

WRF-TMH: predicting transmembrane helix by fusing composition index and physicochemical properties of amino acids

Maqsood Hayat · Asifullah Khan

Received: 10 October 2012 / Accepted: 23 January 2013 / Published online: 14 March 2013
© Springer-Verlag Wien 2013

Abstract Membrane protein is the prime constituent of a cell, which performs a role of mediator between intra and extracellular processes. The prediction of transmembrane (TM) helix and its topology provides essential information regarding the function and structure of membrane proteins. However, prediction of TM helix and its topology is a challenging issue in bioinformatics and computational biology due to experimental complexities and lack of its established structures. Therefore, the location and orientation of TM helix segments are predicted from topogenic sequences. In this regard, we propose WRF-TMH model for effectively predicting TM helix segments. In this model, information is extracted from membrane protein sequences using compositional index and physicochemical properties. The redundant and irrelevant features are eliminated through singular value decomposition. The selected features provided by these feature extraction strategies are then fused to develop a hybrid model. Weighted random forest is adopted as a classification approach. We have used two benchmark datasets including low and high-resolution datasets. tenfold cross validation is employed to assess the performance of WRF-TMH model at different levels including per protein, per segment, and per residue. The success rates of WRF-TMH model are quite promising and are the best reported so far on the same datasets. It is observed that WRF-TMH model might play a substantial role, and will provide essential information for

further structural and functional studies on membrane proteins. The accompanied web predictor is accessible at <http://111.68.99.218/WRF-TMH/>.

Keywords Transmembrane helix · Physicochemical properties · Compositional index · Weighted random forest · Structures of membrane proteins

Introduction

A biological membrane is an anchoring membrane that works as a barrier within or around 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 (Klabunde and Hessler 2002). However, it is complex and difficult to get high-resolution three-dimensional (3D) structures of membrane proteins. Only a few percent of membrane protein structures are available in protein Data Banks (Berman et al. 2000). Membrane protein contains one or more transmembrane (TM) helices, which define the orientation or topology of a membrane protein corresponding to the lipid bilayer. Alpha helical is a prime category of TM proteins, and it performs most of the important biological processes of a cell such as cell signaling, cell-to-cell interaction, cell recognition, and adhesion. Information about TM helix provides some useful clue 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, computational methods are considered as valuable tools for correctly identifying locations of TM helix segments and topology of TM helix proteins.

M. Hayat
Abdul Wali Khan University, Mardan, Pakistan

M. Hayat · A. Khan (✉)
Pattern Recognition Lab, Department of Computer and Information Sciences, Pakistan Institute of Engineering and Applied Sciences, Nilore, 45650 Islamabad, Pakistan
e-mail: asif@pieas.edu.pk

In the past few decades, a series of efforts have been carried out for predicting the topology of TM helix proteins. In the early studies, the analysis was mostly based on the physicochemical properties of amino acids, namely, hydrophobicity (Argos et al. 1982; Cserzo et al. 1997; Eisenberg et al. 1982; Juretic et al. 2002; Kyte and Doolittle 1982; Nakai and Kanehisa 1992; Von Heijne 1992), charge (Claros and Von Heijne 1994; Hirokawa et al. 1998; Juretic et al. 2002), nonpolar phase helicity (Deber et al. 2001), and multiple sequence alignment (Persson and Argos 1996; Rost et al. 1995). DAS-TMfilter (Cserzo et al. 2004), TOP-Pred (Claros and Von Heijne 1994), and SOSUI (Hirokawa et al. 1998) are the most reliable models that give descriptive information about TM helices. Although these methods have efficiently identified TM helix segments, they did not achieve promising results in topology prediction of TM helix proteins. Researchers have used various statistical models such as Hidden Markov Models (HMM) as well as neural networks and support vector machine (SVM) for predicting TM helix. In this regard, several user-friendly web predictors have also been developed for the benefit of academics and researchers. A few of them include TopPred (Claros and Von Heijne 1994), MEMSAT (Jones 2007), PHD (Rost et al. 1996), HMMTOP (Tusnady and Simon 1998, 2001), TMMHMM (Krogh et al. 2001; Sonnhammer et al. 1998), PRODIV_TMMHMM (Viklund and Elofsson 2004), TMMOD (Kahsay et al. 2005), Phobius (Kall et al. 2007), ENSEMBLE (Martelli et al. 2003), PONGO (Amico et al. 2006), HMM-TM (Bagos et al. 2006), MemBrain (Shen and Chou 2008), MEMSAT-SVM (Nugent and Jones 2009a), MEMPACK (Nugent and Jones 2009b), and SVMtop (Lo et al. 2008). The main problem regarding HMM-based methods is that they are computationally expensive. In addition, they use multiple sequence alignments. Additionally, the HMM-based methods fail if TM helix segments are shorter than 16 residues or longer than 35 residues (Shen and Chou 2008). Some researchers have used accuracy along with sensitivity and specificity for evaluation of their proposed methods (Hosseini et al. 2008; Pylouster et al. 2010; Shen and Chou 2008; Zaki et al. 2011a). In addition, several studies have concentrated only on sensitivity and reliabilities of different methods rather than accuracy (Chen et al. 2002; Cuthbertson et al. 2005; Kall and Sonnhammer 2002; Melen et al. 2003; Moller et al. 2001).

In this study, we focus on developing a more effective and accurate TM helix segment prediction system, denoted as WRF-TMH. The proposed approach is based on two kinds of information. The first information is related to the compositional index, whereas the second information is related to the physicochemical properties of amino acids. Compositional index profile is generated by calculating the compositional index of each amino acid in TM helix and

non-TM helix segments. The probability of each amino acid is then calculated in both segments. The physicochemical properties of amino acids such as charge, polarity, aromaticity, size, and electronic are used for exploring the behavior of amino acid sequences. The extraneous information is eliminated through singular value decomposition (SVD) whereby it tries to find such a matrix that has least possible information with strong patterns and trends. Then, highly discriminative features of both the features spaces are combined to form a Hybrid feature space. Weighted random forest (RF) is utilized as a classifier in our proposed system. Weighted RF is an ensemble classifier, where the prediction is made on majority voting; hence, the probability of error becomes less. Further, TM helix segments are less in strength compared with non-TM helix segments. In case of imbalanced data, the classifier will usually be biased towards the majority class. In order to control the bias, each class is assigned different weight, where majority class is assigned low weight and minority class is assigned high weight (Bush et al. 2008). Two standard datasets and tenfold cross validation are used to assess the performance of the proposed WRF-TMH model. The advantage of our proposed approach is that it uses an overlap of 11 residues, while the existing techniques mostly use an overlap of 9. The remaining paper is organized as follows: first, a description of “Materials and methods”; next, an explanation of “the proposed system”; next, a presentation of “performance measures”; and results and discussion and finally a conclusion is drawn.

Materials and methods

Datasets

In order to develop a high quality and reliable prediction model, one needs to construct or select a benchmark dataset according to the problem. Due to the availability of a standard dataset, the learning capability of the model is boosted and thus, the predictions are generally in accordance with the desired output. For this purpose, the dataset must have unbiased homology and less redundancy. Therefore, we used two benchmark datasets. DT1 is a low-resolution membrane protein dataset, which was developed by Moller et al. (2000). It is annotated from SWISS-PROT release 49.0 (Bairoch and Apweiler 1997). Initially, it contained 145 protein sequences, but later two protein sequences were discarded, which had no annotation with membrane proteins. Finally, DT1 consists of 143 protein sequences, which include 687 TMH segments.

DT2 is a high-resolution membrane protein dataset. In this dataset, 101 protein sequences of 3-D structure helix are selected from MPtopo database (Jayasinghe et al.

2001a), while 231 protein sequences are obtained from *TMPDB* database (Ikeda et al. 2002). After merging both the datasets, 30 % CD-hit has been applied to reduce the redundancy. After this screening, *DT2* contains 258 single and multispinning TM protein sequences, which consist of 1,232 TMH segments.

Feature extraction techniques

In this study, we have considered two protein sequence representation methods for extracting pertinent and useful information from the TM protein sequences.

Physicochemical properties

A protein sequence is composed of amino acids. Each amino acid has varying side chain. Amino acids are categorized into different groups according to their nature. Physicochemical properties play vital role in recognizing the behavior of amino acids. In order to extract informative features from protein sequences, we have used some important physicochemical properties of amino acids including charge, polarity, aromaticity, size, and electronic. Each physicochemical property has further sub-types, which differentiate amino acids from each other as shown in Table 1. Physicochemical properties of amino acids perform a significant role in formatting and folding of proteins structure and are largely based on propensity of a

side chain of amino acids. Each property of amino acid has its specific characteristics, which are typically defined by the type of the side chain the amino acids possess. For instance, polar and charged amino acids cover the surface of molecules and are in contact with solvents due to their ability to form hydrogen bonds. Mostly, they interact with each other, for example, positively and negatively charged amino acids form salt bridges, whereas polar amino acid side chains form hydrogen bonds. These interactions are often useful for the stabilization of protein's 3D structures. Polar amino acids are hydrophilic, whereas non-polar amino acids are hydrophobic, which are used to twist protein into useful shapes (Hayat and Khan 2012). In this study, the TM protein sequences are replicated into five sequences and then each amino acid is replaced with its corresponding property. For example, residue r_i at position i can be represented as

$$r_i = (C_i P_i A_i S_i E_i) \quad (1)$$

where C_i , P_i , A_i , S_i , and E_i represent charge, polarity, aromaticity, size, and electronic, respectively. Each amino acid is replaced with its corresponding value, for example, in case of charge the amino acid sequence is replaced by three values: positive, negative, and neutral. Thus, by applying the sliding window, three features are calculated against each position (one residue at a time) and then the window moves to the next position. This process is repeated up to the last residue of a protein sequence. The same procedure is applied for each property of amino acids. Consequently, 16 features are extracted against each position. The feature vector can be expressed as

$$R_i = [C_{ij}]_{1 \times 16} \quad (2)$$

where C_{ij} is the occurrence frequency of property j in window i . Finally, the obtained feature matrix is

$$P = [R_1^T R_2^T \dots R_{L-l+1}^T]_{16 \times L-l+1} \quad (3)$$

where T represents transpose, L is the size of protein sequence, and l is the window size.

Compositional index

Compositional index shows the occurrence frequency of amino acids in protein sequences. High frequency indicates the more existence of that amino acid in protein sequences. To compute compositional indices of amino acids, first, we separated TM and non-TM helix segments from each TM protein sequence. TM helix segments are represented by T1 and non-TM helix segments by T2. Then, the occurrence frequency of each amino acid in T1 and T2, respectively, is computed. The compositional index p_i for each amino acid is calculated as.

Table 1 Physicochemical properties of amino acids

Attributes		Amino acids
Charge	Positive	H, K, R
	Negative	D, E
	Neutral	A, C, F, G, I, L, M, N, P, Q, S, T, V, W
Polarity	Polar	Y
	Nonpolar	C, D, E, H, K, N, Q, R, S, T, Y
Aromaticity	Aliphatic	A, F, G, I, L, M, P, V, W
	Aromatic	I, L, V
	Neutral	F, H, W, Y
Size	Small	A, C, D, E, G, K, M, N, P, Q, R, S, T
	Medium	A, G, P, S
	Large	D, N, T
Electronic	Strong	C, E, F, H, I, K, L, M, Q, R, V, W, Y
	Donor	A, D, E, P
	Weak donor	I, L, V
	Neutral	C, G, H, S, W
	Strong acceptor	F, M, Q, T, Y
	Weak acceptor	K, N, R

$$p_i = -\ln\left(\frac{f_i^{\text{non-TM}} - f_i^{\text{TM}}}{f_i^{\text{TM}}}\right) \quad (4)$$

where $f_i^{\text{non-TM}}$ and f_i^{TM} indicate occurrence frequencies of amino acid i in T1 and T2 datasets, respectively. Subsequently, each amino acid in the sequence is substituted by the corresponding index value. Recently, Compositional index was effectively exploited for TM helix prediction (Zaki et al. 2011a, b). Zaki et al. improved the concept of DomCut method (Suyama and Ohara 2003) by incorporating the amino acid composition knowledge. The compositional index for a TM protein sequence p , with window size w can be computed as

$$m_j^w = \begin{cases} \frac{\sum_{i=1}^{j+(w-1)/2} p_i}{j+(w-1)/2} & 1 \leq j \leq (w-1)/2 \\ \frac{\sum_{i=j-(w-1)/2}^{j+(w-1)/2} p_i}{w} & (w-1)/2 < j \leq L - (w-1)/2 \\ \frac{\sum_{i=j-(w-1)/2}^L p_i}{L-j+1+(w-1)/2} & L - (w-1)/2 < j \leq L \end{cases} \quad (5)$$

We choose window size $w = 7-25$, considering odd size only. The extracted feature vector is thus of size 10-D.

Singular value decomposition

Singular value decomposition (SVD) is a dimensionality reduction technique that plays a vital role in many multivariate data analyses. Using SVD, one can find a reduced dimensional matrix, which has strong correlation with no noise effect. SVD recreates the best possible matrix, which has minimum possible information and emphases, strong patterns, and trends. SVD first transforms correlated variables into uncorrelated variables, which exposes the relationship between the original data and then identifies and orders the dimensions along with exhibition of the most variation in data points. Once highest variation is identified, then it is possible to find the best approximation of original data points in the form of fewer dimensions.

Usually, feature space contains redundant, irrelevant, and mutually dependent information. Therefore, it is needed to transform the feature vector into an orthogonal dimensional space. SVD exposes the matrix or linear transformation in minimum number of dimensions. For instance, if a feature space is of N -dimension laying in a K -dimensional subspace, where $K < N$, then each N -dimensional vector has only K degree of freedom and can be uniquely represented by K number of dimensions. SVD divides the matrix A of size $M \times N$ into three matrices U , W , and V as $A = UWV^T$ when $M > N$. U is an $M \times M$ orthogonal matrix ($UUT = I$) that indicates the left singular vector of A , V is an $N \times N$ orthogonal matrix ($VVT = I$) that represents the right singular vector of A ,

and W is an $M \times N$ diagonal matrix having the singular values of A . If we assume that $M < N$, then linear transformation can be represented by SVD as follows:

$$\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1N} \\ \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \dots & \cdot \\ a_{M1} & a_{M2} & \dots & a_{MN} \end{pmatrix} = \begin{pmatrix} u_{11} & u_{12} & \dots & u_{1M} \\ \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \dots & \cdot \\ u_{M1} & u_{M2} & \dots & u_{MM} \end{pmatrix} \begin{pmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1N} \\ \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \dots & \cdot \\ \sigma_{M1} & \sigma_{M2} & \dots & \sigma_{MN} \end{pmatrix} \times \begin{pmatrix} v_{11} & v_{12} & \dots & v_{1N} \\ \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \dots & \cdot \\ v_{N1} & v_{N2} & \dots & v_{NN} \end{pmatrix} \quad (6)$$

Subsequently, the value of $\sigma_{i,j}$ is zero when $j > M$. So, the product WV^T will produce zero value for rows $M+1$ through N .

$$\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1N} \\ \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \dots & \cdot \\ a_{M1} & a_{M2} & \dots & a_{MN} \end{pmatrix} = \begin{pmatrix} u_{11} & u_{12} & \dots & u_{1M} \\ \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \dots & \cdot \\ u_{M1} & u_{M2} & \dots & u_{MM} \end{pmatrix} \begin{pmatrix} \sigma_1 & 0 & \dots & 0 \\ \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \dots & \cdot \\ 0 & 0 & \dots & \sigma_M \end{pmatrix} \times \begin{pmatrix} v_{11} & v_{12} & \dots & v_{1N} \\ \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \dots & \cdot \\ v_{M1} & v_{M2} & \dots & v_{MN} \end{pmatrix} \quad (7)$$

This indicates that a column A_i of A can be expressed as a linear combination of M vectors in $U(u_{1i}, u_{2i}, \dots, u_{Mi})$ using the singular values in $W(\sigma_1, \sigma_2, \dots, \sigma_M)$ and i^{th} column V_i^T in V^T . The diagonal nonnegative values of W can be ordered such that $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_M$. If some entries on the diagonal of W are zero, then for some K , $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_K \geq \sigma_{K+1} = \dots = \sigma_M = 0$. So, the number of columns in U and the number of rows in V^T can be reduced to K dimensions. However, both the number of rows and columns of W can be reduced to K dimensions.

The rank shows non-zero singular values of matrix A . The required matrix is obtained by multiplying the first K columns of U matrix by first K singular values from W matrix and first K rows of V^T matrix as shown in Fig. 1.

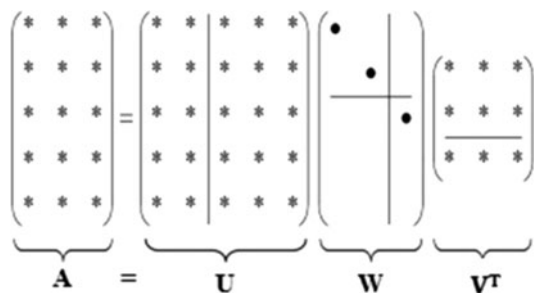


Fig. 1 Graphical representation of SVD

$$\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1N} \\ \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \dots & \cdot \\ a_{M1} & a_{M2} & \dots & a_{MN} \end{pmatrix} = \begin{pmatrix} u_{11} & u_{12} & \dots & u_{1K} \\ \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \dots & \cdot \\ u_{M1} & u_{M2} & \dots & u_{MK} \end{pmatrix} \begin{pmatrix} \sigma_1 & 0 & \dots & 0 \\ \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \dots & \cdot \\ 0 & 0 & \dots & \sigma_K \end{pmatrix} \times \begin{pmatrix} v_{11} & v_{12} & \dots & v_{1N} \\ \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \dots & \cdot \\ v_{K1} & v_{K2} & \dots & v_{KN} \end{pmatrix} \tag{8}$$

Consequently, Matrix A of size $M \times N$ can be equally represented by $K \times N$ matrix WV^T where $K < M$.

Non-zero singular values, which are close to zero are eliminated from the matrix because $\sigma_{K+1} \dots \geq \sigma_M$ show distance from the subspace spanned by U_1, \dots, U_K . A very small distance may not affect the operation that will be performed on the reduced data. In this study, we have picked the first five top dimensions, where 83 % variance is found among these dimensions.

Proposed prediction system for *TM* helices (WRF-TMH)

In this study, we propose an affective model WRF-TMH for the prediction of *TM* helices. The proposed model is based on two informative protein sequence representation methods: physicochemical properties and compositional index. In physicochemical properties based feature extraction, each residue in a protein sequence is first substituted according to the behavior of amino acid by the corresponding value of physicochemical property. After that, the frequency of each value is computed in the specified peptide. The process is repeated to the last residue of a protein sequence. Consequently, 16 features are extracted against each position in a protein sequence.

On the other hand, using compositional index, first we have calculated the occurrence frequency of each amino

acid in *TM* and non-*TM* helix segments. Next, the compositional index against each amino acid is computed. Then, each amino acid in a protein sequence is substituted by its corresponding compositional index. Finally, taking window size of odd number from 7 to 25, as a result, 10 features are extracted against each position. In order to eradicate the redundancy and irrelevant features, we have employed SVD on each feature space separately. Five features with high variations are selected from each feature space and then combined with the selected features of both the feature spaces, to enhance the discriminative power of the feature space. In addition, Weighted RF is used as a learner, which is the collection of tree hypotheses whereby each tree grows with respect to a different bootstrap sample with the same distribution (Afridi et al. 2012; Hayat et al. 2012). The output of final prediction is made on using majority voting; hence, the chances of error are minimized. Recently, RF has been successfully utilized to a wide range of classification problems, especially for predicting protein–protein binding sites (Bordner 2009), residue–residue contact, and helix–helix interaction (Wang et al. 2011) as well as solvent accessible surface area of *TM* helix residues (Wang et al. 2012) in membrane proteins.

In this study, the number of *TM* helix segments is less than the number of non-*TM* helix segments. Usually, it is perceived that in such situation the prediction of learner is often biased towards the majority class, whereas the purpose of our proposed approach is to predict *TM* helix segments more accurately. For this purpose, weight is assigned to each class, whereby high weight is assigned to minority class and low weight is assigned to majority class (Bush et al. 2008). The framework of the proposed approach is illustrated in Fig. 2.

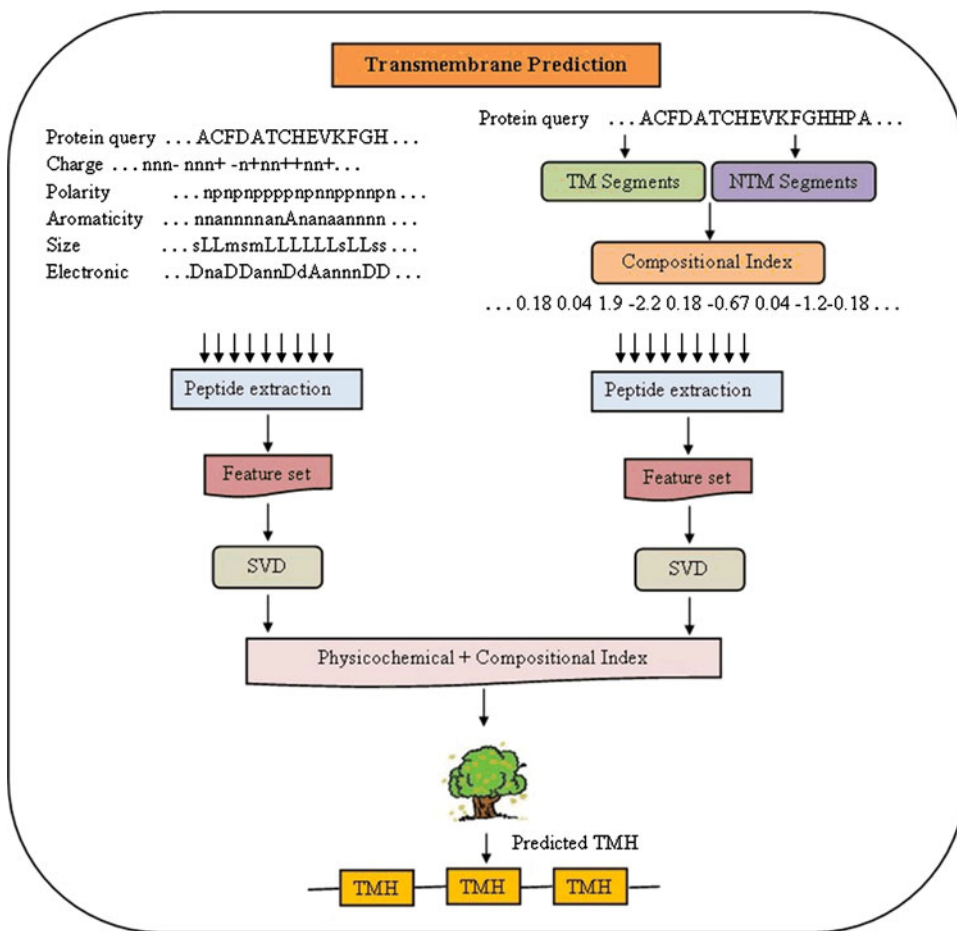
User-friendly web predictor

In order to provide an easy way to access and utilize the developed resource for the prediction of *TM* helix, we have lunched a user-friendly web predictor “WRF-TMH predictor”. This predictor uses a simple format of text and displays the start and end location of each helix along with coloring the residue of each helix in sequence. The main page of WRT-TMH predictor is shown in Fig. 3a, whereas the predicted page is shown in Fig. 3b.

Performance measures

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

Fig. 2 Framework of the proposed approach



$$Q_{htm}^{\%obsd} = \frac{\text{number of correctly predicted TM in dataset}}{\text{Total number of TM in dataset}} \times 100 \tag{9}$$

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

$$Q_{htm}^{\%prd} = \frac{\text{number of correctly predicted TM in dataset}}{\text{number of TM predicted in dataset}} \times 100 \tag{10}$$

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

$$Q_{ok} = \frac{\sum_i^{N_{Prot}} \delta_i}{N_{Prot}} \times 100 \delta_i = \begin{cases} 1, & \text{if } Q_{htm}^{\%obsd} \wedge Q_{htm}^{\%prd} = 100 \text{ for protein } i \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

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

$$Q_2 = \frac{\sum_i^{N_{Prot}} \text{number of residues predicted correctly in protein } i}{N_{Prot}} \times 100 \tag{12}$$

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

$$Q_{2T}^{\%obsd} = \frac{\text{number of residues correctly predicted in TM helices}}{\text{number of residues observed in TM helices}} \times 100 \tag{13}$$

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

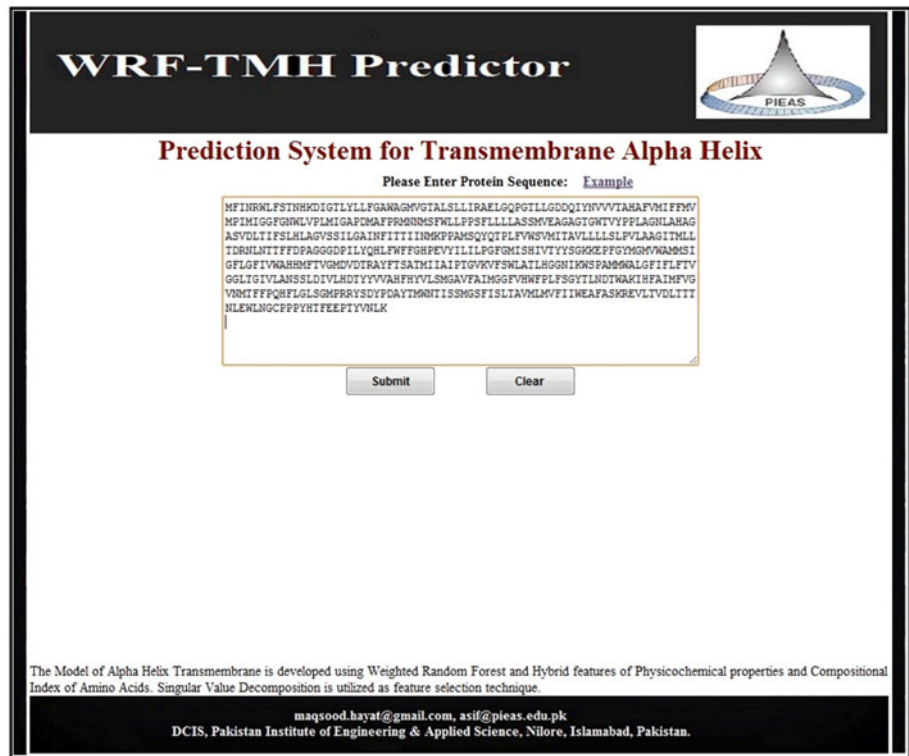
$$Q_{2T}^{\%prd} = \frac{\text{number of residues correctly predicted in TM helices}}{\text{number of residues predicted in TM helices}} \times 100 \tag{14}$$

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

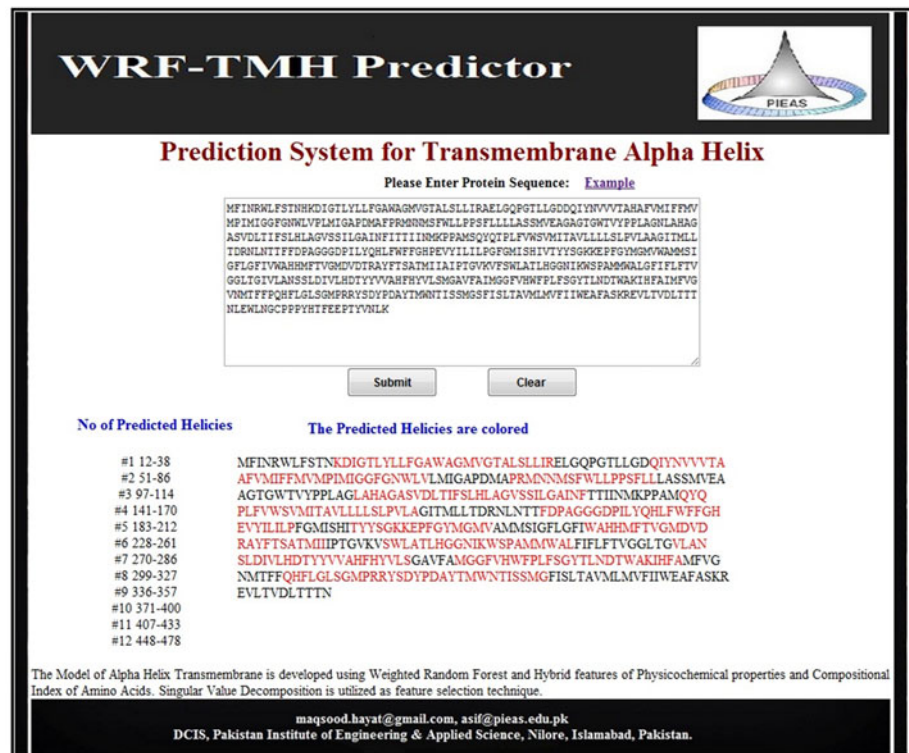
$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{15}$$

MCC is a Mathew correlation coefficient, where the value of MCC is in the range of -1 and 1 . In Eq.15, 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,

Fig. 3 a Illustrates the Main page of WRF-TMH Predictor
b Shows the output of predictor



(a)



(b)

and FN is the number of incorrectly predicted non-TM helix residues.

Results and discussion

Generally, three cross-validation tests including self-consistency, jackknife, and independent dataset are used for evaluating the performance of prediction model. Among these three cross-validation tests, investigators have extensively applied jackknife test due to its special attributes (Khan et al. 2010; Naveed and Khan 2012); however, it is computationally expensive. In order to reduce the computational cost along with considering the important characteristics of jackknife test, we have adopted tenfold cross validation. Jackknife test splits the training dataset into n -fold, while tenfold cross validation randomly partitioned the training dataset into ten approximately equal mutually exclusive folds. In both tests, one fold is used for testing and the remaining folds are used for training. The whole process is repeated ten times where each fold takes place exactly once as testing fold. Finally, the prediction of each fold is averaged to find out the final output. Two benchmark datasets, low-resolution and high-resolution, are used, whereas the performance is assessed on three different levels (per protein, per segment, and per residue).

Performance analysis between selected feature space and full feature space

After identifying patterns and extracting the information from protein sequences, it is possible that sometimes the information contains redundancy and noise, which becomes the cause that disgraces the performance of classification algorithms. In this study, we have performed comparative analysis of selected feature space and full feature space. The full and selected feature spaces are shown in Fig. 4, where black color shows full feature space and red color illustrates the selected feature space.

Performance of weighted RF using full feature space

In this work, first, we have examined the performance of weighted RF in conjunction with full space of individual and hybrid features. Success rates of weighted RF are reported in Table 2. Weighted RF in conjunction with physicochemical properties obtained 66.4 % accuracy at per protein level for low resolution dataset. At per segment level, weighted RF achieved 93.7 % precision and 92.9 % recall. Whereas, in term of per residue, the predicted results of weighted RF are 86.2 % accuracy, 85.1 % precision, 77.6 % recall, and 0.75 MCC. The performance of weighted RF using compositional index is better than

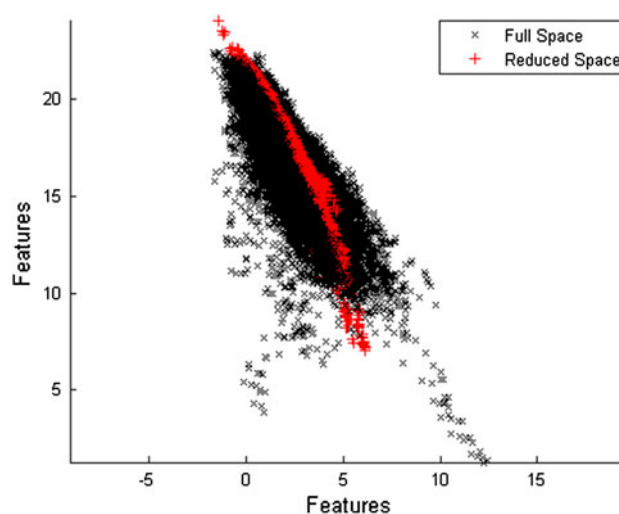


Fig. 4 Full feature space and selected feature space using SVD

physicochemical properties at per protein level; however, it is almost similar at per-segment and per-residue levels. In case of hybrid feature space, weighted RF obtained enhanced results compared to individual feature spaces. Success rates of weighted RF using hybrid space are, 72.7 % accuracy at per protein level, 94.2 % precision, and 94.1 % recall at per segment level, and 88.4 % accuracy, 87.0 % precision, 80.0 % recall, and 0.77 MCC at per residue level. Using high-resolution dataset, the predicted outcome of weighted RF is 69.0 % accuracy in case of per protein level. It achieved 90.3 % precision and 94.0 % recall at segment level while 89.7 % accuracy, 85.4 % precision, 90.6 % recall, and 0.81 MCC at per residue level. In contrast, the performance of weighted RF using compositional index-based features is not so good compared with using physicochemical properties. On utilizing the hybrid feature space the performance of weighted RF is better than that of individual feature spaces. It obtained 71.7 % accuracy at per protein level, whereas 91.2 % precision and 93.9 % recall at per segment level, and 90.3 % accuracy, 88.1 % precision, 91.8 % recall, and 0.83 MCC at per residue level.

Performance of weighted RF using selected feature space

After applying SVD the performance of weighted RF is boosted on each feature space for both the datasets as reported in Table 3. In case of low-resolution dataset, the performance of weighted RF using individual feature spaces is comparable at each level. In contrast, the performance of weighted RF is enhanced using hybrid feature space, which is 76.9 % accuracy in terms of protein level, whereas it is 96.1 % precision and 95.1 % recall at per

Table 2 Success rates of WRF-TMH at different levels using individual and Hybrid feature space

Methods	Per protein (%)	Per segment (%)		Per residue (%)			
	Q_{ok}	Q^{obsd}	Q^{prd}	Q_2	Q^{obsd}	Q^{prd}	MCC
Low resolution							
Physicochemical properties	66.4	93.7	92.9	86.2	85.1	77.6	0.75
Compositional index	67.8	92.3	93.1	85.6	84.7	78.4	0.75
Hybrid features	72.7	94.2	94.1	88.4	87.0	80.0	0.77
High resolution							
Physicochemical properties	69.0	90.3	94.0	89.7	85.4	90.6	0.81
Compositional index	67.8	89.9	93.4	88.1	86.4	90.2	0.80
Hybrid features	71.7	91.2	93.9	90.3	88.1	91.8	0.83

Table 3 Successes rates of WRF-TMH at different levels after applying SVD using individual and Hybrid feature space

Methods	Per protein (%)	Per segment (%)		Per residue (%)			
	Q_{ok}	Q^{obsd}	Q^{prd}	Q_2	Q^{obsd}	Q^{prd}	MCC
Low resolution							
Physicochemical properties	70.6	94.1	93.3	87.5	85.9	79.2	0.76
Compositional index	71.3	94.8	93.8	87.1	86.3	79.0	0.76
Hybrid features	76.9	96.1	95.1	90.8	87.8	81.1	0.78
High-resolution							
Physicochemical properties	70.1	91.1	94.9	90.4	86.9	91.1	0.82
Compositional index	69.7	90.6	94.5	91.1	86.7	90.9	0.81
Hybrid features	74.0	93.3	95.5	92.1	89.3	93.3	0.84

segment level and 90.8 % accuracy, 87.8 % precision, 81.1 % recall, and 0.78 MCC at per residue level.

In case of high-resolution dataset, the performance of weighted RF is better using physicochemical-based feature space than that of compositional index-based feature space. At per protein level, weighted RF obtained 71.1 % accuracy using physicochemical properties, while 69.7 % accuracy using compositional index. Likewise, low-resolution dataset, the performance of weighted RF is further improved using hybrid feature space compared with individual feature spaces. The proposed model achieved 74.0 % accuracy at protein level, 93.3 % precision and 95.5 % recall at segment level and 92.1 % accuracy, 89.3 % precision, 93.3 % recall, and 0.84 MCC at per residue level.

Empirical results revealed that the performance of weighted RF in conjunction with hybrid feature space is promising in case of both the datasets. Hybrid feature space is the fusion of two feature spaces, which compensates the weaknesses of each other. In addition, feature selection technique SVD further improved the performance of weighted RF because it has selected only the high varied features from the feature space.

Performance comparison with existing approaches

Performance comparison of the proposed approach WRF-TMH model with existing approaches at different levels is listed below.

Performance analysis at protein level

The predicted outcomes of WRF-TMH model at per protein level along with already published methods are listed in Table 4. WRF-TMH model has achieved 76.92 % accuracy using low-resolution dataset. In the existing approaches, Arai et al.'s model has obtained the highest accuracy 74.83 % (Arai et al. 2004), whereas Lo et al.'s (2008) developed model *SVMtop* has obtained the accuracy of 73.29 %. In addition, the performance of WRF-TMH model is also compared with the other published methods including *HMMTOP2*, *TMHMM2*, *MEMSAT3*, *Phobius*, *PHDhtm v.1.96*, *Top-Pred2*, *SOSUI 1.1*, and *SPLIT4*. The performance of WRF-TMH model is 3.63 and 2.09 % higher than *SVMtop* and *ConPred-II*, respectively. On the other hand, using high-resolution dataset, WRF-TMH model has yielded 74 % accuracy. In current state-of-the-

Table 4 Performance comparison with existing approaches

	Per protein (%)		Per segment (%)		Per residue (%)			MCC
	Q_{ok}	Q_{TM}	Q^{obsd}	Q^{prd}	Q_2	Q^{obsd}	Q^{prd}	
Low resolution								
WRF-TMH	76.92	72.00	96.06	95.10	90.84	87.81	81.11	0.78
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	82.9	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								
WRF-TMH	74.00	68.6	93.26	95.45	92.13	89.27	93.33	0.84
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

art methods, *SVMtop* has obtained the highest accuracy of 72.09 % (Lo et al. 2008), while *ConPred-II* has obtained 69.14 % of accuracy (Arai et al. 2004). The success rate of WRF-TMH model is 1.91 and 4.86 % higher than that of *SVMtop* and *ConPred-II*, and is more advanced than other existing methods.

Performance analysis at segment level

At per segment level, the performance of WRF-TMH model is measured as recall and precision of the TM helix segments. The recall and precision of WRF-TMH model and other existing methods are shown in column 3–4 of Table 4. The performance of WRF-TMH model is also substantially good at per segment level compared with the already published methods. In case of low-resolution dataset, WRF-TMH model has obtained 96.06 % of recall and 95.10 % of precision. On the other hand, *SVMtop* has achieved 94.76 % recall and 93.94 % precision (Lo et al. 2008). In other existing methods, several methods have

yielded comparable recall but worse precision and vice versa. Besides, the recall and precision of *ConPred-II* are relatively better compared with other state-of-the-art methods (Arai et al. 2004). Whereas WRF-TMH model has achieved 1.84 and 1.16 % higher results than that of *SVMtop* and 1.84 and 2.89 % higher than that of *ConPred-II*, using high-resolution dataset, our proposed approach has achieved 93.26 % recall and 95.45 % precision. In contrast, *SVMtop* has yielded 92.78 % recall and 94.46 % precision.

Performance analysis at residue level

The performance of WRF-TMH model is also analyzed at per residue level. In per protein and per segment levels only TM helix segments are considered; however, in per residue level both TM and non-TM helix segments are measured. At per residue level, the performance of the WRF-TMH model is assessed using four measures such as accuracy, recall, precision, and MCC. The predicted

outcomes of the WRF-TMH model are 90.84, 87.81, 81.11, and 0.78 accuracy, recall, precision, and *MCC*, respectively, using low-resolution dataset. In existing literature, *SVMtop* provides 89.23, 87.50, 80.35, and 0.77 accuracy, recall, precision, and *MCC*, (Lo et al. 2008). In addition, Krogh et al.'s (2001) proposed method *TMHMM2* has achieved 89.23 % accuracy, 82.82 % recall, 83.03 % precision, and 0.76 *MCC*. Arai et al.'s (2004) proposed model *ConPred-II* has obtained 90.07, 84.37, 84.13, and 0.78 %, accuracy, recall, precision, and *MCC*, respectively. On the other hand, using high-resolution dataset, the predicted outcomes of WRF-TMH model are 92.13, 89.27, 93.33, and 0.84 % accuracy, recall, precision, and *MCC*, respectively, whereas, in existing methods the highest success rates have been achieved by *SVMtop*, which are 90.90, 87.84, 84.36, and 0.81 % accuracy, recall, precision, and *MCC*, respectively. Likewise, PolyPhobius has yielded 88.79 % accuracy and 0.77 *MCC*, whereas *SPLIT4* has obtained 83.84 % recall and *ConPred II* has 84.17 % precision. However, in former methods, three residue long helix segments were considered (Jayasinghe et al. 2001a), but later, it was increased to nine residue long helix segment (Jayasinghe et al. 2001b). Likewise, Moller et al. (2000) also considered nine residues long segment in his proposed model, but our proposed model has considered 11 residue-long segment.

Finally, we have concluded that our proposed method has obtained remarkable outcomes at all the three levels. The significance of our proposed approach over existing methods is that it has not only obtained the highest accuracy but also increased the length of overlap segments. These attainments have been conceivable due to merging of two informative protein sequence representation methods and ensemble classifier, i.e. weighted RF.

Conclusion

Owing to the dynamic role of TM helix in living organisms, it is indispensable to develop an accurate, effective, and high-quality prediction model for predicting TM helix. For this purpose, we propose a prediction model WRF-TMH, which has shown superior performance compared with the existing approaches. The proposed model is based on two different types of feature extraction schemes: compositional index and physicochemical properties. In order to avoid training of the model from unnecessary and irrelevant features, SVD is applied. Weighted RF is utilized to handle the problem of bias by assigning different weights to different classes. The performance of the classifier is evaluated through tenfold cross validation using two benchmark datasets. The predicted results of the WRF-

TMH model are higher than that of existing methods at each level, so far. So, it is anticipated that our proposed method might play a significant role and provide vital information for further structural and functional studies on membrane proteins.

Acknowledgments This work was supported by the Higher Education Commission of Pakistan under the indigenous PhD scholarship program 17-5-3 (Eg3-045)/HEC/Sch/2006.

References

- Afridi TH, Khan A, Lee YS (2012) Mito-GSAAC: mitochondria prediction using genetic ensemble classifier and split amino acid composition. *Amino Acids* 42:1443–1453
- Amico M, Finelli M, Rossi I (2006) PONGO: a web server for multiple predictions of all-alpha transmembrane proteins. *Nucleic Acids Res* 34:W169–W172
- Arai M, Mitsuke H, Ikeda M, Xia JX, Kikuchi T, Satake M, Shimizu T (2004) Con Pred II: a consensus prediction method for obtaining transmembrane topology models with high reliability. *Nucleic Acids Res* 32:W390–W393
- Argos P, Rao J, Hargrave P (1982) Structural prediction of membrane bound proteins. *Eur J Biochem* 128:565–575
- Bagos P, Liakopoulos T, Hamodrakas S (2006) Algorithms for incorporating prior topological information in HMMs: application to transmembrane proteins. *BMC Bioinform* 7:189
- Bairoch A, Apweiler R (1997) The SWISS-PROT protein sequence database: its relevance to human molecular medical research. *J Mol Med* 5:312–316
- Berman H, Westbrook J, Feng Z, Gilliland G, Bhat T (2000) *Nucleic Acids Res* 28:235–242
- Bordner A (2009) Predicting protein–protein binding sites in membrane proteins. *BMC Bioinform* 24(10):312
- Bush WS, Edwards TS, Dudek SM, McKinney BA, Ritchie MD (2008) Alternative contingency table measures improve the power and detection of multifactor dimensionality reduction. *BMC Bioinform* 9:238–254
- Chen CP, Kernysky A, Rost B (2002) Transmembrane helix predictions revisited. *Protein Sci* 11:2774–2791
- Claros MG, Von Heijne G (1994) TopPred II: an improved software for membrane protein structure predictions. *Comput Appl Biosci* 10:685–686
- Cserzo M, Wallin E, Simon I, Von Heijne G, Elofsson A (1997) Prediction of transmembrane alpha-helices in prokaryotic membrane proteins: the dense alignment surface method. *Protein Eng Des Sel* 10:673–676
- Cserzo M, Eisenhaber F, Eisenhaber B, Simon I (2004) TM or not TM: transmembrane protein prediction with low false positive rate using DASTMfilter. *Bioinformatics* 20:136–137
- Cuthbertson JM, Doyle DA, Sansom MS (2005) Transmembrane helix prediction: a comparative evaluation and analysis. *Protein Eng Des Sel* 18:295–308
- Deber C, Wang C, Liu L, Prior A, Agrawal S, Muskat B, Cuticchia A (2001) TM finder: a prediction program for transmembrane protein segments using a combination of hydrophobicity and nonpolar phase helicity scales. *Protein Sci* 10:212–219
- Eisenberg D, Weiss RM, Terwilliger TC (1982) The helical hydrophobic moment: a measure of the amphipathicity of a helix. *Nature* 299:371–374
- Hayat M, Khan A (2012) Mem-PHYbrid: hybrid features based prediction system for classifying membrane protein types. *Anal Biochem* 424:35–44

- Hayat M, Khan A, Yeasin M (2012) Prediction of membrane proteins using split amino acid composition and ensemble classification. *Amino Acids* 42:2447–2460
- Hirokawa T, Boon-Chiang S, Mitaku S (1998) SOSUI: classification and secondary structure prediction system for membrane proteins. *Bioinformatics* 14:378–379
- Hosseini SR, Sadeghi M, Pezeshk H, Eslahchi C, Habibi M (2008) Prosign: a method for protein secondary structure assignment based on three-dimensional coordinates of consecutive c(alpha) atoms. *Comput Biol Chem* 32(6):406–411
- Ikeda M, Arai M, Lao DM, Shimizu T (2002) Transmembrane topology prediction methods: a re-assessment and improvement by a consensus method using a dataset of experimentally-characterized transmembrane topologies. In *Silico Biol* 2:19–33
- Jayasinghe S, Hristova K, White SH (2001a) MPTopo: a database of membrane protein topology. *Protein Sci* 10:455–458
- Jayasinghe S, Hristova K, White SH (2001b) Energetics, stability, and prediction of transmembrane helices. *J Mol Biol* 312:927–934
- Jones DT (2007) Improving the accuracy of transmembrane protein topology prediction using evolutionary information. *Bioinformatics* 23:538–544
- Juretic D, Zoranic L, Zucic D (2002) Basic charge clusters and predictions of membrane protein topology. *J Chem Inf Comput Sci* 42:620–632
- Kahsay R, Gao G, Liao L (2005) An improved hidden Markov model for transmembrane protein detection and topology prediction and its applications to complete genomes. *Bioinformatics* 21:1853–1858
- Kall L, Sonnhammer E (2002) Reliability of transmembrane predictions in whole-genome data. *FEBS Lett* 532:415–418
- Kall L, Krogh A, Sonnhammer E (2007) Advantages of combined transmembrane topology and signal peptide prediction—the Phobius web server. *Nucleic Acids Res* 35:W429–W432
- Khan A, Majid A, Choi TS, Acids A (2010) Predicting protein subcellular location: exploiting amino acid based sequence of feature spaces and fusion of diverse classifiers. *Amino Acids* 38:347–350
- Klabunde T, Hessler G (2002) *Chem Bio Chem* 3:928–944
- Krogh A, Larsson B, von Heijne G, Sonnhammer EL (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol* 305:567–580
- Kyte J, Doolittle R (1982) A simple method for displaying the hydrophobic character of a protein. *J Mol Biol* 157:105–132
- Lo A, Chiu HS, Sung TY, Lyu PC, Hsu WL (2008) Enhanced membrane protein topology prediction using a hierarchical classification method and a new scoring function. *J Proteome Res* 7:487–496
- Martelli P, Fariselli P, Casadio R (2003) An ENSEMBLE machine learning approach for the prediction of all-alpha membrane proteins. *Bioinformatics* 19:i205–i211
- Melen K, Krogh A, von-Heijne G (2003) Reliability measures for membrane protein topology prediction algorithms. *J Mol Biol* 327:735–744
- Moller S, Kriventseva EV, Apweiler R (2000) A collection of well characterized integral membrane proteins. *Bioinformatics* 16:1159–1160
- Moller S, Croning MD, Apweiler R (2001) Evaluation of methods for the prediction of membrane spanning regions. *Bioinformatics* 17:646–653
- Nakai K, Kanehisa M (1992) A knowledge base for predicting protein localization sites in eukaryotic cells. *Genomics* 14:897–911
- Naveed M, Khan A (2012) GPCR-MPredictor: multi-level prediction of G protein-coupled receptors using genetic ensemble. *Amino Acids* 42:1809–1823
- Nugent T, Jones D (2009a) Transmembrane protein topology prediction using support vector machines. *BMC Bioinformatics* 10:159
- Nugent T, Jones D (2009b) Predicting transmembrane helix packing arrangements using residue contacts and a force-directed algorithm. *PLoS Comput Biol* 6:e1000714
- Persson B, Argos P (1996) Topology prediction of membrane proteins. *Protein Sci* 5:363–371
- Pylouster J, Bornot A, Etchebest C, Brevern AGD (2010) Influence of assignment on the prediction of transmembrane helices in protein structures. *Amino Acids* 39(5):1241–1254
- Rost B, Casadio R, Fariselli P, Sander C (1995) Transmembrane helices predicted at 95% accuracy. *Protein Sci* 4:521–533
- Rost B, Fariselli P, Casadio R (1996) Topology prediction for helical transmembrane proteins at 86% accuracy. *Protein Sci* 5:1704–1718
- Shen H, Chou JJ (2008) MemBrain: improving the accuracy of predicting transmembrane helices. *PLoS ONE* 3:e2399
- Sonnhammer EL, Von Heijne G, Krogh A (1998) A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc Int Conf Intell Syst Mol Biol* 6:175–182
- Suyama M, Ohara O (2003) Domcut: prediction of inter-domain linker regions in amino acid sequences. *Bioinformatics* 19:673–674
- Tusnady GE, Simon I (1998) Principles governing amino acid composition of integral membrane proteins: application to topology prediction. *J Mol Biol* 283:489–506
- Tusnady GE, Simon I (2001) The HMMTOP transmembrane topology prediction server. *Bioinformatics* 17:849–850
- Viklund H, Elofsson A (2004) Best alpha-helical transmembrane protein topology predictions are achieved using hidden Markov models and evolutionary information. *Protein Sci* 13:1908–1917
- Von Heijne G (1992) Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule. *J Mol Biol* 225:487–494
- Wang XF, Chen Z, Wang C, Yan RX, Zhang Z, Song J (2011) Predicting residue-residue contacts and helix-helix interactions in transmembrane proteins using an integrative feature-based random forest approach. *PLoS ONE* 6:e26767
- Wang C, Xi L, Li S, Liu H, Yao X (2012) A sequence-based computational model for the prediction of the solvent accessible surface area for α-helix and β-barrel transmembrane residues. *J Comput Chem* 33:11–17
- Zaki N, Bouktif S, Sanja LM (2011a) A combination of compositional index and genetic algorithm for predicting transmembrane helical segments. *PLoS ONE* 6(7):e21821
- Zaki N, Bouktif S, Sanja LM (2011b) A genetic algorithm to enhance transmembrane helices topology prediction using compositional index, ACM GECCO'11, Dublin