A Genetic Algorithm to Enhance Transmembrane Helices Prediction

Nazar Zaki*
Intelligent Systems
Faculty of Info. Technology
UAEU, Al Ain 17551, UAE
nzaki@uaeu.ac.ae

Salah Bouktif†
Software Development
Faculty of Info Technology
UAEU, Al Ain 17551, UAE
salahb@uaeu.ac.ae

Sanja Lazarova-Molnar‡
Intelligent Systems
Faculty of Info. Technology
UAEU, Al Ain 17551, UAE
sanja@uaeu.ac.ae

ABSTRACT

A transmembrane helix (TMH) topology prediction is becoming a central problem in bioinformatics because the structure of TM proteins is difficult to determine by experimental means. Therefore, methods which could predict the TMHs topologies computationally are highly desired. In this paper we introduce TMHindex, a method for detecting TMH segments solely by the amino acid sequence information. Each amino acid in a protein sequence is represented by a Compositional Index deduced from a combination of the difference in amino acid appearances in TMH and non-TMH segments in training protein sequences and the amino acid composition information. Furthermore, genetic algorithm was employed to find the optimal threshold value to separate TMH segments from non-TMH segments. The method successfully predicted 376 out of the 378 TMH segments in 70 testing protein sequences. The level of accuracy achieved using TMHindex in comparison to recent methods for predicting the topology of TM proteins is a strong argument in favor of our method.

Categories and Subject Descriptors

I.5 [**PATTERN RECOGNITION**]: [Structural, bioinformatics]; I.2.8 [**Problem Solving, Control Methods, and Search**]: [Heuristic methods]

*Dr. Nazar Zaki is a Director of the Bioinformatics Laboratory and a Coordinator of Intelligent Systems with the Faculty of Information Technology, United Arab Emirates University, Al Ain, P. O. Box 17551, UAE, Email: nzaki@uaeu.ac.ae

†Dr. Salah Bouktif is an Assistant Professor of Software Development in the Faculty of Information Technology at the United Arab Emirates University, Al Ain, P.O. Box 17551, UAE, UAE, Email:salahb@uaeu.ac.ae

‡Dr. Sanja Lazarova-Molnar is is an Assistant Professor of Intelligent Systems in the Faculty of Information Technology at the United Arab Emirates University, Al Ain, P.O. Box 17551, UAE, Email: sanja@uaeu.ac.ae

GECCO'11, July 12–16, 2011, Dublin, Ireland.

General Terms

Algorithms, Performance, Reliability, Verification, Theory

Keywords

Membrane protein, transmembrane helices, genetic algorithm, amino acid composition, compositional index.

1. INTRODUCTION

1.1 Background and motivation

A biological membrane or biomembrane is an enclosing or separating membrane that acts as selective barricade within or around a cell in which cells may maintain specific chemical or biochemical environment. Membrane proteins play key roles in biological systems as pores, ion channels and receptors. Being important in intracellular communication and coordination, membrane proteins may serve as good drug targets. For instance, varying the function of signaling, proteins may assist in correcting defects in signaling that are the root of many diseases. Biological membrane is usually spanned by a TM protein which makes them important targets of both basic science and pharmaceutical research [1]. The major category of TM proteins is the Alpha-helical proteins. This protein category constitutes roughly 30% of a typical genome and is usually present in the inner membranes of bacterial cells, the plasma membrane of eukaryotes or in the outer membranes. In fact, alpha-helical transmembrane proteins are involved in a wide range of important biological processes such as cell signaling, transport of membrane-impermeable molecules, cell-cell communication, cell recognition and adhesion. Since many TM are also prime drug targets, it has been estimated that more than half of the currently commercialized drugs target membrane proteins [2]. Therefore, the prediction of TMH could play an important role in the study of membrane proteins. The importance of this role is emphasized by the lack of high-resolution structures for such proteins, available for no more than 0.5% of the Protein Data Bank (PDB). Knowledge of the TMH topology can help in identifying binding sites and infer functions for membrane proteins. However, because membrane proteins are hard to solubilize and purify, only a very small amount of membrane proteins have structure and topology experimentally determined. This has motivated various computational methods for predicting the topology of membrane proteins [3]. These methods enclose important applications in genome analysis, and can be used to extract global trend in membrane protein evolution.

1.2 Existing methods

In the last two decades, researchers have developed a battery of successively more powerful methods for predicting TMH. This development can be broken into three main categories. In the first category, early TMH prediction methods were based on experimentally determined hydropathy indices of hydrophobic properties for each residue in the protein sequence. Examples of this category include TOPPred [4], DAS-TMfilter [1] and SOSUI [5] which are among the most reliable methods in providing descriptive information about TMHs. These methods use hydrophobicity analysis alone and therefore, they can not predict TMHs with length greater than 25 residues [6]. The recent high-resolution structures production of helical membrane proteins revealed that TMH could have a wide length distribution of more than 25 residues.

In the second category, further accuracy was achieved by employing probabilistic approaches such as Hidden Markov Models (HMMs). In this case actual biological structural knowledge was incorporated into the model's architecture in order to increase its prediction power. Methods such as HMM-TOP [7], TMHMM [8], THUMBU [9] and Phobius [10], allowed researchers to predict reliable integral membrane proteins in a large collection of genome. However, HMM based methods are considered computationally expensive since they involve multiple sequences alignments, calculation of the profile HMM topology and parameterization, and training via expectation maximization. Moreover, the HMM based methods are unable to correctly predict TMHs shorter than 16 residues or longer than 35 residues [6]. As for distantly related protein sequences, a profile alignment may not be possible if, for example, the sequences contain shuffled domains.

In the third category, additional accuracy was gleaned by leveraging machine learning techniques such as neural networks, support vector machines and k-nearest neighbor. Examples of this category include PHD [11], MemBrain [6] and MEMSAT-SVM [12]. Despite their success, the mentioned machine learning methods have two major limitations. First, the learning ability drops when the datasets are small. Second, the feature extraction step requires extensive computations, and thereby a simple algorithm that does not require sequence alignments in the feature extraction step is desirable.

1.3 Proposed solution

In this paper, we focus on the determination of TMH spanning segments and the amino-terminal orientations. We introduce TMHindex which predicts TMH segments solely by the amino acid information. The prediction is performed by using TMH compositional index deduced from the dataset of TMH segments and the amino acid composition. A TMH preference profile is then generated by calculating the average TMH index values along the amino acid sequence using a sliding window of different sizes. Finally, a genetic algorithm was employed to refine the prediction by detecting the optimal set of threshold values that separate the TMH segments from non-TMH segments.

2. METHOD

In this section we introduce the proposed method for predicting TMH proteins topology, referred to as TMHindex. An overview of TMHindex method is shown in Figure 1. TMHindex consists of the two following major steps detailed further in Sections 2.1 and 2.2, respectively:

1. Calculation of the TMH compositional index: In this step we extract the TMH segments and non-TMH segments from the training dataset, compute the difference in amino acid appearances in TMH segments and non-TMH segments, compute the amino acid composition of the protein testing sequence and finally calculate the TMH compositional index.

2. Employing a Genetic Algorithm (GA) to find the optimal set of threshold values: In this step we tailor a GA in order to find an optimal set of threshold values that will accurately segregate TMH and non-TMH segments.

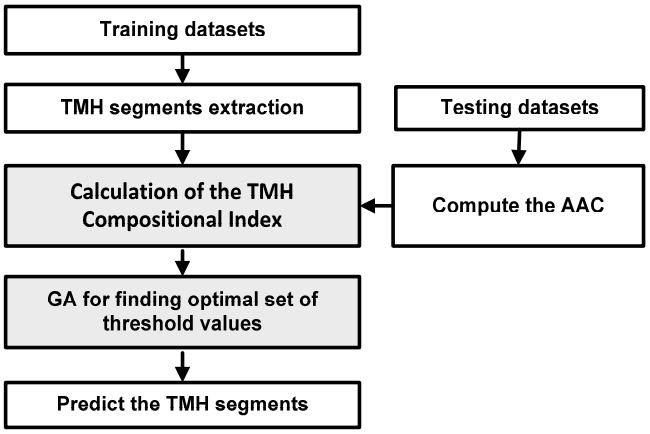

Figure 1: TMHindex overview.

2.1 TMH compositional index

We started by analyzing the amino acid composition in TMH segments and non-TMH segments. We denoted by S^* the enumerated set of sequences in the database of membrane protein sequences. From each protein sequence s_i in S^*, we extracted known TMH and non-TMH segments and store them in datasets S_1 and S_2, respectively. To represent the preference for amino acid residues in TMH segments, we defined an index t. The index t_i for the amino acid $i \in$ {A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T,W, Y, V}, is calculated as follows:

$$t_i = -ln\left(\frac{f_i^{non-helix}}{f_i^{helix}}\right) \quad (1)$$

where $f_i^{non-helix}$ and f_i^{helix} are respectively the frequencies of amino acid i in the datasets S_1 and S_2. The negative value of t_i (threshold value of 0) indicates that the amino acid i preferably exists in TMH segment. This is rather analogous way to DomCut method [13] which was developed to predict the inter-domain linker segments in amino acid sequences. However, the information contained in the index values t_i has demonstrated as insufficient to accurately predict the TMH segments, thus we incorporated the amino acid composition knowledge to the t_i index. The conventional Amino Acid Compositions (AAC) contain 20 components, each of

which reflects the normalized occurrence frequency for one of the 20 native amino acids in a sequence. Owing to its simplicity, the AAC model was widely used in many earlier statistical methods for predicting protein attributes. Therefore, it was previously used in many bioinformatics applications such as inferring the lifestyle of an organism from the characteristic properties of its genome [14] and compensating for the lack of domain information in predicting protein-protein interaction [15].

To this end, we recalculated the compositional index r_i as follows:

$$r_i = -ln\left(\frac{f_i^{non-helix}}{f_i^{helix}}\right) \times a_i, \quad (2)$$

where a_i is the AAC of amino acid i. We then represented each residue in each of the testing protein sequences by its corresponding compositional index r_i. The index values are then averaged over a window that slides along the length of each protein sequence. To calculate the averaged compositional index values $m_{k,j}^w$ for a protein sequence s_k, given a single window size w, we apply Algorithm 1 where L_k is the length of the k^{th} protein and $s_{k,j}$ is the amino acid at position j in protein sequence s_k:

Algorithm: CompositionalIndexAlgorithm(w)
for $j \leftarrow 1$ **to** L_k **do**
  if $j > (w-1)/2$ and $j \leq L_k - (w-1)/2$ **then**
    $m_{k,j}^w \leftarrow \frac{\sum_{i=j-(w-1)/2}^{j+(w-1)/2} r_{s_{k,i}}}{w}$
  else if $j \leq (w-1)/2$ **then**
    $m_{k,j}^w \leftarrow \frac{\sum_{i=1}^{j+(w-1)/2} r_{s_{k,i}}}{j+(w-1)/2}$
  else if $j > L_k - (w-1)/2$ **then**
    $m_{k,j}^w \leftarrow \frac{\sum_{i=j-(w-1)/2}^{L_k} r_{s_{k,i}}}{L_k - j + 1 + (w-1)/2}$
  end
  end
end
end

Algorithm 1: Averaged Compositional Index Algorithm

As revealed in MemBrain method [6], fusion of various window sizes provides more flexibility in accounting for length variation of TMHs, and thus reduces the bias towards a fixed TMH length introduced by using only one window size (as treated in most of the previous TMH topology predictors). Therefore, the averaging is carried across a sequence of odd window sizes ranging from b to e ($3 \leq b < e$), yielding the set of values $\overline{m}_{k,j}$ for each sequence k:

$$\overline{m}_{k,j} = \frac{\sum_{l=0}^{(e-b)/2} m_{k,j}^{b+2l}}{((e-b)/2)+1}, j = 1, ..., L_k, \quad (3)$$

where l is the summation index that ranges across the $\frac{e-b}{2}+1$ window sizes.

The values $\overline{m}_{k,j}$ are then used in conjunction with GA to refine the prediction by detecting short loops and turns that separate the TMH segments.

2.2 Dynamic threshold using GA

Finding an optimal threshold which separates TMH segments from non-TMH segments is crucial to the accuracy of the topology prediction. It is a challenging matter that remains unsolved by many existing predictors. Most of the existing methods were using fixed thresholds to segment the scores (e.g. residues with scores higher than a defined threshold value, are assigned as helix segment). Indeed, this is a weakness because optimal threshold for defining two TMH segments separated by long loops is different from a threshold required for identifying TMH segments separated by short loops or tight turns. High-resolution structures show that two consecutive TMH segments are often connected by very short loops or turns and that is why in MemBrain [6] for instance, the authors have utilized a dynamic threshold value in which a base threshold propensity of 0.4 was used to initially define TMH fragments. Then the threshold was raised according to the shape of the local propensity profile for identifying short loops or helical breaks in fragments. Despite the success shown by utilizing dynamic threshold, it is noticeable that rising the threshold could improve the predictions of the TMH segments in part of the sequence and could reduce the prediction accuracy in another part of the sequence.

In our present work, we consider the amino acid sequence as a set of sequence chunks. Each chunk would have its proper dynamic threshold value. Therefore, the problem turns out to be a search problem of a set of dynamic threshold values that will better reflect the structure of the amino acid sequence and predict accurately the TMH and non-TMH segments. Such a search problem can be seen as a kind of partition problems [16], known as NP-complete almost certainly unsolvable with a polynomial time algorithm. The application of metaheuristic search techniques to this class of problems is a promising solution [16–18]. Metaheuristics are high-level frameworks that employ heuristics to find solutions for combinatorial problems at a reasonable computational cost. Moreover, they are strategies ready for adaptation to specific problems. In particular, GA is one of the most commonly used techniques and has proven its effectiveness in combinatorial optimization and complex prediction [16, 19]. Besides, GA is easily customizable for our problem. In the following sections we will focus on the description of GA and its adaptation to our TMH segment prediction method.

2.2.1 *Overview of Genetic Algorithms*

The basic idea of GA is to typically start from a set of initial solutions, and use biologically inspired evolution mechanisms to derive new and possibly better solutions ([17]). The derivation starts by an initial solution set P_0 (called the initial population), and generates a sequence of populations $P_1, , P_T$, each obtained by "mutating" the previous population. The elements of the solution sets are called chromosomes. The fitness of each chromosome is measured by an objective function called fitness function. Each chromosome (possible solution) consists of a set of genes. At each generation, the algorithm selects a number of pairs of chromosomes using a selection method that gives priority to the fittest chromosomes. To each selected pair, the algorithm applies one of two operators, crossover and mutation, with probability pc and pm, respectively, where pc and pm are input parameters of the algorithm. The crossover operator combines the genes of the two chromosomes, while the mutation operator randomly modifies certain genes. Each selected pair of chromosomes produces a new pair of chro-

mosomes that constitute the next generation. The N_e fittest chromosomes of each generation are automatically added to the next generation. The algorithm stops if a convergence criterion is satisfied or if a fixed number of generations is reached. The GA is summarized in Algorithm 2.

Algorithm: GeneticAlgorithm(T, pc, pm, N_e)
Initialize P_0
BestFit $\leftarrow$ fittest chromosome of P_0
BestFitEver $\leftarrow$ BestFit
for $t \leftarrow 0$ **to** T **do**
  $Q \leftarrow$ pairs of members selected by roulette-wheel from P_t
  $Q' \leftarrow$ offsprings of pairs in Q derived by crossover and mutation
  $P_{t+1} \leftarrow Q' \cup$ {the N_e fittest members of P_t}
  BestFit $\leftarrow$ fittest chromosome in P_{t+1}
  if BestFit is fitter than BestFitEver **then**
    BestFitEver $\leftarrow$ BestFit
  end
end
return BestFitEver

Algorithm 2: Summary of a genetic algorithm

To apply GA to a specific problem, all elements of the generic algorithm must be customized and adapted to the problem. In particular, the solutions must be encoded into chromosomes, the two operators (crossover and mutation) and the fitness function must be defined.

2.2.2 Encoding protein sequence as chromosome

To properly apply GA to our problem, we define a chromosome encoding for the protein sequence represented by a vector of $m_{k,j}$ values calculated in Equation 3. As each chromosome is a set of genes of size N, we encode a gene as a pair (λ, μ), where λ is a threshold value and μ is the upper position in the protein sequence before which λ is used as threshold. For more details let $(\lambda_{i-1}, \mu_{i-1})$, (λ_i, μ_i) and $(\lambda_{i+1}, \mu_{i+1})$ be three consecutive genes in the chromosome representing the sequence of a given protein. The value λ_i is interpreted as the threshold applied from the position μ_{i-1} to the position μ_i in the protein's sequence and λ_{i+1} is the threshold applied from the position μ_i to the position μ_{i+1} in the sequence. In particular, the threshold λ_1 would be applied from the beginning of the sequence to the position μ_1 as illustrated in Figure 2.

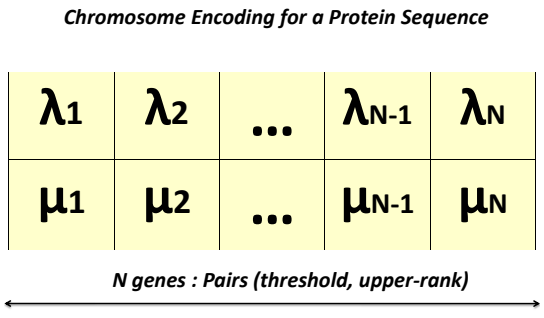

Figure 2: Encoding protein sequence as chromosome

2.2.3 Crossover Operator

Crossover is a reproduction operator that occurs with high probability pc. It takes two parent chromosomes and produces one or two child chromosomes. Our encoding scheme of chromosomes allows greater freedom to use different ways of performing crossover and mutation. Based on the chromosome real-representation, we define two types of crossover techniques for our problem; one-cut point crossover, and uniform crossover.

One-cut point crossover is a standard way to perform crossover between the chromosomes. It consists of cutting at a position i one of the two parent chromosomes into two subsets of genes (vector of pairs λ and μ). Then the second chromosome is cut at the position j into two other subsets. The cutting point j is determined as the rank of the pair (λ_i, μ_i) where the position μ_j is the smallest position in the second parent chromosome greater than μ_i. Two new chromosomes are then created by interleaving the subsets. Figure 3 shows an illustrative example of the one-cut point crossover between chromosomes representing a protein sequence of length 250.

In the uniform crossover, two parent chromosomes give birth to a single offspring. Each gene of the new offspring is a copy of a gene from one of the parents selected in the following way:

The gene i of the offspring (λ_i, μ_i) is a copy of the gene i of one of the two parent selected randomly. If μ_i in the gene i of the selected parent is not greater than μ_{i-1} of the previously copied gene, then the gene i is selected from the second parent.

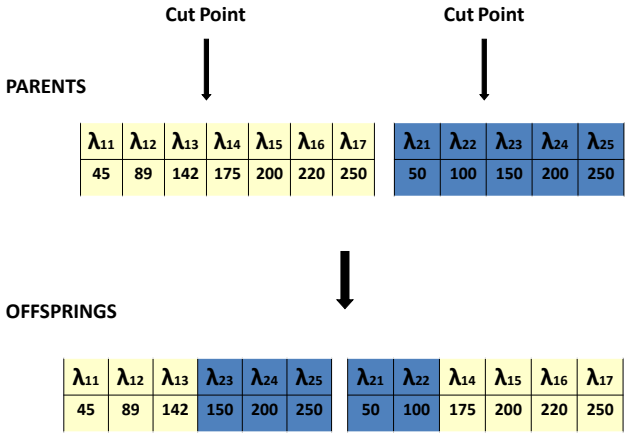

Figure 3: Crossover operator

The crossover ways described above are motivated by the fact that moving the intervals boundaries or modifying the threshold values will diversify the population and then by the fitness evaluation, selection process and elitism (see Algorithm 2), the fittest solutions will be chosen to reproduce and form the next improved generation.

2.2.4 Mutation Operator

Mutation is the second reproduction operator that occurs with a small probability pm. It is employed to extend the search space by creating new points that could be potential solutions. When a chromosome is selected for mutation, a

small number of its genes are randomly chosen in order to be modified. With our chromosome encoding, two ways of modifying a gene (λ_i, μ_i) are possible (See Figure 4). In the first, the threshold λ is modified by making a positive or negative variation of its value, while in the second way, the upper bound μ_i is moved either towards μ_{i-1} or μ_{i+1}.

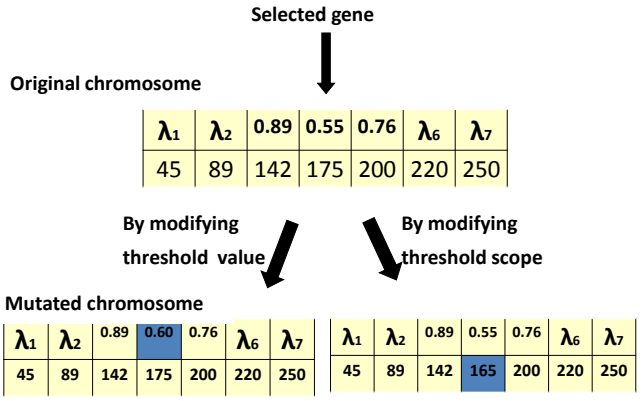

Figure 4: Mutation operator

2.2.5 *Initial Population and Selection Method*

Before starting its evolutionary process, the GA needs to build an initial population P_0 of solutions (vectors of pairs). An individual of the initial population P_0 is obtained by firstly choosing a random number N of ranges within the protein sequence size. Next, the N ranges are arbitrarily defined by slicing the protein sequence into N chunks. Finally, for each range, a threshold is defined by selecting randomly one of the scores assigned to one of the residues within the range. The size of the initial population is a parameter of our algorithm that will be set after several tuning runs. The process of evolution starts by selecting a pair of chromosomes according to the *roulette-wheel* selection technique, whereby one can imagine a roulette wheel where all chromosomes are placed. Each chromosome is assigned a portion of the wheel that is proportional to its fitness. A marble is thrown and the chromosome where the marble halts is selected.

3. EXPERIMENTAL RESULTS

3.1 Evaluation measures

To test the TMHindex method and compare its performance to the existing state-of-the-art predictors, we used four commonly used evaluation measures:

1. TMH segments prediction success rate (r_{psr}),

$$r_{psr} = r_c/r_t, (r_t = 378) \quad (4)$$

Where r_c and r_t are the number of TMH segments correctly predicted and total number of TMH segments in the test dataset, respectively. A prediction is considered correct if there is an overlap of at least nine amino acids between the predicted and experimentally known TM segment. This threshold length is quite reasonable in comparison to the typical TMH which are on average 21 residues long. Different residues overlap was used in the past such as 3 residues [8], 5 residues [20] and 9 residues [6].

2. Protein prediction success rate (p_{psr}),

$$p_{psr} = p_c/p_t, (p_t = 70) \quad (5)$$

Where p_c and p_t are the number of correctly predicted proteins and total number of proteins in the test dataset, respectively. A protein is considered correctly predicted if all its TMH segments are correctly predicted.

3. Residue prediction success rate (s_{psr}),

$$s_{psr} = s_c/s_t, (s_t = L_k) \quad (6)$$

Where s_c and s_t are the number of correctly predicted residues and the total number of residues in a protein sequence, respectively. This evaluation measure is equally used as a fitness function to the proposed GA.

4. The N-score and C-score,

These two scores (illustrated in Figure 5) evaluate the accuracy of predicting the in and out ends of TMHs [21]. N and C scores are the number of N- and C-terminal residues that do not match when comparing the predicted TMH segment and the known TMH segment. A lower score in this case yields a more accurate prediction. If the prediction of this TMH segment is an exact match, then the prediction score is equal to 0. It means that its N-score and C-score are null.

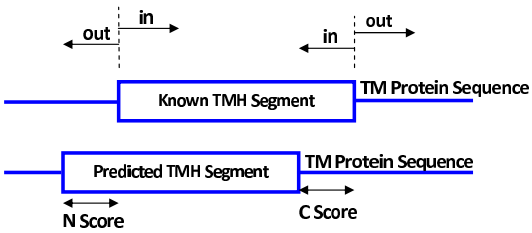

Figure 5: The N and C scores.

3.2 Illustration

To illustrate the experimental work, in Figure 6 and Figure 7 we show the way the TMH segment is detected in a sample protein 1OCC using the index t_i with a threshold value of 0. We used odd window sizes, from $b = 5$ to $e = 19$ and the final values $\overline{m}_{k,j}$ representing each amino acid in the sequence are calculated. The maximum window size was chosen to be 19 because a 19-residue segment is close to the thickness of the hydrocarbon core of a lipid bilayer [22]. In this case, the known TMH segment (in bold) starts in residue 12 and ends in residue 35. The length of the protein sequence $L_k = 46$ and therefore $s_{psr} = 0.78$, C-score = 6 and N-score = 4.

To improve the prediction accuracy we incorporated the compositional index r_i and the results are shown in Figure 8,

>1OCC:M|PDBID|CHAIN|SEQUENCE
ITAKPAKTPTSPKEQAIGLSVTFLSFLLPAGWVLYHLDNYKKSSAA
← Known TMH Segment →

Figure 6: Sample protein 1OCC.

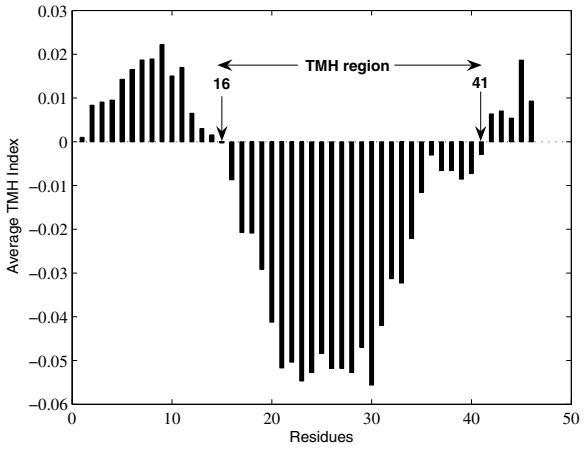

Figure 7: TMH segment detection in protein 1OCC using the index t_i.

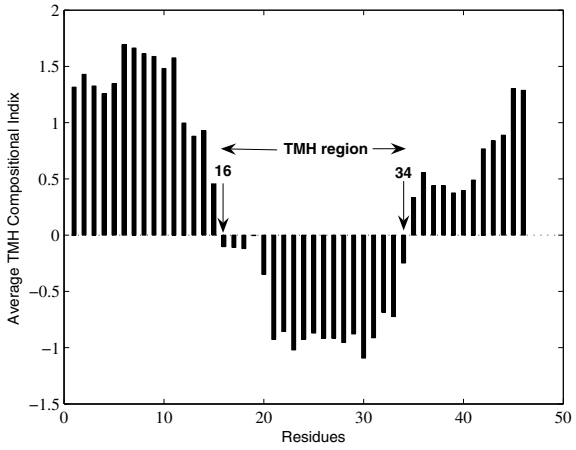

Figure 8: TMH segment detection in protein 1OCC using the compositional index r_i.

therefore the obtained accuracies were improved, i.e., $s_{psr} = 0.89$, C-score = 1 and N-score = 4.

As a second enhancement of our approach, GA was applied to find the optimal threshold set separating TMH segments from the non-TMH segments as illustrated in Figure 9. Prior to the application of GA, several runs were performed in order to tune the different parameters. As a result of parameters tuning, our algorithm converged towards a near-optimal solution after $T = 80$ generations and with a population size set to 80. During the reproduction process, crossover and mutation occur with probabilities pc equal to 0.6 and pm equal to 0.2, respectively. The elitism strategy was used by which the N fittest chromosomes of one generation are cloned and copied to the next generation. After applying GA to the sequence of the protein 1OCC, the latter is divided into 2 equal parts. Each part consists of 23 residues and the two upper boundary positions, μ_1 and μ_2, are respectively set to 23 and 46. The threshold values λ_1 and λ_2 are computed to be 1 and 0.25, respectively. The obtained structure of the protein 1OCC computed by GA achieved high accuracies, i.e., $s_{psr} = 1$, C-score = 0 and N-score = 0.

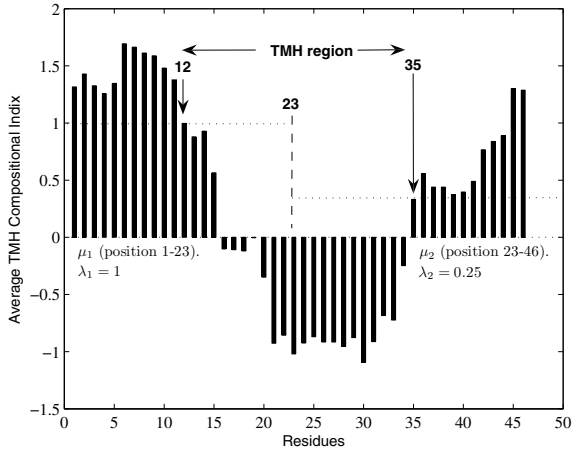

Figure 9: TMH segment detection in protein 1OCC using GA.

3.3 Comparison with existing methods

The aim of the TMH segments predictions method is to obtain high accuracy when applied to unknown proteins. For predicting the TMH segment within a protein, we first computed the index t_i. We collected the TMH and non-TMH segments from a training dataset. The training dataset contains 50 protein sequences which consist of 327 known TMH segments. The testing dataset used contains 70 protein sequences which consist of 378 known TMH segments. The training and testing datasets have no sequence overlap. The datasets have experimentally determined TMH topology and they were used by most of other TMH predictors such as MemBrain [6], Phobius [10], THUMBU [9] and TMHMM [8]. The datasets are available at ftp://ftp.ebi.ac.uk/pub/datasets/testsets/transmembrane.

The performance of TMHindex was measured by r_{psr}, p_{psr}, N-score, C-score and the number of TMH segments which were correctly predicted. The comparisons of the performance of TMHindex with those of THUMBU, SOSUI , DAS-TMfilter , TOP-PRED, TMHMM, Phobious and MemBrain, are reported in Table 1. The results show that TMHindex is successful in making fewer mis-classifications of TM helices. It outperforms the compared methods according to all of the measures used for performance evaluations. TMHindex was able to predict 376 of the total 378 TMH segments in the test dataset. The unpredicted TMH were from proteins 2IUB:A and 2B5F:A. Furthermore, the residue prediction success rate s_{psr} was 0.905.

The distributions of helix lengths in the testing datasets were also examined (Figure 10). The investigation shows

Table 1: Performance comparison of various TMH predictors.

Predictor	r_{psr} (%)	p_{psr} (%)	N-Score	C-Score	Correct TMHs
THUMBU	85.5	47.1	6.9 ± 4.9	058 ± 0.19	316
SOSUI	89.1	57.1	5.0 ± 4.2	0.44 ± 0.21	334
DAS-TMfilter	90.7	64.3	5.5 ± 5.3	0.58 ± 0.16	341
TOP-PRED	92.6	60	4.6 ± 3.9	0.45 ± 0.15	352
TMHMM	91	65.7	4.5 ± 3.9	0.44 ± 0.15	343
Phobious	91.8	71.4	4.4 ± 4.1	0.44 ± 0.19	345
MemBrain	97.9	87.1	3.1 ± 2.8	0.35 ± 0.14	371
TMHindex	99.46	91.1	2.19 ± 0.04	2.04 ± 0.03	376

that the prediction methods typically search for TM helices with length raging between 17 and 25 residues. In fact, out of the 378 TM helices in the dataset, only 204 (54%) of the helices fall within this range, 29 (7.7%) have length less than 17 and 145 (38.3%) of the helices with length exceeding 25 residues. Several membrane proteins contain TM helices that do not span the bilayer, for example the pore (P) helix of the potassium channel KcsA (1K4C) and the NPA-containing loops of the aquaporins. These 'half-TMs' are shorter in length than conventional TM helices and are expected to be more difficult to predict [21]. The distributions of TM helices given in Figure 10 reveal a small but significant population of half-TMs are present in the testing dataset. Similarly, there are many TMH segments which are longer than 25 residues in length that often ended unpredicted or partially predicted by most of the available methods. Figure 10 clearly show that Phobius is unable to detect TMH segments shorter than 16 and longer than 30 residues. DAS-TMfilter and THUMBU are unable to detect many TMH segments longer than 25 residues. MemBrain is unable to detect many TMH segments longer than 30 residues. The only remark that needs more inversigation of the TMHindex method is related to the prediction of some TMH segments of length 23 and 24, respectively. Their predictions show more errors than any other segments.

With respect to CPU time, the current version of TMHindex needs approximately 20 minutes for predicting and converging towards accurate structures of the available 70 protein sequences using a computer equipped with Intel Core 2 Duo CPU T7250 @ 2.00 GHz and 2.99 GB of RAM.

4. CONCLUSION

The prediction of TMH has proven to be important in the study of membrane proteins which play many roles in cells, including TM transport, signaling and energy transduction. In this paper we introduce TMHindex method which was able to successfully predict 376 out of the 378 TMH segments in 70 benchmark protein sequences. The level of the accuracy achieved using TMHindex in comparison to known methods for predicting the topology of TM proteins is a strong indication of the TMHindex capability. The improvement of the proposed method was due to two main reasons. First, the employment of the TMH compositional index which was deduced from a dataset of priory known TMH segments and the incorporation of the amino acid composition knowledge. Second, tailoring a GA which offered a flexible way to model an intelligent predictor of TM proteins segments based on more dynamic thresholds.

In the future, we will extend the TMHindex method to predict signal peptides. Predicting TMH and signal peptides is challenging because of the high similarity between the hydrophobic regions of a transmembrane helix and that of a signal peptide [10]. Although, GA customization significantly improved the prediction, further tuning and other strategies choices within the metaheuristic could achieve more capable and flexible prediction.

5. REFERENCES

[1] M. Cserzo, F. Eisenhaber, B. Eisenhaber, and I. Simon. TM or not TM: transmembrane protein prediction with low false positive rate using DAS-TMfilter. Bioinformatics, 20(1):136–137, 2004.

[2] T. Nugent and D. T. Jones. Transmembrane protein topology prediction using support vector machines. BMC Bioinformatics, 10, 2009.

[3] R. Y. Kahsay, G. R. Gao, and L. Liao. An improved hidden markov model for transmembrane protein detection and topology prediction and its applications to complete genomes. Bioinformatics, 21(9):1853–1858, 2005.

[4] M. G. Claros and Gunnar von Heijne. Toppred II: an improved software for membrane protein structure predictions. Computer Applications in the Biosciences, 10(6):685–686, 1994.

[5] Takatsugu Hirokawa, S. Boon-Chieng, and Shigeki Mitaku. SOSUI: classification and secondary structure prediction system for membrane proteins. Bioinformatics, 14(4):378–379, 1998.

[6] H. Shen and J. J. Chou. Membrain: Improving the accuracy of predicting transmembrane helices. Plos One, 6(3), 2008.

[7] G. E. Tusnady and I. Simon. Principles governing amino acid composition of integral membrane proteins: application to topology prediction. J Mol Biol., 289(2):489–506, 1998.

[8] A. Krogh, B. Larsson, G. von Heijne, and E. L. Sonnhammer. Predicting transmembrane protein topology with a hidden markov model: Application to complete genomes. J. Mol. Biol., 305(3):567–580, 2001.

[9] H. Zhou and Y. Zhou. Predicting the topology of transmembrane helical proteins using mean burial

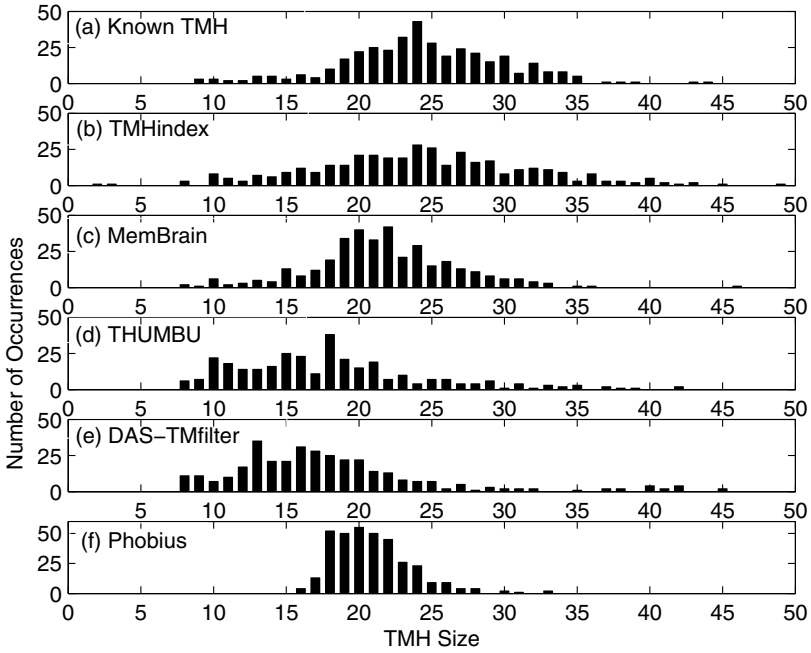

Figure 10: TMH length distribution in (a) 70 known membrane protein structures in the testing dataset, (b) TMHs predicted by TMHindex, (c) Membrain (d) THUMBU (e) DAS-TMfilter (f) Phobius.

propensity and a hidden-markov-model-based method. Protein Sci., 12(7):1547–1555, 2003.

[10] L. Käll, A. Krogh, and E. L. Sonnhammer. A combined transmembrane topology and signal peptide prediction method. J. Mol. Biol., (338):1027–1036, 2004.

[11] B. Rost, R. Casadio, and P. Fariselli. Refining neural network predictions for helical transmembrane proteins by dynamic programming. In David J. States, Pamkaj Agarwal, Terry Gaasterland, Lawrence Hunter, and Randall Smith, editors, Proceedings of the Fourth International Conference on Intelligent Systems for Molecular Biology, pages 192–200, Menlo Park, June 12–15 1996. AAAI Press.

[12] T. Nugent and D.T. Jones. Transmembrane protein topology prediction using support vector machines. BMC Bioinformatics, 26(159), 2009.

[13] M. Suyama and O. Ohara. Domcut: prediction of inter-domain linker regions in amino acid sequences. Bioinformatics, 19(5):673–674, 2003.

[14] F. Tekaia, E. Yeramian, and B. Dujon. Amino acid composition of genomes, lifestyles of organisms, and evolutionary trends: a global picture with correspondence analysis. Gene, 297(4):51–60, 2002.

[15] S. Roy, D. Martinez, O. Platero, T. Lane, and M. Werner-Washburne. Exploiting amino acid composition for predicting protein-protein interactions. PLoS One, 11(4), 2009.

[16] E. Falkenauer. Genetic algorithms and grouping problems. John Wiley and Sons, 1998.

[17] J.H. Holland. Adaptation in Natural Artificial Systems. University of Michigan Press, 1975.

[18] R. Garey. Computers and Intractability: A guide to the theory of NP-completeness. Freeman and Co, 1979.

[19] Salah Bouktif, Faheem Ahmed, Issa Khalil, and Giuliano Antoniol. A novel composite model approach to improve software quality prediction. Information & Software Technology, 52(12):1298–1311, 2010.

[20] E. L. Sonnhammer, G. von Heijne, and A. Krogh. A hidden markov model for predicting transmembrane helices in protein sequences. In Proceedings of the 6th International Conference on Intelligent Systems for Molecular Biology (ISMB-98), Montréal, Québec, Canada, June 28 - July 1, 1998, pages 175–182. AAAI, 1998.

[21] J. M. Cuthbertson, D. A. Doyle, and M. S. Sansom. Transmembrane helix prediction a comparative evaluation and analysis. Protein Eng Des Sel., 18(6):295–308, 2005.

[22] S. Jayasinghe, K. Hristova, and S.H. White. Energetics, stability, and prediction of transmembrane helices. J. Mol. Biol., 312(5):927–934, 2001.