold recognition. One of the layers in multi-layer SVM detects the super family and family in the Structural Classification of Proteins (SCOP) [11] hierarchy by using fine-tuned binary SVM classification rules and Bio-kernel function. Another layer of multi-layer SVM was used to detect protein fold level in SCOP hierarchy using discriminative SVM with string kernel. However, the high dimensional feature vector affects the accuracy and processing time of homology detection and fold recognition process [12]. A tool is developed to detect protein remote homology using Markov Random Fields (MRF) and stochastic search. The MRF was used to capture standard Hidden Markov Model (HMM) and pairwise association between amino acid residues bonded together in β-sheets. Nevertheless, in many real cases MRF was computationally impractical. So, stochastic search was used which provided optimal or near optimal solution for protein homology detection. However, this tool required a template which was built from a set of protein chains [13].

The feature extraction technique utilized Position Specific Scoring Matrix (PSSM) to calculate the tri-grams of protein sequence and predicts the protein fold recognition [14]. Based on the tri-grams, a matrix was constructed using PSSM which determined the fold of a protein sequence. However, this technique still needs further improvement in terms of recognition accuracy. Also, the Soft Ngrams technique is utilized for protein homology detection [15]. Ngram was a profile-based representation for protein sequences that permitted to consider whole information in the profile. Then, the representation was converted into a feature vector by using a hybrid generative-discriminative scheme. Finally, the feature vectors were processed in SVM to detect the protein homology. However, soft Ngrams is computationally expensive. HMM-HMM arrangement and dynamic programming is used for effective recognition of protein fold [16]. Initially, Profile HMM (PHMM) matrix was extracted from the protein sequence by applying HMM-HMM alignment on the protein sequence. After that, kernalized dynamic programming was explored to calculate the distance between the corresponding PHMM matrices. Based on the

distance between the two proteins, the protein fold was recognized. By including other features from physicochemical attributes, the recognition accuracy will be improved.

Protein fold recognition is achieved using the Computational Predicator [17]. In the computational predictor, the sequence features were extracted from the protein sequences and then a dictionary was constructed which holds the extracted features. The dictionary was given as input to Sparse Representation Classifier (SRC) for protein fold recognition. Advanced machine learning methods will be used to enhance the fold recognition. The characteristics of protein sequences was extracted to enhance Deep Extreme Learning Machine (DELM) based protein fold prediction [18]. Bacterial Foraging Optimization-Genetic Algorithm (BFO-GA) algorithm is using for the purpose of multiple sequence alignment of measures carrying out and improve the multi objective [19]. Deep learning technique named Protein Remote Homology Detection based on Bidirectional Long Short-Term Memory (ProDec-BLSTM) [21] and ensemble classifier named SVM-Ensemble [22] are used to detect the protein remote homologies. PATSIM [23] tool is used to analyze the protein patterns based on the Self Optimized Prediction Method (SOPM) server. The computational methods for protein remote homology detection is discussed and it can be divided into three groups such as discriminative, alignment and ranking methods [24]. CONVERT method concerns homology detection as a translation task and presents a concept of illustrative protein [25]. A discriminative method named ReFold –MAP extracts the comprehensive features based on Motif-PSSM, ACC-PSSM and PDT Profile [26]. Machine learning algorithms are used to predict the protein homology of un-annotated sequences [27]. Principal Component Analysis (PCA) was applied in the extracted features to reduce the dimensionality of extracted features. The extracted features and the original features were processed in DELM and Linear Discriminant Analysis (LDA) to recognize the protein fold [16]. However, it is limited to high dimensional data. To overcome the disadvantages in the existing research works, this research work planned to propose SOFM-SMSW algorithm for protein remote homology detection.

3. METHODOLOGY

Initially, SOFM is constructed for the protein sequences based on Multiple Sequence Alignment (MSA). Then, the PVS is applied on the SOFM of each sequence to get the most important sequences (i.e., target sequence) which reduces the computational complexity for

PRHD. Smith-Waterman technique is used to refine the sequence alignment of target sequences using local alignment and structural alignment. From the structural alignment and local alignment of target sequence, a feature vector and label are generated. Then, the feature vectors and label are trained using kNN to forecast the match of the position (i.e., substitution score). The substitution score is used in Smith-Waterman algorithm for refining the sequence alignment and the alignment score of the refined matrix is trained in SVM for efficient PRHD. In this section, the proposed SOFM-SMSW is described in detail for protein remote homology detection. The overall framework for the proposed SOFM-SMSW algorithm is shown in Fig.1.

3.1 Input Protein sequence

The Structural Classification of Proteins (SCOP) 1.53 and SCOP 1.67 benchmark datasets are used for the experimental analysis. Fig.1(a) shows the sample subset of protein sequence for the remote homology detection.

3.2 Construction of Sequence-Order Frequency Matrix based on Multiple Sequence Alignment

The construction process of SOFM consists of two steps such as i) extraction of amino acid patterns along with sequence-order information from MSA and ii) computation of alignment scores in SOFM. Consider a protein P with n amino acid residues

$$P = Q_1 \quad Q_2 \quad Q_3 \quad Q_4 \dots \quad Q_{n-1} \quad Q_n \tag{1}$$

In Eq. (1), $Q_i (i = 1,2,3,4 \dots n)$ – i^{th} amino acid residues in protein P. Its equivalent MSA is given as follows:

$$\begin{aligned}
P_1&: \quad Q_{1,1} \quad Q_{1,2} \quad Q_{1,3} \; Q_{1,4} \; \dots \quad Q_{1,n-1} \quad Q_{1,n} \\
P_2&: \quad Q_{2,1} \quad Q_{2,2} \quad Q_{2,3} \; Q_{2,4} \quad \dots \quad Q_{2,n-1} \quad Q_{2,n} \\
P_3&: \quad Q_{3,1} \quad Q_{3,2} \quad Q_{3,3} \; Q_{3,4} \; \dots \quad Q_{3,n-1} \quad Q_{3,n} \\
&\quad \vdots \\
P_4&: \quad Q_{k,1} \quad Q_{k,2} \quad Q_{k,3} \; Q_{k,4} \; \dots \quad Q_{k,n-1} \quad Q_{k,n}
\end{aligned} \tag{2}$$

In Eq. (2), $P_i(i = 1,2,3, \dots k)$ denotes the $i - th$ sequence in MSA and $Q_{i,j}(i = 1,2,3, \dots, k; j = 1,2,3, \dots n)$ denotes amino acid or a gap at j-th position in sequence P_i.

Substring $s_{i,j}$ with h amino acids at location j in sequence P_i is denoted as,

$$s_{i,j} = Q_{i,j}Q_{i,j+1} \dots Q_{i,j+h-1} \tag{3}$$

In Eq. (3), $h(h = 1,2, \dots n)$ denotes the length of substring $s_{i,j}$. Assume S_j denote the group of all substrings with h amino acid at position j.

$$S_j = \{Q_{i,j}Q_{i,j+1} \dots Q_{i,j+h-1} | \forall i \in [1,2, \dots, k\} \tag{4}$$

where,

elements in S_j are repeatable and size of S_j equals to the sum of protein sequences in MSA.

In order to create the profile, the sequence-order information is combined when these substrings in every column of the MSA are used. According to the probability of the substring $s_{i,j}$ appearing S_j, the SOFM alignment scores are calculated. SOFM can be represented in matrix format which is given as follows:

$$G = \begin{bmatrix} g_{1,1} & g_{1,2} & \cdots & g_{1,n-h+1} \\ g_{2,1} & g_{2,2} & \cdots & g_{2,n-h+1} \\ \vdots & \vdots & \vdots & \vdots \\ g_{20^h,1} & g_{20^h,2} & \cdots & g_{20^h,n-h+1} \end{bmatrix} \tag{5}$$

In Eq. (5), n - length of the protein sequence P, 20 - standard amino acids and 20^h denotes the total number of all possible substrings $\beta_i(i = 1,2, \dots 20^h)$ of length h. The $g_{i,j}(0 < g_{i,j} < 1)$ is occurring probability of substring β_i in position $j(j = 1,2, \dots n - h + 1)$ during the evolutionary process, which is given as follows:

$$g_{i,j} = \frac{\mathcal{F}_{i,j}(\beta_i, S_j)}{\sum_{i=1}^{20^h} \mathcal{F}_{i,j}(\beta_i, S_j)} \tag{6}$$

$$\mathcal{F}_{i,j}(\beta_i, S_j) = \begin{cases} f(\beta_i, S_j), & \beta_i \in S_j \\ 0, & \beta_i \notin S_j \end{cases} \tag{7}$$

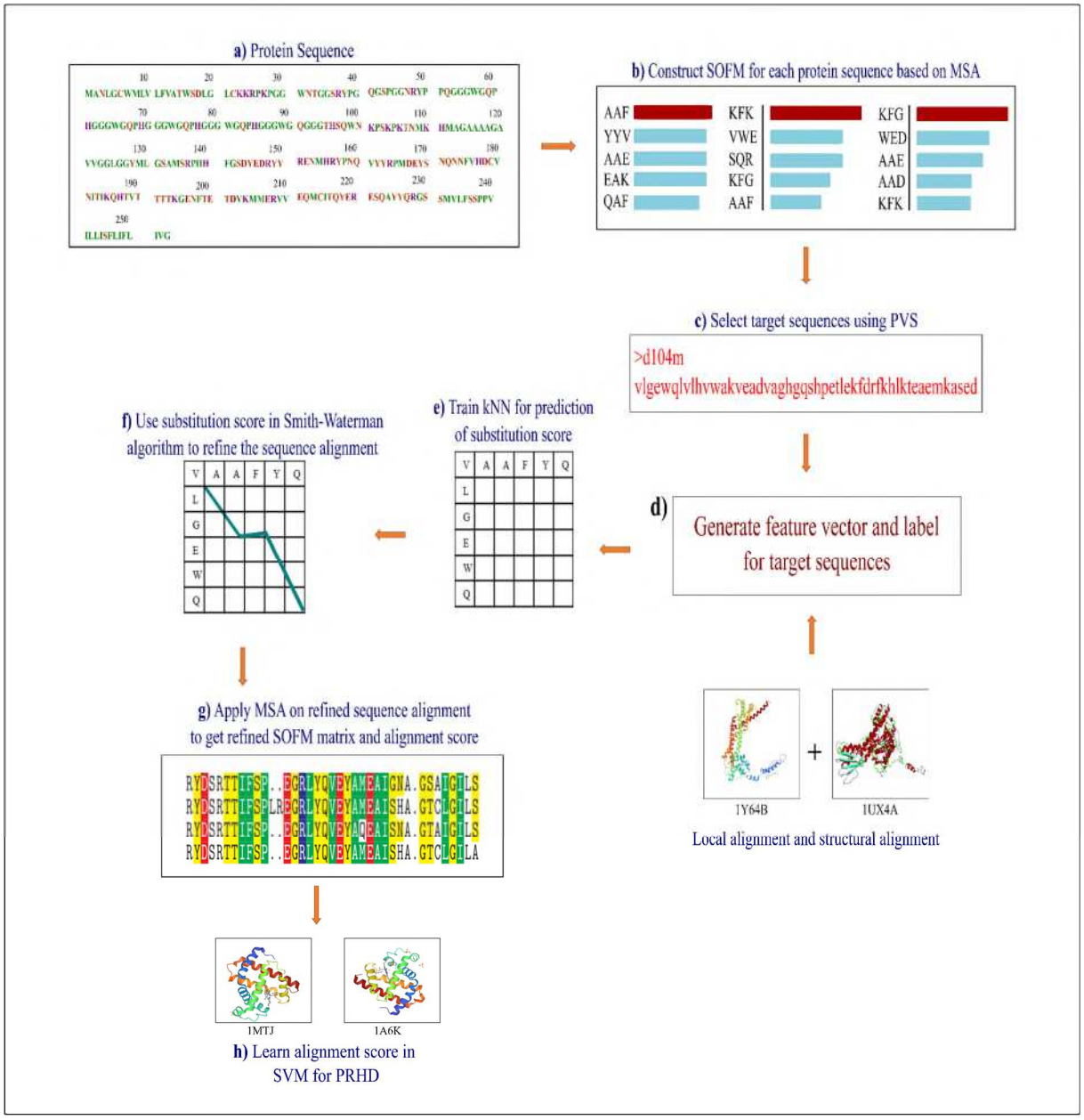

Fig.1: Overall framework for the proposed SOFM-SMSW algorithm

In Eq. (7), $f(\beta_i, S_j)$ denotes the occurring frequency of substring β_i in collection S_j. $\mathcal{F}_{i,j}(\beta_i, S_j)$ equals to $f(\beta_i, S_j)$ when the substring β_i appears in the j-th position of MSA or else 0. $\sum_{i=1}^{20^h} \mathcal{F}_{i,j}(\beta_i, S_j)$ denotes the total occurring frequency of all substrings in position j. The highest score $g_{i,j}$ indicates more probable appearing of substring β_i at j-th position of protein sequence P. The construction of Sequence-Order Frequency Matrix based on Multiple Sequence Alignment is shown in Fig.1 (b).

3.3 Selection of target sequence using PVS

After the computation of alignment score for each protein sequence, the most important sequence is selected using PVS. It chooses target sequence C of size s with probability proportional to $\mu(C)$ times $det(\sum_{i\in C} SOFM_i SOFM_i^T)$ for a measure μ (uniform distribution). Consider $SOFM_s$ is the collection of SOFM of each protein sequence of $[m]$ size exactly s and $SOFM_{\leq s}$ is the subsets of $[m]$ of size atmost s. If μ be a probability measure on set of protein sequence in $SOFM_s$ or $SOFM_{\leq s}$. Then the PVS with measure μ picks a set $C \in SOFM_s$(or $SOFM_{\leq s}$) with probability proportional to (C) $det(\sum_{i\in C} SOFM_i SOFM_i^T)$. Assume the notation $P^C = \prod_{i\in C} p_i$, $SOFM_C$ a matrix of target sequence for $i \in C$, $SOFM_C(P)$ a matrix of column vectors. Consider $e_s(P_1, P_2, .. P_n)$ be the degree s elementry symmetric polynomial in the protein sequences $P_1, P_2, .. P_n$. By principle, $e_0(P) = 1$ for any P. $E_s(SOFM)$ can be defined for any positive semi-definite $m \times m$ matrix to be $e_s(\lambda_1, \lambda_2, \dots \lambda_n)$, where $\lambda(SOFM) = (\lambda_1, \lambda_2, \dots \lambda_n)$ is the vector of eignevalues of $SOFM$. Observe that $E_1(SOFM) = tr(SOFM)$ and $E_n(SOFM) = det(SOFM)$.

Assume $H \subseteq [m]$ be of size no more than an integer d. Then,

$$det(SOFM_T(P)^T SOFM_T(P)) = P^T det(SOFM_T^T SOFM_T) \quad (8)$$

In the PVS method, an SOFM of each protein sequence is given as input along with an integer s and then find the uniform distribution measure μ on SOFM. Then, convex relaxation is solved to get a fractional solution with $\sum_{i=1}^{m} P_i = s$. After that, the SOFM of each protein sequence is sampled with $\Pr[SOFM_{xy} = C] \propto \mu(C) det\left(SOFM_C SOFM_C^T\right)$ and $\mu(C)$ may be defined using the solution of P. Add the $s - |SOFM|$, when $|SOFM| < s$ and finally it returns a set δ which has optimal information in S_{xy}. The selection of target sequence using Proportional Volume Sampling algorithm is shown in Fig.1 (c).

Proportional Volume Sampling Algorithm

Step 1: Given an input $SOFM = [SOFM_1, SOFM_2, \dots SOFM_n]$, s a positive integer and measure on $SOFM_s$

Step 2: Solve convex relaxation to obtain a fractional solution with $\sum_{i=1}^{m} P_i = s$

Step 3: Sample set δ where $\Pr[SOFM_{xy} = C] \propto \mu(C)det\left(SOFM_C SOFM_C^T\right)$ and $\mu(C)$ may be defined using the solution of P.

Step 4: Return δ (target sequence).

3.4 Generate feature vector and label for target sequences

The local alignment and structural alignment of known homologous are used to learn the substitution score which is used in Smith-Waterman algorithm [20] for refining protein sequence alignment. Assume (Q, T) be the query sequence and target protein sequences (1Y64 and 1UX4) [28-29] correspondingly. Initially, feature vector V_{xy} at Q_x and T_y is the concatenation of query and target's residues feature vectors which are given as follows:

$$V_{xy}(P) = \left(H_x^{query}, H_x^{target}\right) \tag{9}$$

where, H is the concatenation of query and target protein sequences around the residue which is given as follows:

$$H_i = \left(h_{i-\frac{w}{2}}, \dots, h_i, \dots, h_{i+\frac{w}{2}}\right) \tag{10}$$

In Eq. (10), w is the window size. This feature vector is defined at each residue pair of the query sequence and target protein sequences. But it is calculated within the areas, where the window moves along with alignment path since information from the residue pairs that are far from the alignment path is not informative. A label L_{XY} is assigned as 0 or 1 at X and Y. The generation of feature and label for target sequence is shown in Fig.1 (d).

$$L_{XY} = \begin{cases} 1, & if\ X\ and\ Y\ matches \\ 0, & otherwise \end{cases} \tag{11}$$

3.5 Train KNN for prediction of substitution score

After labeling the sequences, the pairwise protein structural alignment is calculated using Smith-waterman method. It needs a substitution score for every residue pair which is learned using k-Nearest Neighbor (k-NN). The substitution score is used to forecast the match of the position. It calculates the substitution score by find the distance between V_{XY} and $V_{XY}^{training}$, where V_{XY} is the feature vector of query and target protein sequences in testing dataset and

$V_{XY}^{training}$ is the feature vector of X and Y protein sequences in training dataset. The distance values are sorted and choose the minimum K minimum distances. Then, mean of the values in the K distances is assigned as a substitution score for testing data. Training of KNN algorithm for prediction of substitution score is shown in Fig.1 (e).

Input: V_{XY}, $V_{XY}^{training}$, L_{XY}, Size of k

Output: Substitution score τ_{XY}

Step 1: For query and target protein sequences calculate the distance between V_{XY} and $V_{XY}^{training}$ by using

$$D(V_{00}, V_{ij}) = \left(|V_{01} - V_{i1}|^2 + |V_{02} - V_{i2}|^2 + \cdots + |V_{0j} - V_{ij}|^2\right)^{\frac{1}{2}} \quad (12)$$

Step 2: Sort the distance in descending order and select k minimum distances.

Step 3: Take the mean of the value and it is returned as substitution score of X and Y protein sequences.

3.6 Use substitution score in Smith-Waterman algorithm to refine the sequence alignment

The substitution score is used in the Smith-Waterman method for PRHD. Smith-waterman technique performs sequence alignment for identifying the similar region among two strings of protein sequences such as $X = x_1x_2 \ldots x_n$ and $Y = y_1y_2 \ldots y_n$. A similarity $sim(x, y)$ is given between sequence elements x and y. A matrix M is constructed to find pairs of segments with high degree similarity. Initially set,

$$M_{r0} = M_{oj} = \tau_{xy} = 0 \; for \; 0 \le k \le n \; and \; 0 \le l \le m \quad (13)$$

The preliminary values of M have the interpretation that M_{ij} is the maximum similarity of two segments ending in x_i and y_j, correspondingly. These values are obtained from the relationship

$$M_{ij} = max\left\{M_{i-1,j-1} + \tau_{xy} + sim(x_i, y_j), \max_{k \ge 1}\{M_{i-k,j} - W_k\}, \max_{l \ge 1}\{M_{i,j-1} - W_l\}, 0\right\} \quad (14)$$

In Eq. (13), $l \le i \le n$ and $l \le j \le n$, $M_{i-1,j-1}$ is the score of aligning x_i and y_j, τ_{xy} is the substitution score, $M_{i-k,j} - W_k$ is the score if x_i is at the end of a gap of length k, $M_{i,j-1} -$

W_l is the score if y_j is at the end of gap of length and 0 means there is no similarity between x_i and y_j. Starting at highest score matrix M and ending at a matrix cell which has a score of 0, trace back based on the source of every score recursively to produce the best sequence alignment. Use of substitution score in Smith-Waterman algorithm to refine the sequence alignment is shown in Fig.1 (f).

1. Assume $X = x_1x_2 \dots x_n$ and $Y = y_1y_2 \dots y_m$ be the protein sequences to be aligned, where n and m are the lengths of X and Y correspondingly.
2. Construct SOFM matrix for the protein sequences based on adjacent amino acid substrings.
3. Choose a set of columns in SOFM using PVS.
4. Get the local alignment through constructing $SOFM_X$ and $SOFM_Y$ and their substitution matrix S_{XY}.
5. Apply PVS method on SOFM of each protein sequence to get the target protein sequence.
6. Generate feature vector and label of target sequences using Eq. (9) and Eq. (11).
7. Process the feature vector and label of target sequences in kNN for prediction of substitution score.
8. Calculate the substitution score using k-NN.
9. Construct a scoring matrix $b_{i,j}$ using Smith-Waterman algorithm.
10. Apply MSA on optimal protein sequence alignment and get a refined alignment score.
11. Train the alignment score in SVM for PRHD.
12. SVM learns $b_{i,j}$ for protein remote homology detection.

Fig.2: Pseudocode of the proposed SOFM-SMSW Algorithm

3.7 Apply MSA on refined sequence alignment to get refined SOFM matrix and alignment score

After obtaining the best sequence alignment, MSA is applied on it to get the refined SOFM and refined alignment score. Multiple Sequence Alignment is applied on refined sequence alignment to get refined SOFM matrix and alignment score is shown in Fig.1 (g). The

refined alignment score is given as input to SVM for PRHD (1A6K & 1MTJ) and it is shown in Fig. 1 (h) [30-31]. LIBSVM package is used for the protein remote homology detection. The Radial Basis Function (RBF) kernel is used for the SVM algorithm to predict the remote homologues. Regularization parameter of the SVM is set to 1.0 and the kernel co-efficient gamma is set to 'scale'. The pseudocode of the proposed algorithm is shown in Fig.2.

4. RESULTS AND DISCUSSION

This section describes the effectiveness of proposed SOFM-SMSW algorithm with the existing algorithms such as SVM Top-N-gram, SVM pairwise, GPkernal, LSTM, SOFM-Top and SOFM-SW are tested in terms of Accuracy, Precision, Recall, Receiver Operating Characteristics (ROC) and ROC50. For the experimental purpose, Structural Classification of Proteins (SCOP) 1.53 and SCOP 1.67 benchmark datasets are used. The SCOP 1.53 contains 4532 proteins of 54 families and the SCOP 1.67 contains 4019 sequences of 102 families. The experiments were carried out in HP Intel ® Pentium ® CPU N3710 @ 1.60GHz, 4 GB RAM with running Windows 10 operating system and the proposed SOFM-SMSW algorithm was implemented in java programming language.

4.1 Accuracy

Accuracy metric measures the ratio of correct protein remote homology detection over the total number of proteins evaluated. It is calculated as:

$$Accuracy = \frac{True\ Positive\ (TP) + True\ Negative\ (TN)}{TP + TN + False\ Positive\ (FP) + False\ Negative\ (FN)} \quad (15)$$

Table 1 shows the assessment of accuracy for the proposed and the existing algorithms such as SVM Top-N-gram, SVM pairwise, GPkernal, LSTM, SOFM-Top, SOFM-SW and SOFM-SMSW for two benchmark datasets such as SCOP 1.53 & SCOP 1.67.

Table 1: Comparison of Accuracy for the proposed SOFM-SMSW algorithm with the existing algorithms

Algorithms/Datasets	SCOP 1.53	SCOP 1.67
SVM Top Ngram (SVM TN)	80	79
SVM pairwise (SVM PW)	83	80

GPkernal	84	82
LSTM	81	83
SOFM-Top	82	80
SOFM-SW	86	84
SOFM-SMSW	**89**	**88**

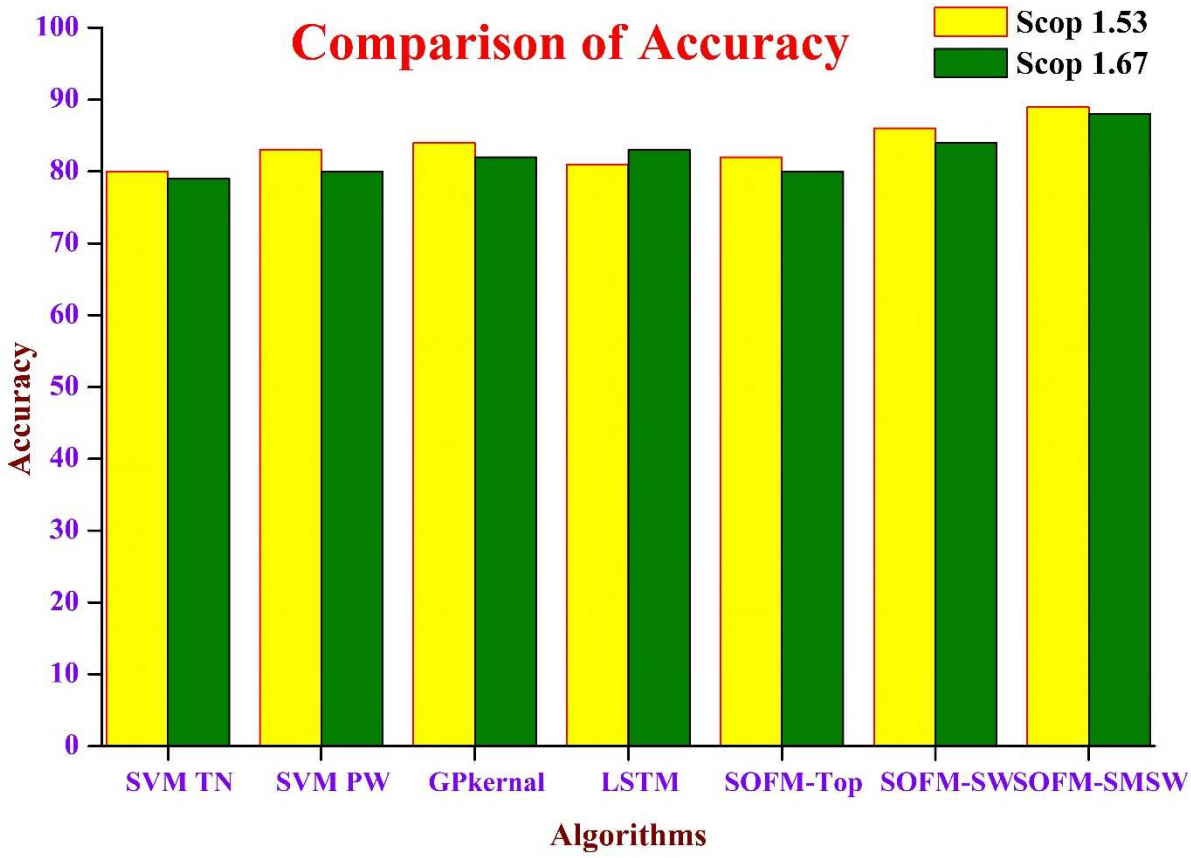

Fig. 3: Comparison of Accuracy for the proposed SOFM-SMSW algorithm with the existing algorithms

Figure 3 shows the accuracy of the proposed and the existing algorithms for SCOP 1.53 and SCOP 1.67 datasets. For SCOP 1.53 dataset, the accuracy of SOFM-SMSW is 10.11% greater than SVM Top Ngram, 6.74% greater than SVM pairwise, 5.61% greater than GPkernal, 8.98% greater than LSTM, 7.86% greater than SOFM-Top and 3.37% greater than SOFM-SW. For the SCOP 1.67 dataset, the accuracy of SOFM-SMSW is 10.22% greater than

SVM Top Ngram, 9.09% greater than SVM pairwise, 6.81% greater than GPkernal, 5.68% greater than LSTM, 9.09% greater than SOFM-Top and 4.54% greater than SOFM-SW. From this analysis, it is proved that the proposed SOFM-SMSW algorithm has the highest accuracy than other methods for SCOP 1.53 and SCOP 1.67 datasets.

4.2 Precision

Precision is defined as the fraction of aligned positions that are correctly aligned based on SVM Top-N-gram, SVM pairwise, GPkernal, LSTM, SOFM-Top, SOFM-SW and SOFM-SMSW methods. Precision is calculated using Eq.16. Table 2 shows the comparison of precision values for the proposed SOFM-SMSW algorithm with the existing algorithms such as SVM Top-N-gram, SVM pairwise, GPkernal, LSTM, SOFM-Top, SOFM-SW and SOFM-SMSW for two benchmark datasets such as SCOP 1.53 & SCOP 1.67.

$$Precision = \frac{TP}{TP+FP} \tag{16}$$

Table 2: Comparison of Precision for the proposed SOFM-SMSW algorithm with the existing algorithms

Algorithms/Datasets	SCOP 1.53	SCOP 1.67
SVM Top Ngram (SVM TN)	79	81
SVM pairwise (SVM PW)	82	81
GPkernal	83	80
LSTM	81	78
SOFM-Top	83	79
SOFM-SW	85	83
SOFM-SMSW	**90**	**88**

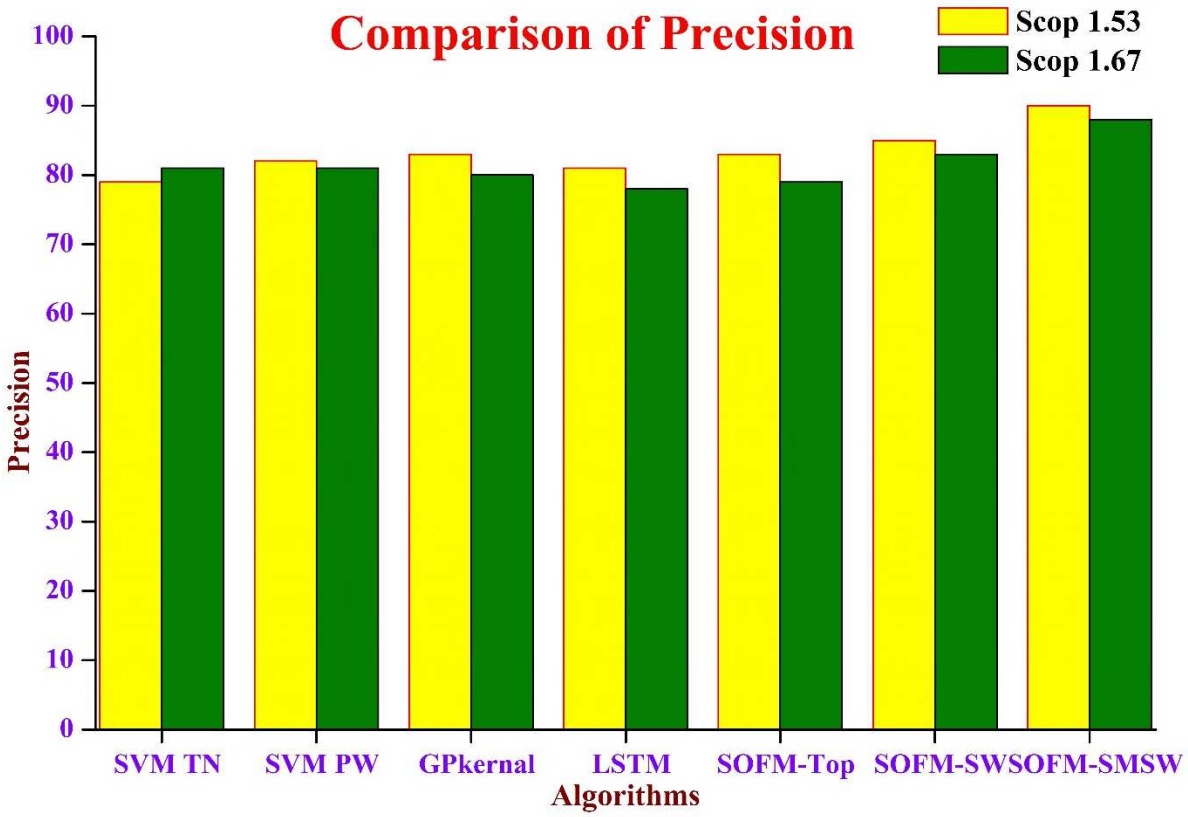

Fig. 4: Comparison of Precision for the proposed SOFM-SMSW algorithm with the existing algorithms

Figure 4 shows the precision values of proposed and the existing algorithms for SCOP 1.53 and SCOP 1.67 datasets. For SCOP 1.53 dataset, the precision of SOFM-SMSW is 12.22% greater than SVM Top Ngram, 8.88% greater than SVM pairwise, 7.77% greater than GPkernal, 10% greater than LSTM, 7.77% greater than SOFM-Top and 5.55% greater than SOFM-SW. For the SCOP 1.67 dataset, the precision of SOFM-SMSW is 7.95% greater than SVM Top Ngram, 7.95% greater than SVM pairwise, 9.09% greater than GPkernal, 11.36% greater than LSTM, 10.23% greater than SOFM-Top and 5.68% greater than SOFM-SW. From this analysis, it is proved that the proposed SOFM-SMSW has higher precision values than other methods for SCOP 1.53 and SCOP 1.67 datasets.

4.3 Recall

Recall is the fraction of align able residues that are correctly aligned based on SVM Top-N-gram, SVM pairwise, GPkernal, LSTM, SOFM-Top, SOFM-SW and SOFM-SMSW methods.

It is calculated using Eq.17. Table 3 shows the comparison of recall values for the proposed SOFM-SMSW algorithm with the existing algorithms such as SVM Top-N-gram, SVM pairwise, GPkernal, LSTM, SOFM-Top, SOFM-SW and SOFM-SMSW for two benchmark datasets such as SCOP 1.53 & SCOP 1.67.

$$Recall = \frac{TP}{TP+TN} \tag{17}$$

Table 3: Comparison of Recall for the proposed SOFM-SMSW algorithm with the existing algorithms

Algorithms/Datasets	SCOP 1.53	SCOP 1.67
SVM Top Ngram (SVM TN)	78	80
SVM pairwise (SVM PW)	82	80
GPkernal	83	81
LSTM	81	79
SOFM-Top	84	80
SOFM-SW	87	84
SOFM-SMSW	**91**	**89**

Figure 5 shows the recall values of proposed and the existing algorithms for SCOP 1.53 and SCOP 1.67 datasets. For SCOP 1.53 dataset, the precision of SOFM-SMSW is 14.28% greater than SVM Top Ngram, 8.89% greater than SVM pairwise, 8.79% greater than GPkernal, 10.99% greater than LSTM, 7.69% greater than SOFM-Top and 4.39% greater than SOFM-SW. For the SCOP 1.67 dataset, the precision of SOFM-SMSW is 10.11% greater than SVM Top Ngram, 10.11% greater than SVM pairwise, 8.99% greater than GPkernal, 11.23% greater than LSTM, 10.11% greater than SOFM-Top and 5.62% greater than SOFM-SW. From this analysis, it is proved that the proposed SOFM-SMSW has higher recall values than other methods for SCOP 1.53 and SCOP 1.67 datasets.

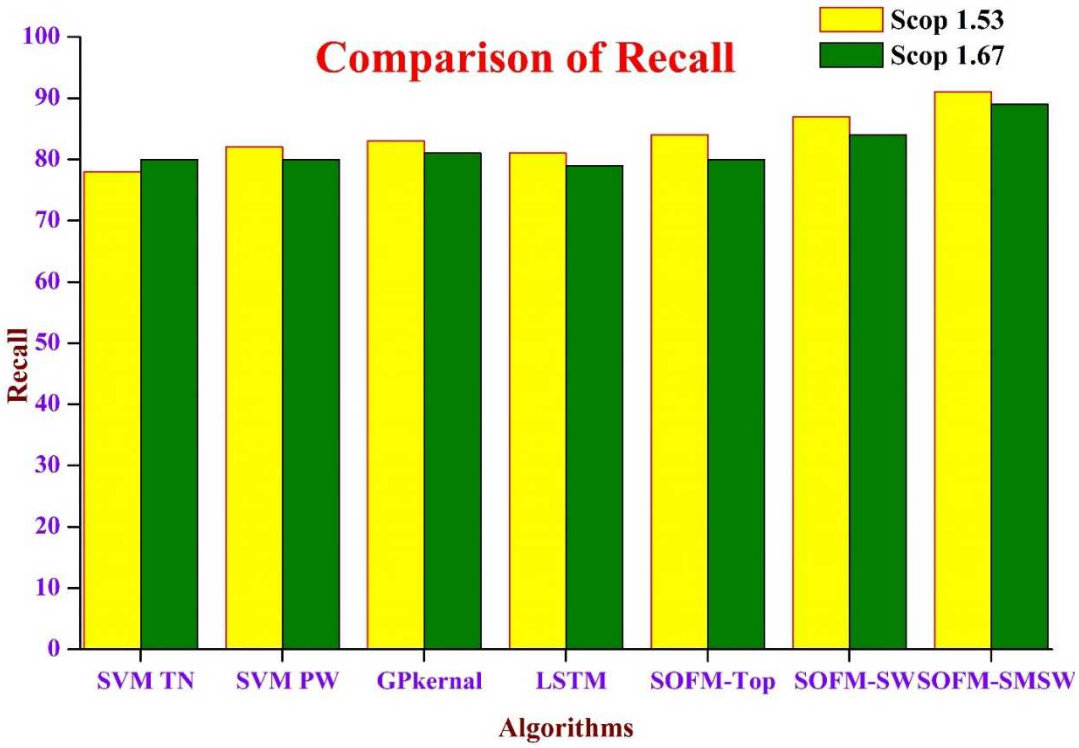

Fig. 5: Comparison of Recall for the proposed SOFM-SMSW algorithm with the existing algorithms

4.4 Receiver Operating Characteristic (ROC)

ROC score is used to calculate the trade-off among specificity and sensitivity. It plots true positives against false positives in the normalized area under a curve. Table 4 illustrates the comparison of ROC & ROC50 values for the proposed SOFM-SMSW algorithm with the existing algorithms such as SVM Top-N-gram, SVM pairwise, GPkernal, LSTM, SOFM-Top, SOFM-SW and SOFM-SMSW for two benchmark datasets such as SCOP 1.53 & SCOP 1.67.

Table 4: Comparison of ROC and ROC50 values for SCOP 1.53 & 1.67 datasets

Algorithms / Datasets	SCOP 1.53		SCOP 1.67	
	Roc	Roc50	Roc	Roc50
SVM Top Ngram (SVM TN)	0.7172	0.4136	0.7378	0.7578
SVM pairwise (SVM PW)	0.7329	0.4465	0.7581	0.7781
GPkernal	0.7621	0.4699	0.7893	0.8093

LSTM	0.8024	0.5090	0.8211	0.8411
SOFM-Top	0.821	0.842	0.714	0.764
SOFM-SW	0.921	0.921	0.753	0.852
SOFM-SMSW	**0.941**	**0.969**	**0.864**	**0.913**

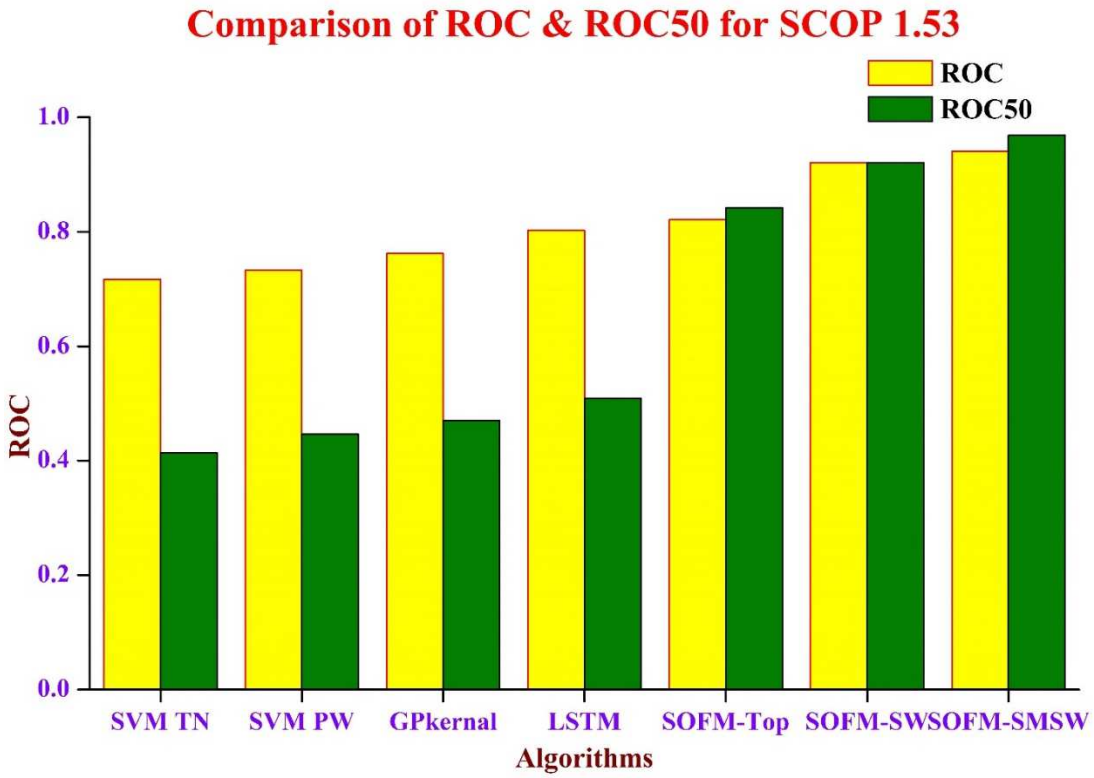

Fig. 6: Comparison of ROC & ROC50 values for the proposed SOFM-SMSW algorithm with the existing algorithms for SCOP 1.53 dataset

Figure 6 & 7 shows the ROC & ROC50 values for the proposed SOFM-SMSW algorithm with the existing algorithms for SCOP 1.53 & 1.67 dataset. For SCOP 1.53 dataset, the ROC of SOFM-SMSW is 23.78% greater than SVM Top Ngram, 22.11% greater than SVM pairwise, 19.01% greater than GPkernal, 14.72% greater than LSTM, 12.75% greater than SOFM-Top and 2.13% greater than SOFM-SW. And the ROC50 of SOFM-SMSW is 57.31% greater than SVM Top Ngram, 53.92% greater than SVM pairwise, 51.51% greater than GPkernal, 47.47% greater than LSTM, 13.11% greater than SOFM-Top and 4.95% greater than SOFM-SW. For the SCOP 1.67 dataset, the ROC of SOFM-SMSW is 14.60% greater than SVM

Top Ngram, 12.26% greater than SVM pairwise, 8.65% greater than GPkernal, 4.97 % greater than LSTM, 17.36% greater than SOFM-Top and 12.85% greater than SOFM-SW. And the ROC50 of SOFM-SMSW is 16.99% greater than SVM Top Ngram, 14.78% greater than SVM pairwise, 11.36% greater than GPkernal, 7.88% greater than LSTM, 16.32% greater than SOFM-Top and 6.68% greater than SOFM-SW. From this analysis, it is proved that the proposed SOFM-SMSW algorithm has better ROC & ROC 50 values than other methods for SCOP 1.53 & 1.67 dataset.

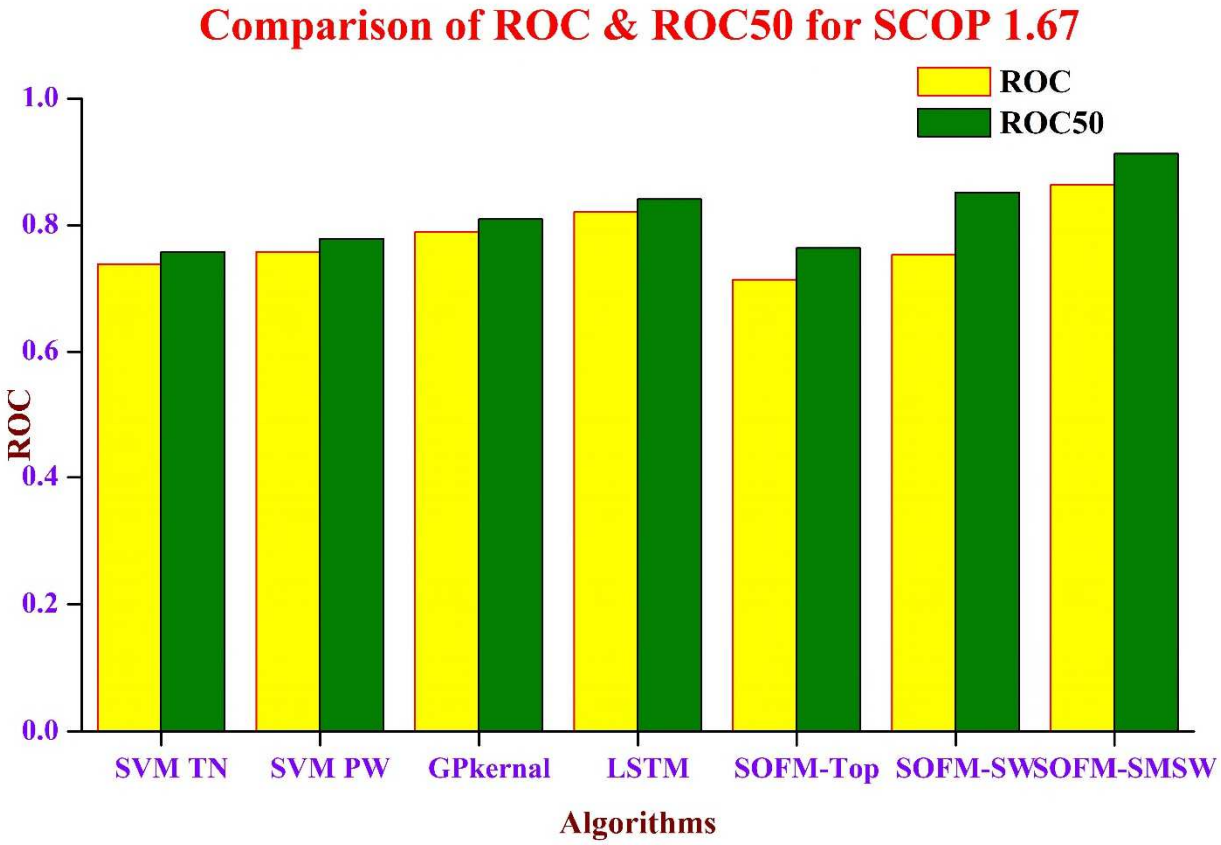

Fig. 7: Comparison of ROC & ROC50 values for the proposed SOFM-SMSW algorithm with the existing algorithms for SCOP 1.67 dataset

5. CONCLUSION

In this research work, the SOFM-SMSW algorithm is proposed for predicting PRHD. The SOFM-SMSW algorithm used the PVS method to select the optimum target sequences based on the uniform distribution measure. Initially, a SOFM matrix is constructed from MSA

and then a uniform distribution of each protein's SOFM is calculated. Based on it, the target sequence is obtained. After that, labeling is performed to find the concatenation position of two protein sequences and it is processed in kNN for prediction of substitution score. It is processed in Smith-Waterman algorithm to refine the sequence alignment and it is processed over MSA which returns refined alignment score. Finally, the alignment score is processed in SVM for PRHD. The experimental results illustrate that the proposed SOFM-SMSW algorithm has better accuracy, precision, recall, ROC and ROC 50 for PRHD than the other existing algorithms**.**

Availability of Data and Materials

The source of data is collected from the Astral Sequences & Subsets in SCOPe online repository.

Scop 1.53 Dataset Link: https://scop.berkeley.edu/astral/ancient/scopseq-1.53/astral-scopdom-seqres-all-1.53.fa

Scop 1.67 Dataset Link: https://scop.berkeley.edu/downloads/scopseq-1.67/astral-scopdom-seqres-gd-sel-gs-bib-40-1.67.fa

Funding

Not applicable

Ethics Approval and Consent to Participate

Not applicable

Human and Animal Rights

No Animals/Humans were used for this study.

Conflict of Interest

The authors declare no conflict of interest.

Consent for Publication

Not applicable

References

[1] Alberts B, Johnson A, Lewis J, et al. Molecular Biology of the Cell, 4th edition. Garland Science: New York, 2002.

[2] Liu X, Zhao L, Dong Q. Protein remote homology detection based on auto-cross covariance transformation. Comput Biol Med. 2011, 41(8), 640-7.

[3] Liu B, Li S. ProtDet-CCH: Protein Remote Homology Detection by Combining Long Short-Term Memory and Ranking Methods. IEEE/ACM Trans Comput Biol Bioinform. 2019 ,16(4), 1203-1210.

[4] Liu B and Zhu Y. ProtDec-LTR3.0: Protein Remote Homology Detection by Incorporating Profile-Based Features Into Learning to Rank. IEEE Access. 2019, 7, 102499-102507.

[5] Mensi A, Bicego M, Lovato P, Loog M, & Tax DM. A dissimilarity-based multiple instance learning approach for protein remote homology detection. Pattern Recognit. Lett., 2019, 128, 231-236.

[6] Ma J, Wang S, Wang Z, Xu J. MRFalign: protein homology detection through alignment of Markov random fields. PLoS Comput Biol. 2014, 10(3), e1003500.

[7] Chen J, Long R, Wang XL, Liu B, Chou KC. dRHP-PseRA: detecting remote homology proteins using profile-based pseudo protein sequence and rank aggregation. Sci Rep. 2016, 6, 32333, 1-7.

[8] Xie S, Li P, Jiang Y, Zhao Y. A discriminative method for protein remote homology detection based on N-Gram. Genet Mol Res. 2015, 14(1), 69-78.

[9] Liu B, Chen J, Guo M, Wang X. Protein Remote Homology Detection and Fold Recognition Based on Sequence-Order Frequency Matrix. IEEE/ACM Trans Comput Biol Bioinform. 2019, 16(1), 292-300.

[10] Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol. 1981, 147(1), 195-197.

[11] Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol. 1995, 247(4), 536-540.

[12] Muda HM, Saad P, Othman RM. Remote protein homology detection and fold recognition using two-layer support vector machine classifiers. Comput Biol Med. 2011, 41(8), 687-99.

[13] Daniels NM, Gallant A, Ramsey N, Cowen LJ. MRFy: Remote Homology Detection for Beta-Structural Proteins Using Markov Random Fields and Stochastic Search. IEEE/ACM Trans Comput Biol Bioinform. 2015,12(1), 4-16.

[14] Paliwal KK, Sharma A, Lyons J, Dehzangi A. A tri-gram based feature extraction technique using linear probabilities of position specific scoring matrix for protein fold recognition. IEEE Trans Nanobioscience. 2014,13(1), 44-50.

[15] Lovato P, Cristani M, Bicego M. Soft Ngram Representation and Modeling for Protein Remote Homology Detection. IEEE/ACM Trans Comput Biol Bioinform. 2017, 14(6), 1482-1488.

[16] Lyons J, Paliwal KK, Dehzangi A, Heffernan R, Tsunoda T, Sharma A. Protein fold recognition using HMM-HMM alignment and dynamic programming. J Theor Biol. 2016, 393, 67-74.

[17] Yan K, Xu Y, Fang X, Zheng C, Liu B. Protein fold recognition based on sparse representation based classification. Artif Intell Med. 2017, 79,1-8.

[18] Ibrahim W, Abadeh MS. Extracting features from protein sequences to improve deep extreme learning machine for protein fold recognition. J Theor Biol. 2017, 421, 1-15.

[19] P. Manikandan, D. Ramyachitra. Bacterial Foraging Optimization –Genetic Algorithm for Multiple Sequence Alignment with Multi Objectives, Scientific Reports, 2017, 7, 8833, 1-14.

[20] Zhang P, Tan G, Gao GR. Implementation of the Smith-Waterman algorithm on a reconfigurable supercomputing platform. In Proceedings of the 1st international workshop on High-performance reconfigurable computing technology and applications: held in conjunction with SC07, 2007, 39-48.

[21] Li S, Chen J, Liu B. Protein remote homology detection based on bidirectional long short-term memory. BMC Bioinformatics. 2017, 18(1), 443.

[22] Chen J, Liu B, Huang D. Protein Remote Homology Detection Based on an Ensemble Learning Approach. Biomed Res Int. 2016, 2016, 5813645

[23] Manikandan P, Ramyachitra D. PATSIM: Prediction and analysis of protein sequences using hybrid Knuth-Morris Pratt (KMP) and Boyer-Moore (BM) algorithm. Gene. 2018, 657, 50-59.

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

[25] Song Gao, Shui Yu & Shaowen Yao. An efficient protein homology detection approach based on seq2seq model and ranking. Biotechnology & Biotechnological Equipment. 2021, 35(1), 633-640.

[26] Guo Y, Yan K, Wu H, Liu B. ReFold-MAP: Protein remote homology detection and fold recognition based on features extracted from profiles. Anal Biochem. 2020, 611, 114013.

[27] Routray M, N. Ray. Remote homology detection using GA and NSGA-II on physicochemical properties. Journal of Computer Applications in Technology. 2021, 64 (4): 393-402.

[28] Otomo T, Tomchick DR, Otomo C, Panchal SC, Machius M, Rosen MK. Structural basis of actin filament nucleation and processive capping by a formin homology 2 domain. Nature. 2005, 433(7025), 488-94

[29] Xu Y, Moseley JB, Sagot I, Poy F, Pellman D, Goode BL, Eck MJ. Crystal structures of a Formin Homology-2 domain reveal a tethered dimer architecture. Cell. 2004, 116(5), 711-23.

[30] Lai HH, Li T, Lyons DS, Phillips GN Jr, Olson JS, Gibson QH. Phe-46(CD4) orients the distal histidine for hydrogen bonding to bound ligands in sperm whale myoglobin. Proteins. 1995, 22(4), 322-39.

[31] Vojtechovský J, Chu K, Berendzen J, Sweet RM, Schlichting I. Crystal structures of myoglobin-ligand complexes at near-atomic resolution. Biophys J. 1999, 77(4), 2153-74.