Molecular BioSystems

PSOFuzzySVM-TMH: Identification of Transmembrane Helix Segments Using Ensemble Feature Space by Incorporated Fuzzy Support Vector Machine

Journal:	*Molecular BioSystems*
Manuscript ID:	MB-ART-03-2015-000196.R1
Article Type:	Paper
Date Submitted by the Author:	04-May-2015
Complete List of Authors:	Hayat, Maqsood; Abdul Wali Khan University Mardan, Computer Science Tahir, Muhammad; Abdul Wali Khan University Mardan, Computer Science

PSOFuzzySVM-TMH: Identification of Transmembrane Helix Segments Using Ensemble Feature Space by Incorporated Fuzzy Support Vector Machine

Maqsood Hayat*, Muhammad Tahir

Department of Computer Science, Abdul Wali Khan University, Mardan, Pakistan

*Corresponding Author: Maqsood Hayat

Email Address: m.hayat@awkum.edu.pk, maqsood.hayat@gmail.com

Office Ph: 92-937-542194. Fax 92-937-542194

Abstract:

Membrane protein is a central component of the cell, which manages intra and extracellular processes. Membrane proteins execute diversity of functions, which are vital to the survival of organisms. Topology of transmembrane proteins describes number of transmembrane (*TM*) helix segments and its orientation. However, owing to the lack of its recognized structures, identification of *TM* helix and its topology through experimental methods is laborious and low throughput. In order to identify *TM* helix segments reliably, accurately, and effectively from topogenic sequences, we propose *PSOFuzzySVM-TMH* model. In this model, evolutionary based information position specific scoring matrix and discrete based information 6-letter exchange group are used to formulate transmembrane protein sequences. The noisy and extraneous attributes are eradicated using optimization selection technique particle swam optimization from both feature spaces. Finally, the selected feature spaces are combined in order to form in order to form ensemble feature space. Fuzzy-Support vector Machine is utilized as a classification algorithm. Two benchmark datasets including low and high resolution datasets are used. At various levels the performance of *PSOFuzzySVM-TMH* model is assessed through 10-fold cross validation test. The empirical results reveal that the proposed framework *PSOFuzzySVM-TMH* outperforms in term of classification performance in the examined datasets. It is ascertained that the proposed model might be a useful and high throughput tool for academia and research community for further structure and functional studies on transmembrane proteins.

Keywords: Transmembrane helix; PSSM; 6-letter exchange group; PSO; Fuzzy-SVM.

1. Introduction

Membrane proteins are a major constituent of the cell, which control internal and external processes of a cell. It plays a central role in cellular processes ranging from basic molecule transport to sophisticated signaling pathways. Currently, in market, more than half of all drugs are directly targeted against the membrane proteins [1]. However, it is complex and challenging to get high-resolution three-dimensional (*3D*) structures of membrane proteins. Only a limited number of membrane protein structures are available in protein Data Banks [2]. Membrane proteins contain one or more transmembrane (*TM*) helices, which express the orientation or topology of a membrane protein corresponding to the lipid bilayer. Alpha helix is a prime category of transmembrane proteins, which perform most of the important biological processes of a cell such as cell signaling, cell-to-cell interaction, cell recognition, and adhesion. Information regarding *TM* helix provides some useful intimation in determining the function of membrane proteins. Since, the determination of the crystal structure of membrane proteins by X-ray or nuclear magnetic resonance (*NMR*) is extremely difficult; therefore, computational methods are considered as valuable tools for correctly identifying locations of *TM* helix segments and topology of *TM* helix proteins.

In the last few decades, a series of efforts have been carried out for identifying the orientation of *TM* helix segments. In the early studies, usually the investigation was made on the basis of physicochemical properties of amino acids namely, hydrophobicity [3-9], charge [6, 10, 11], nonpolar phase helicity [12], and multiple sequence alignment [13, 14]. DAS-TMfilter [15], TOP-Pred [10], and SOSUI [11] and found the most substantial models that give descriptive information about *TM* helices. The performance of these models were quite promising merely for *TM* helix segments rather than topology prediction of *TM* helix proteins. In addition, several researchers have used

various statistical models such as Hidden Markov Models (*HMM*), support vector machine (*SVM*), and artificial neural networks for predicting *TM* helix segments. In addition, several user-friendly web predictors have also been developed for the benefit of academies and researchers. A few of them include *TopPred* [10], *MEMSAT* [16], *PHD* [17], *HMMTOP* [18, 19], *TMHMM* [20, 21], *PRODIV_TMHMM* [22], *TMMOD* [23], Phobius [24], *ENSEMBLE* [25], *PONGO* [26], *HMM-TM* [27], *MemBrain* [28], *MEMSAT-SVM* [29], *MEMPACK* [30], and *SVMtop* [31]. Multiple sequence alignments and computational cost are remained the target issues of *HMM* based models. However, the main drawback of *HMM* based models is unexecutable in case of shorter than16 residues *TM* helix segments or longer than 35 residues *TM* helix segments [28]. Few researchers have concentrated on accuracy as well as sensitivity and specificity for analyzing their proposed models [28, 32-34]. Furthermore, several studies have emphasized only on sensitivity and reliabilities of different models rather than accuracy [35-39].

In this study, a more powerful, accurate and high throughput model is proposed for identification of TM helix segments. In this model, two protein sequences representation methods namely: position specific scoring matrix and 6-letter exchange group are used to extract salient features. After that, evolutionary feature selection technique particle swarm optimization is applied to select noisy free features. Finally, the selected features of both the spaces are combined in order to form an ensemble feature space. Fuzzy-SVM is utilized as classification algorithm. 10-fold cross validation is applied to assess the performance of proposed model.

The remaining paper is structured as follows: first, Section 2 describes Materials and Methods. Next, Section 3 explains the proposed system. Then, Section 4 describes performance measures, while Section 5 presents results and discussion. Finally, Section 6 draws the conclusion.

2. Materials and Methods

2.1 Benchmark datasets

In this study, we have used two benchmark datasets. *Dataset1* is a low-resolution transmembrane protein dataset, which was developed by Moller et al. [40]. It is annotated from *SWISS-PROT* release 49.0 [41]. Initially, it contained 145 protein sequences, but later two protein sequences were discarded, which had no annotation with transmembrane proteins. Finally, *Dataset1* consists of 143 protein sequences, which include 687 *TMH* segments.

Dataset2 is a High-resolution membrane protein dataset. In this dataset, 101 transmembrane protein sequences of *3-D* structure helix are selected from *MPtopo* database [42], while 231 transmembrane protein sequences are obtained from *TMPDB* database [43]. After combining both the datasets, 30% *CD*-Hit has been applied to reduce the redundancy and similarity. After this screening, *Dataset2* contains 258 single and multispanning transmembrane protein sequences, which consist of 1,232 *TMH* segments.

2.2 Sample Formulation Techniques

In this work, we have used two different types of protein sequence representation methods including Position-specific scoring matrix (*PSSM*) and 6-letter exchange group representation to extract pertinent and useful information from transmembrane protein sequences.

2.2.1 Position-specific scoring matrix

PSSM is evolutionary profiles and motif based descriptive, which exploits multiple alignments and profiles about protein families. In *PSSM*, each amino acid residue has against 20 values, which determine the frequencies of substitutions detected at the specific position in the protein family. *PSSM* matrix consists of negative and positive scores; negative indicates less

substitution in the alignment while the positive shows more substitutions have been taken place in the alignment. Let us consider a protein sequence *P* with *N* residues long, *PSSM* can be generated as:

$$P_{PSSM} = \begin{bmatrix} X_{1\to1} & X_{1\to2} & \dots & X_{1\to j} & \dots & X_{1\to20} \\ X_{2\to1} & X_{2\to2} & \dots & X_{2\to j} & \dots & X_{2\to20} \\ \vdots & \vdots & \dots & \vdots & \dots & \vdots \\ X_{i\to1} & X_{i\to2} & \dots & X_{i\to j} & \dots & X_{i\to20} \\ \vdots & \vdots & \dots & \vdots & \dots & \vdots \\ X_{N\to1} & X_{N\to2} & \dots & X_{N\to j} & \dots & X_{N\to20} \end{bmatrix} \quad (1)$$

where $X_{i\to j}$ shows the i^{th} position residue in the protein sequence, which is substituted by amino acid type *j* in the biological evolutionary process. The values of *j=1… 20* represent the alphabetical order of 20 native amino acids. The P_{PSSM} is obtained by executing *PSI-BLAST* [44, 45], which explored the Swiss-Prot database in three iterations with the cutoff *E*-value of 0.001 for multiple sequence alignment against the sequence of the protein query *P*. Consequently, $N \times 20$ scoring matrix is generated.

In order to extract attributes from *PSSM* matrix, we have taken sliding window centered on a target residue with four residues on each side of the target residue. As a result 180-D feature space is produced. The original score in each position is normalized by using logistic function as given below [46]:

$$f(x) = 1/(1+e^{-x}) \quad (2)$$

where x is the original score.

2.2.2 *6*-letter Exchange Group Representation

Protein sequence is composed of twenty amino acids. In these amino acids some of them have similarity in their structure. In this feature extraction technique, amino acids are categorized into six different classes, which are called 6-letter exchange group representation. The exchange groups show the effects of evolution. First, the protein sequence is converted into its equivalent 6-letter exchange group representation sequence [47] shown in Table 1, which has been derived from the PAM matrix [48]. For example, all *K*, *H*, and *R* amino acids in the original sequence are replaced by e_1 and *D, E, N,* and *Q* are replaced by e_2 and *C* is substituted by e_3 etc. After substituting the amino acids by 6-letter the resulting sequence contains only six different characters. Then, we applied sliding window, 6 features are extracted against each position and moves the window to the next position. This process is repeated up to the last residue of the protein sequence.

$$p_i = (e_1, e_2, e_3, e_4, e_5, e_6) \quad (3)$$

$$S_i = [C_{ij}]_{1*6} \quad (4)$$

where C_{ij} is the occurrence frequency of exchange group e_j in window i. Finally, the resultant matrix is

$$P = [S_1^T S_2^T \ldots S^T_{N-m+1}]_{6*N-m+1} \quad (5)$$

where T represents transpose, N is length of protein sequence and m is the window size.

2.3 Particle Swarm Optimization (*PSO*)

Usually, high throughput prediction model requires precise and correction observation. In this regards, feature selection method is required to select high discriminative features, reduce noise, and enhance speed and performance. For this purpose, we have adopted intelligent feature selection technique, *PSO* in which selection and training are processed concurrently and consequently the computational cost is reduced.

PSO is a population based stochastic evolutionary approach introduced by Eberhart and Kennedy in 1995 [49]. It is inspired by nature of social behavior simulation found among different species. *PSO* is a global optimization algorithm, which was applied mostly for nonlinear function optimization, neural network training, and pattern recognition [50]. In the algorithm, a swarm contains N particles moving around in M-dimensional search space. The location of the i^{th} particle at *PSO* iteration t is denoted as:

$$X_i^t = \left(\mathrm{x}_{i1}, x_{i2}, \ldots, x_{iD}\right) \tag{6}$$

During the search process the particle successively adjusts its position toward the global optimum according to the two factors: the best position encountered by itself (*pbest*) denoted as $P_i = \left(p_{i1}, \mathrm{p}_{i2}, \ldots, \mathrm{p}_{iD}\right)$ and the best position encountered by the whole swarm (*gbest*) denoted as $P_g = \left(p_{g1}, \mathrm{p}_{g2}, \ldots, \mathrm{p}_{gD}\right)$. Its velocity at iteration t is represented by $V_i^t = \left(v_{i1}, \mathrm{v}_{i2}, \ldots, \mathrm{v}_{iD}\right)$.

The position at next iteration is calculated according to the following equations:

$$V_i^{(t)} = \lambda\left(w \times V_i^{(t-1)} + c_1 \times rand \times \left(P_i - X_i^{(t-1)}\right) + c_2 \times rand \times \left(P_g - X_i^{(t-1)}\right)\right) \tag{7}$$

$$X_i^{(t)} = X_i^{(t-1)} + V_i^{(t)} \quad (8)$$

where c_1 and c_2 are two positive constants, called cognitive learning rate and social learning rate respectively; *rand*() is a random function in the range [0, 1], w is the inertia factor; and λ is the constriction factor. In addition, the velocities of particles are confined within $[V_{min}, V_{max}]^D$. If an element of velocities exceeds the threshold V_{min} or V_{max}, it is set equal to the corresponding threshold.

2.4 Fuzzy Support Vector Machine

Support vector machine (*SVM*) is a popular learning hypothesis mostly utilized in pattern classification and nonlinear regression [51, 52]. The core idea of *SVM* is to draw a hyperplane in such a way to maximize the margin of separation between two classes. *SVM* was initially used for binary problems but later it was applied for multiclass problems. In case of multiclass problem it converts *n*-classes into *n*-two classes, which classify i^{th} class from the remaining classes. Let the i^{th} decision hyperplane that classifies class *i* and the remaining classes be

$$D_i(x) = w_i^t + b_i \quad (9)$$

The hyperplane will be optimal hyperplane when $D_i(x) = 0$, the instances belonging to class *i* that satisfy $D_i(x) = 1$ and the instances belonging to the remaining class that satisfy $D_i(x) = -1$. If $D_i(x) > 0$ is satisfied for one *i* then *x* is classified into class *i*.

On the other hand, if $D_i(x) > 0$ is satisfied for more *i's* or there is no *i* that satisfies this equation then *x* is unclassifiable. In order to resolve this issue, fuzzy membership function is introduced. In this method, one dimensional membership functions $m_{i,j}(x)$ on the directions orthogonal to the optimal separating hyperplane $D_j(x) = 0$ is defined for class *i* as given

1. $For \quad i = j$

$$m_{i,i}(x) = \begin{cases} 1 & for\ D_i(x) > 1 \\ D_i(x) & otherwise \end{cases} \tag{10}$$

2. $For \quad i \neq j$

$$m_{i,j}(x) = \begin{cases} 1 & for\ D_j(x) < -1 \\ -D_j(x) & otherwise \end{cases} \tag{11}$$

$D_i(x) > 1$ means there is only class *i* training data is available so the degree of class *i* is 1 and otherwise $D_i(x)$. For $i \neq j$ class *i* is on the negative side of $D_j(x) = 0$. In this case support vectors may not include class *i* data but when $D_i(x) < -1$ we assume that the degree of membership of class *i* is 1 and otherwise $-D_j(x)$. We define the class *i* membership function if *x* using the minimum operator for $m_{i,j}(x) \qquad (j = 1,...,n)$

$$m_i(x) = \min_{j=1...n} m_{i,j}(x) \tag{12}$$

Now the datum *x* is classified into the class

$$arg \max_{i=1...n} m_i(x) \tag{13}$$

If *x* satisfies

$$D_k(x) \begin{cases} > 0 & For\ k=i \\ \leq 0 & For\ k \neq i\ k=1,...,n \end{cases} \tag{14}$$

From (10) and (11) $m_i(x) > 0$ and $m_j(x) \leq 0 (j \neq i, j = 1,...,n)$ hold. Thus *x* is classified into class *i*. Now suppose $D_i(x) > 0$ is satisfied for $i_1, \ldots, i_l (l > 1)$ Then from (10) to (12) $m_k(x)$ is given as follow

$$1.\ k \in \{i_1,...,i_l\}$$

$$m_k(x) = \min_{j=i_1,...i_l, j \neq k} -D_j(x) \tag{15}$$

2. $k \neq j (j = i_1, ..., i_l)$

$$m_k(x) = \min_{j=i_1,...,i_l} -D_j(x) \quad (16)$$

Thus the maximum degree of membership is achieved among $m_k(x), k = i_1, \ldots, i_l$ $D_k(x)$. Namely, $D_k(x)$ is maximized in $k \in \{i_1, \ldots, i_l\}$. Let $D_i(x) > 0$ be not satisfied for any class then

$$D_i(x) < 0 \quad \text{for } j=1,...,n \quad (17)$$

Then (11) is given

$$m_i(x) = D_i(x) \quad (18)$$

The procedure of classification is given below

1. For x, if $D_i(x) > 0$ is satisfied only for one class, the input is classified into the class. Otherwise go to step2.
2. If $D_i(x) > 0$ is satisfied for more than one class $i (i = i_1, \ldots, i_l, l > 1)$ classify the datum into the class with the maximum $D_i(x) (i \in \{i_1, \ldots, i_l\})$ otherwise go to step 3.
3. If $D_i(x) \leq 0$ is satisfied for all the classes, classify the datum into the class with the minimum absolute value of $D_i(x)$.

3. Proposed Prediction System for *TM* Helix Segments (*PSOFuzzySVM-TMH*)

In this work, we develop a more powerful, accurate and reliable prediction model *PSOFuzzySVM-TMH* for the identification of transmembrane helix segments. In this model, features are extracted by two different protein sequence representation methods including evolutionary information *PSSM* and 6-letter exchange group representation. In *PSSM* 20 values are extracted against each residue of protein sequence using *PSI-Blast*, which determine the frequencies of substitutions detected at the specific position in a protein family.

Then applied sliding window centered on a target residue with four residues on each side of the target residue. Consequently 180-D feature space is generated.

In contrast, using 6-letter exchange group representation, first, the protein sequence is converted into 6 letters exchange group. Then applied sliding window, 6 features are extracted against each position and moves the window to the next position. This process is repeated up to the last residue of the protein sequence. In order to select high discriminative features and eliminate the redundancy and extraneous features; we have utilized an evolutionary feature selection method *PSO* on each feature space separately. Consequently, 90-D features are selected from *PSSM* feature space and 4-D features are selected from 6-letter exchange group. Furthermore, both the selected feature spaces are combined to form an ensemble feature space. As a resultant, the dimension of the ensemble feature space is 94-D. In addition, *Fuzzy SVM* is utilized as a classification. The framework of the proposed approach is illustrated in Fig. 1.

4. Metrics for measuring Prediction Quality

Various performance measures including accuracy, recall, precision, and *MCC* are used to evaluate the performance of *PSOFuzzySVM-TMH* model at different levels such as per protein, per segment, and per residue, respectively.

$$Q_{htm}^{\%obsd} = \left(number\ of\ correctly\ predicted\ TM\ in\ dataset \Big/ Total\ number\ of\ TM\ in\ dataset \right) \times 100 \quad (19)$$

where $Q_{htm}^{\%obsd}$ indicates the recall of *TM* helix segments.

$$Q_{htm}^{\%\,prd} = \left(number\ of\ correctly\ predicted\ TM\ in\ dataset \Big/ number\ of\ TM\ predicted\ in\ dataset \right) \times 100 \quad (20)$$

where $Q_{htm}^{\%\,prd}$ represents the precision of *TM* helix segments.

$$Q_{ok} = \left(\sum_{i}^{N_{Prot}} \delta_i \Big/ N_{Prot} \right) \times 100 \quad \delta_i = \begin{cases} 1, \ if\ Q_{htm}^{\%obsd} \wedge Q_{htm}^{\%\,prd} = 100\ for\ protein\ i \\ 0, \ otherwise \end{cases} \quad (21)$$

where Q_{ok} indicates the number of protein sequences in which all its *TM* helix segments are correctly predicted.

$$Q_2 = \left(\sum_{i}^{N_{Prot}} \left(number\ of\ residues\ predicted\ correctly\ in\ protein\ i / number\ of\ residues\ in\ poriten\ i \right) \Big/ N_{Prot} \right) \times 100 \quad (22)$$

where Q_2 shows the percentage of correctly predicted residues in both the *TM* helix and non-*TM* helix segments.

$$Q_{2T}^{\%obsd} = \left(number\ of\ residues\ correctly\ predicted\ in\ TM\ helices \Big/ number\ of\ residues\ observed\ in\ TM\ helices \right) \times 100 \quad (23)$$

where $Q_{2T}^{\%obsd}$ measures that how many residues are correctly predicted in the observed residues.

$$Q_{2T}^{\%\,prd} = \left(number\ of\ residues\ correctly\ predicted\ in\ TM\ helices \Big/ number\ of\ residues\ predicted\ in\ TM\ helices \right) \times 100 \quad (24)$$

where $Q_{2T}^{\%\,prd}$ measures that how many residues are correctly predicted in the predicted residues.

$$MCC = \left(TP \times TN - FP \times FN \Big/ \sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)} \right) \quad (25)$$

MCC is a Mathew correlation coefficient, where the value of *MCC* is in the range of -1 and 1. In equation 25, *TP* is the number of correctly predicted *TM* helix residues; *FP* is the number of incorrectly predicted *TM* helix residues, *TN* is the number of correctly predicted non- *TM* helix residues, and *FN* is the number of incorrectly predicted non-*TM* helix residues.

5. Result and Discussion

In order to evaluate the performance of computational model, mostly researchers have applied various cross validation tests. In cross validation tests, jackknife test has extensively been applied by many investigators due to its distinguishable characteristics [50, 51, 53]. Despite of its special features, it is computationally expensive with respect to time and space. In order to utilize the special characteristics of jackknife test along with minimum computational cost, we have applied 10-fold cross validation. The performance of our proposed model with full feature space and selected feature space of the both feature extraction schemes along with their ensemble space are mentioned below. In this work, the performance of computational model is analyzed at three levels, per protein, per segment, and per residue, respectively.

5.1 Prediction performance of *PSOFuzzySVM-TMH* on *PSSM* feature space

The success rates of *PSOFuzzySVM-TMH* by incorporated *PSSM* based full and selected feature spaces are listed in Table 2. In case of low resolution dataset, the obtained accuracy of proposed model at protein level is 67.8% whereas at segment level the precision and recall are 94.3% and 93.6%, respectively. At per residue level the predicted results of proposed model are 88.0% accuracy, 87.2% precision, 79.2% recall and 0.77 *MCC*. Using high resolution dataset, the accuracy of proposed model at protein level is 70.1%. At segment level the precision is 96.1% and recall is 95.2%, whereas at residue level the accuracy of proposed model is 90.9%, precision 86.7%, recall 91.4% and *MCC* 0.82.

In order to enhance the discrimination power of classification algorithm, we have applied *PSO* to select highly discriminative features from feature space. Consequently, 90 features are selected from PSSM feature space, which shows the highest success rates so far. The predicted results of *PSOFuzzySVM-TMH* using selected feature space are reported in Table 2. In case of

low resolution dataset, the proposed model yielded 71.3% accuracy at protein level. At segment level, the precision and recall of proposed model are 94.6% and 95.3%, whereas, the accuracy, precision, recall and *MCC* of the proposed model at residue level are 89.5%, 88.9%, 81.2%, and 0.78, respectively. On the other hand, the accuracy of proposed model using high resolution dataset is 72.6% at protein level, the precision and recall at segment level are 97.0% and 96.7%. At residue level the proposed model achieved 92.0% accuracy, 88.4% precision, 92.6% recall, and 0.83 *MCC*.

5.2 Prediction performance of *PSOFuzzySVM-TMH* on 6-letter Exchange Group feature space

The prediction results of our proposed model using full and selected feature spaces of 6-letter exchange group are shown in Table 3. In case of low resolution dataset, the predicted accuracy of proposed model at protein level is 69.2%, at segment level, its precision and recall are 94.7% and 94.1% respectively. Its success rates at residue level are 88.3% accuracy, 87.9% precision, 80.2% recall, and 0.77 *MCC*. In contrast, using high resolution dataset, the performance of proposed model at protein and segment level is 70.1% accuracy, 95.2% precision and 96.0% recall. Its predicted results at residue level are 90.2%, 86.9%, 91.8%, 0.81 accuracy, precision, recall and *MCC* respectively. In case of selected feature space, at protein level, the accuracy of proposed model is 72.0% and 73.9% using low and high resolution datasets. The empirical results revealed that the performance of our proposed model in conjunction with selected feature space is higher than that of using full feature space.

5.3 Prediction performance of *PSOFuzzySVM-TMH* using ensemble feature space

In order to enhance the classification performance of our proposed model, we have fused both the feature spaces and formed an ensemble feature space. The performance of our proposed model using ensemble space is reported in Table 4. Using low resolution dataset, the accuracy of proposed model at protein level is 75.5% whereas at segment level its precision and recall are 95.6% and 95.7%, respectively. Its performance at residue level is 90.7% accuracy, 89.1% precision, 83.4% recall and 0.79 *MCC*. In case of high resolution dataset, the performances of proposed model at protein and segment level are 77.5% accuracy, 96.6% precision and 96.3% recall. On the hand, its success rates at residue level are 92.5%, 90.3%, 93.2%, and 0.84 accuracy, precision, recall, and MCC, respectively.

Furthermore, we have concatenated the selected feature spaces and formed an ensemble feature space. After that, the performance of our proposed model is more enhanced compared to unselected feature spaces is shown in Table 4. In case of low resolution dataset, our proposed model achieved 77.6% accuracy at protein level whereas, 97.0% and 97.1% precision and recall, respectively at segment level. At residue level, the accuracy, precision, recall, and *MCC* of our proposed model are 93.8%, 91.8%, 85.1%, and 0.81, respectively. On the other hand, using high resolution dataset, the success rates of our proposed model are also listed in Table 4. Our proposed model obtained 79.3% accuracy at protein level and 94.1% accuracy at residue level. Its precision and recall at segment and residue levels are 97.5%, 98.2%, 92.8% and 95.7%, respectively.

After analyzing the results, we have observed that the performance of our proposed model is better using selected feature spaces compared to unselected feature spaces. Further, the performance of our proposed model is sound using 6-letter exchange group in case of individual feature space. In contrast, the performance of our proposed model in conjunction with ensemble

feature space is more enhanced than that of individual feature space, because, we have combined the discrimination power of the two feature spaces. In addition, the performance of proposed model using high resolution dataset is better compared to low resolution dataset. Low resolution dataset contains low reliability annotation proteins. The second main issue in low resolution dataset, signal peptides are not removed from some low resolution transmembrane proteins.

5.4 Performance Comparison with Existing Models

Performance comparison between proposed model *PSOFuzzySVM-TMH* model and existing models at different levels is mentioned below:

The success rates of *PSOFuzzySVM-TMH* model and existing models at per protein level are reported in Table 5. Our proposed model has obtained the highest accuracy 77.61% compared to existing models using low-resolution dataset. In current state of art methods. Arai et al.'s model has yielded the accuracy of 74.83% [54]. Whereas, 73.29% accuracy has been achieved by Lo et al. proposed model *SVMtop* [31]. Similarly, the success rate of proposed model is compared with other existing methods namely: *HMMTOP2, TMHMM2, MEMSAT3, Phobius, PHDhtm v.1.96, Top-Pred2, SOSUI 1.1,* and *SPLIT4*. In contrast, the proposed model still achieved the highest accuracy 79.32% in case of high-resolution dataset. In literature. Lo et al., proposed model *SVMtop* has achieved the accuracy of 72.09% [31]. So, the performance of our proposed model at protein level is 4.32% higher in case of low resolution dataset and 7.23% higher in case of high resolution dataset.

At segment level the performance of proposed model is evaluated using two measures precision and recall. These measures are also shown in Table 5. Our proposed model has achieved the highest recall and precision 97.07% and 97.12%, respectively using low-resolution

dataset. In contrast, the existing model, *SVMtop* has obtained 94.76% recall and 93.94% precision [31]. In case of high-resolution dataset, our proposed model has obtained 97.57% recall and 98.21% precision. The performance of the proposed model is also measured at per residue level. Various performance measures are used at this level such as accuracy, recall, precision and *MCC*. Our proposed model has yielded 93.81% accuracy 91.82% recall, 85.15% precision, and 0.81 *MCC*. On the other hand the highest performance of existing model *SVMtop* are 89.23% 87.50%, 80.35%, and 0.77 accuracy, recall, precision, and *MCC*, respectively using low-resolution dataset [55]. In case of high-resolution dataset, our proposed model has obtained 94.13%, 92.82%, 95.73%, 0.86, accuracy, recall, precision, and *MCC*, respectively, whereas the predicted outcomes of the *SVMtop* model are 90.90%, 87.84%, 84.36%, and 0.81, accuracy, recall, precision, and *MCC*, respectively [31].

After comparison, we have concluded that the performance of our proposed model is outstanding at each level in both datasets. These significant achievements have been ascribed to the fusion of two informative feature representation schemes, the selection of valuable features, and the best classification algorithm.

Conclusion

In this work, we have developed a more powerful, robust and high throughput identification model for *TM* helix segments. In this model, two sample formulation methods namely: *PSSM* and 6-letter exchange group representation are applied to extract numerical features from protein sequences. In order to select high discriminative features, evolutionary feature selection approach *PSO* is applied on both feature spaces. After that the selected feature spaces are combined to form an ensemble feature space. *Fuzzy SVM* is utilized as classification algorithm. The performance of the classifier is assessed through 10-fold cross validation using two

benchmarks datasets. After analyzing, we observed that the performance of *PSOFuzzySVM-THM* model is quite promising and higher than that of existing methods at each level, so far. So, it is ascertained that the proposed model may become a useful and high throughput tool for academia and research community for further structure and functional studies on transmembrane proteins.

References

1. T. Klabunde and G. Hessler, *Chem Bio Chem*, 2002, 3, 928-944.
2. H. Berman, J. Westbrook, Z. Feng, G. Gilliland and T. Bhat, *Nucleic Acids Res.*, 2000, 28, 235-242.
3. P. Argos, J. Rao and P. Hargrave, *Eur. J. Biochem*, 1982, 128, 565-575.
4. M. Cserzo, E. Wallin, I. Simon, G. Von Heijne and A. Elofsson, *Protein Eng. Des. Sel.*, 1997, 10, 673-676.
5. D. Eisenberg, R. M. Weiss and T. C. Terwilliger, *Nature*, 1982, 299, 371-374.
6. D. Juretic, L. Zoranic and D. Zucic, *J. Chem. Inf. Comput. Sci.*, 2002, 42, 620–632.
7. J. Kyte and R. Doolittle, *J. Mol. Biol.*, 1982, 157, 105–132.
8. K. Nakai and M. Kanehisa, *Genomics*, 1992, 14, 897–911.
9. G. Von Heijne, *J. Mol. Biol.*, 1992, 225, 487-494.
10. M. G. Claros and G. Von Heijne, *Comput. Appl. BioSci.*, 1994, 10, 685-686.
11. T. Hirokawa, S. Boon-Chieng and S. Mitaku, *Bioinformatics*, 1998, 14, 378–379.
12. C. Deber, C. Wang, L. Liu, A. Prior, S. Agrawal, B. Muskat and A. Cuticchia, *Protein Sci.*, 2001, 10, 212–219.
13. B. Persson and P. Argos, *Protein Sci.*, 1996, 5, 363–371.
14. B. Rost, R. Casadio, P. Fariselli and C. Sander, *Protein Sci.*, 1995, 4, 521–533.
15. M. Cserzo, F. Eisenhaber, B. Eisenhaber and I. Simon, *Bioinformatics*, 2004, 20, 136–137.
16. D. T. Jones, *Bioinformatics*, 2007, 23, 538-544.
17. B. Rost, P. Fariselli and R. Casadio, *Protein Sci.*, 1996, 5, 1704-1718.
18. G. E. Tusnady and I. Simon, *J. Mol. Biol.*, 1998, 283, 489-506.
19. G. E. Tusnady and I. Simon, *Bioinformatics*, 2001, 17, 849-850.
20. A. Krogh, B. Larsson, G. von Heijne and E. L. Sonnhammer, *J Mol Biol*, 2001, 305, 567-580.
21. E. L. Sonnhammer, G. Von Heijne and A. Krogh, *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, 1998, 6, 175-182.
22. H. Viklund and A. Elofsson, *Protein Sci.*, 2004, 13, 1908-1917.
23. R. Kahsay, G. Gao and L. Liao, *Bioinformatics*, 2005, 21, 1853–1858.
24. L. Kall, A. Krogh and E. Sonnhammer, *Nucl. Acids Res.*, 2007, 35, W429–432.
25. P. Martelli, P. Fariselli and R. Casadio, 2003; 19:, *Bioinformatics*, 2003, 19, i205–211.
26. M. Amico, M. Finelli and I. Rossi, *Nucl. Acids Res*, 2006, 34, W169–172.
27. P. Bagos, T. Liakopoulos and S. Hamodrakas, *BMC Bioinformatics*, 2006, 7, 189.
28. H. Shen and J. J. Chou, *PLoS ONE*, 2008, 3, e2399.
29. T. Nugent and D. Jones, *BMC Bioinformatics*, 2009, 10, 159.
30. T. Nugent and D. Jones, *PLoS Comput. Biol.*, 2009, 6, e1000714.
31. A. Lo, H. S. Chiu, T. Y. Sung, P. C. Lyu and W. L. Hsu, *Journal of Proteome Research*, 2008, 7, 487-496.

32. S. R. Hosseini, M. Sadeghi, H. Pezeshk, C. Eslahchi and M. Habibi, *Comput Biol Chem*, 2008, 406-411.
33. J. Pylouster, A. Bornot, C. Etchebest and A. G. D. Brevern, *Amino Acids*, 2010, 1241-1254.
34. N. Zaki, S. Bouktif and L. M. Sanja, *PLoSONE*, 2011, 6.
35. C. P. Chen, A. Kernytsky and B. Rost, *Protein Sci.*, 2002, 11, 2774-2791.
36. J. M. Cuthbertson, D. A. Doyle and M. S. Sansom, *Protein Eng. Des. Sel.*, 2005, 18, 295-308.
37. L. Kall and E. Sonnhammer, *FEBS Lett.*, 2002, 532, 415-418.
38. K. Melen, A. Krogh and G. von-Heijne, *J. Mol. Biol.*, 2003, 327, 735-744.
39. S. Moller, M. D. Croning and R. Apweiler, *Bioinformatics*, 2001, 646-653, 17.
40. S. Moller, E. V. Kriventseva and R. Apweiler, *Bioinformatics* 2000, 16, 1159-1160.
41. A. Bairoch and R. Apweiler, *J. Mol. Med.* , 1997, 5, 312-316.
42. S. Jayasinghe, K. Hristova and S. H. White, *Protein Sci.*, 2001, 10, 455-458.
43. M. Ikeda, M. Arai, D. M. Lao and T. Shimizu, *In Silico. Biol.*, 2002, 2, 19-33.
44. T. Liu, X. Zheng and J. Wang, *Biochimie*, 2010, 92, 1330-1334.
45. A. A. Schaffer, L. Aravind, T. L. Madden, S. Shavirin and J. L. Spouge, *Nucleic Acids Res*, 2001, 29, 2994–3005.
46. A. Fuchs, A. Kirschner and D. Frishman, *Proteins*, 2009, 74, 857-871.
47. S. K. Golmohammadi, L. Kurgan, B. Crowley and M. Reformat, *Frontiers in the Convergence of Bioscience and Information Technologies*, 2007, 153-158.
48. M. O. Dayhoff, R. M. Schwartz and B. C. Orcutt, *Atlas Protein Sequence Struct.*, 1978, 5, 345-352.
49. J. Kennedy and R. Eberhart, Perth, Western Australia, 1995.
50. M. Hayat, M. Tahir and S. A. Khan, *Journal of Theoretical Biology*, 2013, 346C, 8-15.
51. M. Hayat and A. Khan, *Analytical Biochemistry*, 2012, 424, 35-44.
52. M. Hayat, A. Khan and M. Yeasin, *Amino Acids*, 2012, 42, 2447-2460.
53. M. Hayat and A. Khan, *IET Communications*, 2012, 6, 3257-3264.
54. M. Arai, H. Mitsuke, M. Ikeda, J. X. Xia, T. Kikuchi, M. Satake and T. Shimizu, *Nucleic Acids Res.*, 2004, 32, W390–393.
55. M. Hayat and A. Khan, *Amino Acids*, 2013, 44, 1317-1328.
56. M. Hayat and A. Khan, *Journal of Theoretical Biology*, 2011, 271, 10-17.

List of Figures

Figure 1. Framework of the proposed approach

List of Tables

Table 1 Categorization of amino acids [56]

Table 2. Prediction performance of proposed model using PSSM method based full and condensed features at different levels**.**

Table 3. Prediction performance of proposed model using 6-letter exchange group method based full and condensed features at different levels.

Table 4. Prediction performance of proposed model using hybrid method based full and condensed features at different levels.

Table 5. Performance comparison with existing models

Table 1 Categorization of amino acids [56]

Group	Sub-Group	Amino Acids
Exchange group	e_1	KHR
	e_2	DENQ
	e_3	C
	e_4	AGPST
	e_5	ILMV
	e_6	FYW

Table 2. Prediction performance of proposed model using PSSM method based full and condensed features at different levels

Methods	Per Proteins (%)	Per segments (%)		Per residue (%)			
	Q_{ok}	Q^{obsd}	Q^{prd}	Q_2	Q^{obsd}	Q^{prd}	MCC
Low resolution							
Full Space	67.8	94.3	93.6	88.0	87.2	79.2	0.77
Selected Space	71.3	94.6	95.3	89.5	88.9	81.2	0.78
High resolution							
Full Space	70.1	96.1	95.2	90.9	86.7	91.4	0.82
Selected Space	72.6	97.0	96.7	92.0	88.4	92.6	0.83

Table 3. Prediction performance of proposed model using 6-letter exchange group method based full and condensed features at different levels

Methods	Per Proteins (%)	Per segments (%)			Per residue (%)		
	Q_{ok}	Q^{obsd}	Q^{prd}	Q_2	Q^{obsd}	Q^{prd}	MCC
Low resolution							
Full Space	69.2	94.7	94.1	88.3	87.9	80.2	0.77
Selected Space	72.0	95.2	95.8	89.1	88.3	81.0	0.78
High resolution							
Full Space	70.1	95.2	96.0	90.2	86.9	91.8	0.81
Selected Space	73.9	96.7	97.3	91.9	88.0	92.9	0.82

Table 4. Prediction performance of proposed model using ensemble method based full and condensed features at different levels

Methods	Per Proteins (%)	Per segments (%)			Per residue (%)		
	Q_{ok}	Q^{obsd}	Q^{prd}	Q_2	Q^{obsd}	Q^{prd}	MCC
Low resolution							
Full Space	75.5	95.6	95.7	90.7	89.1	83.4	0.79
Selected Space	77.6	97.0	97.1	93.8	91.8	85.1	0.81
High resolution							
Full Space	77.5	96.9	96.3	92.5	90.3	93.2	0.84
Selected Space	79.3	97.5	98.2	94.6	92.8	95.7	0.86

Table 5. Performance comparison with existing models

	Per Protein (%)		Per segment (%)		Per residue (%)			
	Q_{ok}	Q_{TM}	Q^{obsd}	Q^{prd}	Q_2	Q^{obsd}	Q^{prd}	MCC
Low resolution								
PSOFuzzySVM-TMH	**77.61**	**74.20**	**97.07**	**97.12**	**93.81**	**91.82**	**85.15**	**0.81**
SVMtop	73.29	69.23	94.76	93.94	89.23	87.50	80.35	0.77
TMHMM2	68.53	58.74	90.39	93.52	89.23	82.82	83.03	0.76
HMMTOP2	64.34	55.94	89.96	93.78	87.89	79.36	84.37	0.75
PHDhtm v.1.96	39.86	29.37	76.27	85.76	85.35	81.71	76.59	0.71
MEMSAT3	70.63	67.83	91.56	90.24	87.91	84.54	77.63	0.73
TopPred2	57.34	42.66	86.75	91.13	88.00	76.85	82.90	0.72
SOSUI 1.1	63.64	-	88.36	91.55	87.00	80.41	78.66	0.71
SPLIT4	72.73	64.34	93.45	91.32	88.07	87.56	76.88	0.74
ConPred II	74.83	65.04	94.76	92.21	90.07	84.37	84.13	0.78
Phobius	72.03	60.84	92.87	93.14	88.92	83.92	82.57	0.77
PolyPhobius	71.33	61.54	94.47	91.54	89.75	86.84	83.11	0.79
High resolution								
PSOFuzzySVM-TMH	**79.32**	**75.61**	**97.57**	**98.21**	**94.13**	**92.82**	**95.73**	**0.86**
SVMtop	72.09	62.79	92.78	94.46	90.90	87.84	84.36	0.81
TMHMM2	59.30	46.12	86.93	93.78	87.70	78.59	83.55	0.74
HMMTOP2	65.89	52.71	90.34	89.98	87.68	78.30	82.30	0.73
PHDhtm v.1.96	38.37	25.58	74.43	84.59	84.55	78.28	78.03	0.70
MEMSAT3	64.84	56.64	87.67	91.09	87.16	79.64	78.84	0.71
TopPred2	50.39	37.21	84.50	90.05	86.96	74.06	82.47	0.71
SOSUI 1.1	56.98	-	85.06	92.17	86.15	76.88	80.02	0.71
SPLIT4	65.12	54.65	89.77	91.56	87.12	83.84	78.00	0.73
ConPred II	69.14	55.43	90.94	91.31	88.63	79.99	84.17	0.75
Phobius	67.05	54.65	88.72	93.58	87.81	79.42	83.76	0.75
PolyPhobius	67.44	55.81	90.91	91.28	88.79	82.66	83.34	0.77

Figure

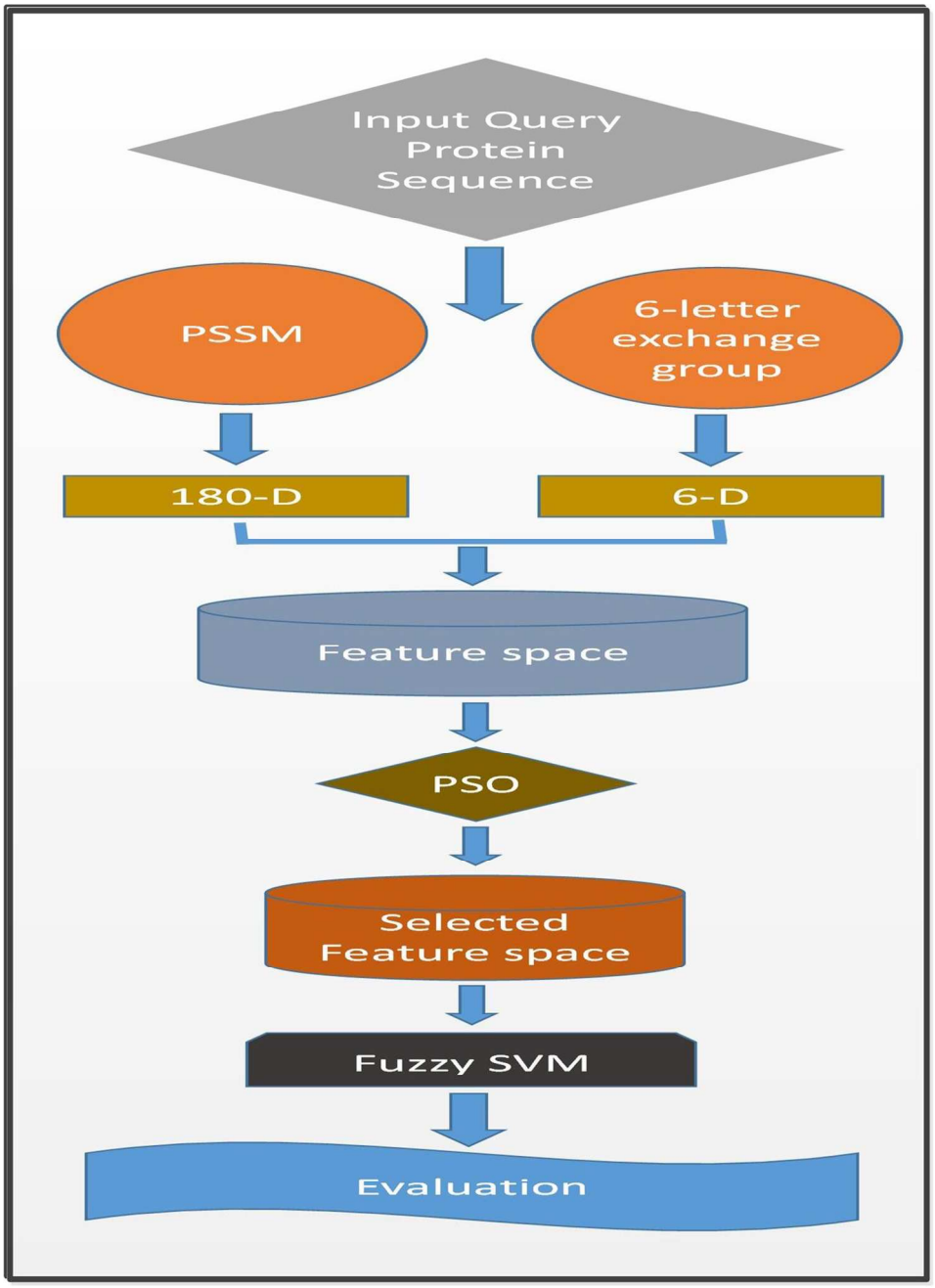

Figure 1. Framework of Proposed Model