Proceedings of the 28th IEEE
EMBS Annual International Conference
New York City, USA, Aug 30-Sept 3, 2006

Linear discrimination of transmembrane from non-transmembrane segments in proteins using higher-order crossings

Ilias K. Kitsas *Student Member, IEEE*, Stavros M. Panas, *Member, IEEE*, and Leontios J. Hadjileontiadis, *Member, IEEE*

***Abstract*—Identification of transmembrane segments in protein sequences is an important issue in the field of bioinformatics. In this study, a method is proposed for linear discrimination between transmembrane and non-transmembrane segments, combining chemical and statistical features of the proteins with higher-order crossings analysis for protein segment classification. The method was tested on human proteins extracted from public available databases and the results have shown a remarkable efficiency of the proposed algorithm to correctly classify the sequence segments under study into two linearly separated classes, for a wide range of transmembrane segment lengths.**

I. Introduction

PREDICTION of transmembrane (TM) helices in integral membrane proteins is an important aspect of bioinformatics and the existence of reliable methods for their discrimination from the non-transmembrane (NTM) segments of the protein has many applications in genome analysis [1]. So far, a number of algorithms designed to identify transmembrane helices in the primary amino acid sequence have been developed [1], [2]. For most of them, the prediction accuracy raises up to 98% [3], [4]. However, the most common observed error is the under- or over-prediction of TM helices, which is very important, as the number and orientation of the helices typically reveal aspects about function of one TM helix [1].

A number of methods have also been proposed for the distinction between TM and globular proteins with the majority using neural networks [5], atomic coordinates [6], and linear discrimination rules [7]. In the proposed scheme, a new method for linear discrimination of TM from NTM segments was developed, involving the hydrophobicity [8] of the protein, a statistical analysis of the aminoacid sequence (propensities) and higher-order crossings (HOC) [9] analysis. Results of the method applied on human proteins show that the method linearly discriminates TM from NTM segments, providing a reliable and simple criterion for the verification of TM segment prediction.

II. Methodology

A. Sequence Transformation

The authors are with the Department of Electrical & Computer Engineering, Aristotle University of Thessaloniki, GR 54124, Thessaloniki, Greece. Corresponding author: I. Kitsas, Tel.: +302310 996319, FAX: +302310 996312, E-mail: ikitsas@auth.gr

For efficient analysis of a protein sequence a transformation of the sequence from a string of characters (i.e., amino acids) to a numerical sequence should be employed. Two transformations were adopted here; the propensity- and the hydrophobicity-based ones.

In the propensity-based transformation, a training set should be involved. This was composed by 117 human single TM proteins, with a single TM segment clearly verified, extracted from the SWISS-PROT protein database, release 46.0 [10]. Initially, statistical amino acid composition was calculated for each protein of the training set. In particular, the relative frequency of occurrence of residue of type *i* in a protein segment *k* of length $n(k)$, in the state *j* is estimated as

$$f_{i(j)}(k) = \frac{N_{i(j)}(k)}{n(k)}, \quad j = TM, NTM. \qquad (1)$$

Next, the average composition (across all segments) of residue of type *i* in each structural state, is given by

$$C_{i(j)} = \frac{\sum_k f_{i(j)}(k)}{N_{(j)}}, j = TM, NTM, \qquad (2)$$

where $N_{(j)}$ is the number of protein segments that belong to the corresponding state. By using (2), the numerical propensity-based transformation for each amino acid type is calculated as

$$P_i = 100 \log[C_{i(TM)} / C_{i(NTM)}]. \qquad (3)$$

The resulting propensity scale is shown in Fig. 1(a).

Hydrophobicity is a property of the amino acids that determines where the amino acid will be located in the final structure of the protein relative to the membrane [8]. A hydrophobicity scale is a list of amino acids with corresponding hydrophobicity values. So far, many hydrophobicity scales have been proposed in the literature. The scale used in this study is the one most commonly used, proposed by Kyte and Doolittle [8], shown in Fig. 1(b).

These two transformations result in numerical sequences denoted as P_i and H_i for the propensity- and hydrophobicity-based ones, respectively; an example is given in Fig. 2. The ratio P_i / H_i is then estimated for each protein, consisting the analysis signal. In this way, a numerical representation combining both statistical and

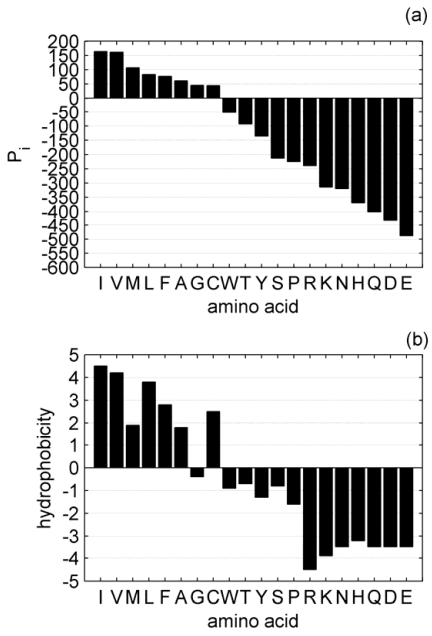

Fig. 1. Scales for (a) propensity (P_i) and (b) hydrophobicity (H) used in the propensity- and hydrophobicity-based numerical transformations.

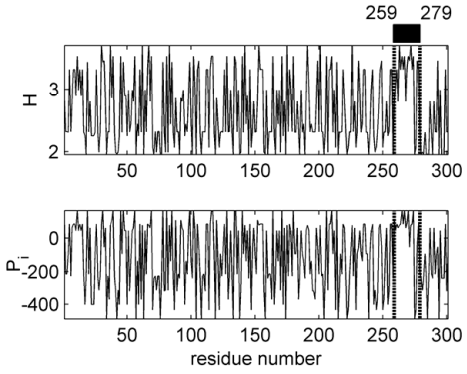

Fig. 2. Example of the application of the proposed protein sequence transformation to the 'Natural cytotoxicity triggering receptor 1 precursor' protein [SWISS-PROT ID: NCTR1_HUMAN, AC: O76036]. Transmembrane helix segment is denoted with the thick black line located at 259-279 residues, crossing all subfigures with dotted lines. P_i corresponds to the output of the propensity-based transformation; and H corresponds to the output of the hydrophobicity-based transformation.

chemical properties of the protein sequences is achieved.

B. Higher-Order Crossings (HOC)

All observed protein sequences display local and global up and down movements as the number of residue increases after their numerical transformation. This behavior, seen in a finite zero-mean sequence $\{Z_n\}, n = 1,\ldots,N,$ oscillating about level zero can be expressed through the zero-crossing count.

Let ∇ be the backward difference operator defined by

$$\nabla Z_n \equiv Z_n - Z_{n-1} \tag{4}$$

The difference operator ∇ is a high-pass filter. If we define the following sequence of high-pass filters

$$\Im_k \equiv \nabla^{k-1},\ k = 1,2,3,\ldots, \tag{5}$$

with $\Im_1 \equiv \nabla^0$ being the identity filter, and with a transfer function given by

$$H(\omega;k) = (1-\exp(-j\omega))^{k-1}, \tag{6}$$

we can estimate the corresponding HOC, namely simple HOC [9], by

$$D_k = NZC\{\Im_k(Z_n)\},\ k = 1,2,3,\ldots;\ n = 1,\ldots,N, \tag{7}$$

where $NZC\{\cdot\}$ denotes the estimation of the number of zero-crossings and

$$\nabla^{k-1} Z_t = \sum_{j=1}^{k} \binom{k-1}{j-1} (-1)^{j-1} Z_{n-j+1}. \tag{8}$$

Given that, we only have finite sequence, we lose an observation with each difference. Hence, to avoid this effect we must index the data by moving to the right, i.e., for the evaluation of k simple HOC, the index $t = 1$ should be given to the k th or a later observation. For the estimation of the number of zero-crossings in (7), a binary sequence $X_t(k)$ is initially constructed given by

$$X_n(k) = \begin{cases} 1 & if\ \Im_k(Z_n) \geq 0 \\ 0 & if\ \Im_k(Z_n) < 0 \end{cases}, \tag{9}$$

and the desired simple HOC are then estimated by counting symbol changes in $X_1(k),\ldots,X_N(k),$ i.e.,

$$D_k = \sum_{n=2}^{N} [X_n(k) - X_{n-1}(k)]^2. \tag{10}$$

From (9) it is obvious that $D_{k+1} \geq D_k$, a fact that reveals the monotonic character of simple HOC. In finite sequences what we really have is the sure inequality $D_{k+1} \geq D_k - 1$ [9]. In addition, as k increases, the discrimination power of simple HOC diminishes, since different processes yield almost the same D_k [9]. A maximum value of $k = 20$ suffices for effective discrimination purposes [9].

Scatter plots of HOC of the optimum order, k_{op}, versus HOC of other (usually neighboring) orders can reveal the presence of two classes. If the classes do not overlap in the scattergrams, linear discrimination could be achieved.

For quantitative definition of the two class-ranges in the HOC scattergrams, the centers of each cluster are estimated, using fuzzy logic-based *c*-means (FCM), which is a data clustering technique wherein each data point belongs to a cluster to some degree.

C. HOC analysis features

In many cases, the distance of two signals from a given reference is used for discrimination. In many respects, white Gaussian noise is the most convenient reference signal, for which the expected HOC are given by [9]

$$E\{D_k\} = (N-1)\left\{\frac{1}{2} + \frac{1}{\pi}\sin^{-1}\left(\frac{k-1}{k}\right)\right\}. \quad (11)$$

The limits for which the D_k falls within, with a probability of approximately 95% for each k are [9]

$$\pm 1.96(N-1)^{1/2}\left\{\frac{1}{4} - \left[\frac{1}{\pi}\sin^{-1}\left(\frac{k-1}{k}\right)\right]^2\right\}^{1/2}. \quad (12)$$

Using (11) and (12), a white noise (WN) test is constructed to examine whether a protein sequence segment oscillates as Gaussian white noise. The hypothesis of white noise should be rejected when at least one D_k, $k = 1,2,\ldots,K$, for some K, falls outside the limits of (12) [9].

D. HOC clusters distance

In order to examine the possibility of potential overlapping between two classes of protein segments, the mean (MN) and the standard deviation (σ) of the distance from cluster centre, c, alongside the standard error of the MN (s), for the two classes in the HOC scattergram, are estimated. Consequently, the borderline of the two clusters is defined as the perpendicular line at the $C(l_1,l_2)$ point of their center-distance, with the C vector satisfying the following equations:

$$\frac{dis(C,c_1)}{MN_1+\sigma_1+s_1} = \frac{dis(C,c_2)}{MN_2+\sigma_2+s_2}, \quad (13)$$

$$dis(C,c_1) + dis(C,c_2) = dis(c_1,c_2), \quad (14)$$

where $dis(a_1,a_2)$ denotes the Euclidean distance between vectors a_1 and a_2, and $c_i, MN_i, \sigma_i, s_i, i = 1,2$, denote the estimated centre vector, the mean, the standard deviation, and the standard error of the mean for the two clusters, respectively.

III. Dataset Characteristics

The dataset used in this work consists of documented transmembrane proteins extracted from SWISS-PROT Release 46.0 [10]. From the initial set of 12108 human protein sequences, automatically selected based on the presence in the feature table of the 'TRANSMEM' keyword, a subset of 1390 sequences was extracted containing all the proteins with a single-membrane segment (both membrane spanning and anchored ones). This set was used to extract the abovementioned training set used in the propensity-based transformation, as well as the training and test sets used for the evaluation of the algorithm. The TM helix length for the proteins of the test sets varied form 10 to 30, whereas for the training set from 17 to 25 residues. Two sets of proteins, i.e., a 'low length' and a 'high length' dataset, were created, including proteins comprising TM and NTM segments with less or equal to 500 residues and more than 500 residues, respectively. The former contained 324 and the latter 164 proteins. Furthermore, two subsets of the 'low-length' and 'high-length' datasets with 80 and 40 proteins, respectively, were randomly extracted and used in the procedure of testing the efficiency of the algorithm. All sequence data were tested against global alignment using the Needleman-Wunsch algorithm [12] resulting in a global similarity less than 33%.

IV. Results and Discussion

Results from the algorithm when applied to the training dataset used in the propensity-based transformation are shown in Fig. 3. In particular, Fig. 3(a) illustrates the WN test, as well as the HOC sequences for the TM and NTM segments of the set. The HOC order k ranged from 1 to 20 and its optimum value was found $k_{op}^{NTM-TM} = 7$, since there, the distance (denoted by an arrow) between the most top HOC line of TM and the most bottom HOC line of NTM is maximised. This implies that scatter plots of the k_{op}^{NTM-TM}-order crossings could provide a clear (linear) discrimination between NTM and TM segments. This is true from the HOC scatter plot of Fig. 3(b), where the HOC pair (D_7, D_8) from Z^{TM} (stars) and Z^{NTM} (circles) is depicted. Clearly, the two processes give rise to distinctly different clusters of pairs (D_7, D_8), which can be linearly discriminated. In this scattergram, the coordinates of the two cluster centers, estimated by the FCM technique, were found equal to $c_{tr}^{NTM}(412.6,406.2)$ and $c_{tr}^{TM}(184.6,155.6)$ with a distance of $dis(c_{tr}^{NTM}, c_{tr}^{TM}) = 338.8$. The same procedure was followed for the 'low-length' and 'high-length' training datasets as well. Figures 4(a) and 5(a) illustrate the WN test, as well as the HOC sequences for the TM and NTM segments of the 'low-length' and 'high-length' training sets, respectively. It is apparent that the optimum HOC order is $k_{op-l}^{NTM-TM} = 8$ and $k_{op-h}^{NTM-TM} = 4$, respectively. Figures 4(b) and 5(b) present the HOC scatter plots for the two training sets. A linear discrimination between the two groups is achieved in both datasets. In the upper left corner of Fig. 5(b), a detailed segment of the diagram for the group of TM segments is shown.

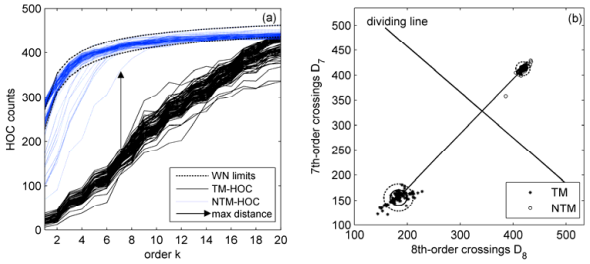

Fig. 3. Results from the algorithm when applied to the training dataset used in the propensity-based transformation. (a) The WN test and (b) the HOC scatter plot of (D_8,D_7) for the TM (stars) and NTM (circles) set segments.

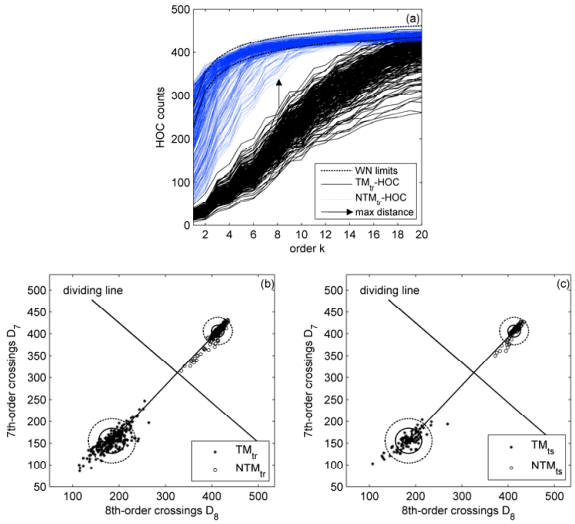

Fig. 4. Results from the algorithm when applied to the 'low length' training dataset. (a) The WN test, (b) the HOC scatter plot of (D_8,D_7) for the TM_{tr} (stars) and NTM_{tr} (circles) segments of the training set, (c) the HOC scatter plot of (D_8,D_7) for the TM_{ts} (stars) and NTM_{ts} (circles) segments of the 'low length' testing set.

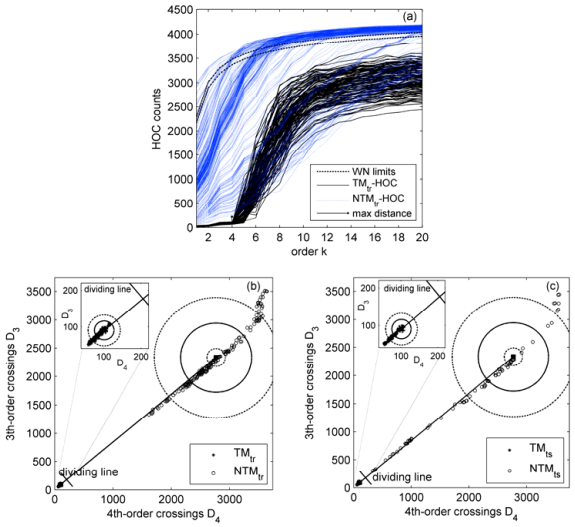

Fig. 5. Results from the algorithm when applied to the 'high length' training dataset. (a) The WN test, (b) the HOC scatter plot of (D_4,D_3) for the TM_{tr} (stars) and NTM_{tr} (circles) segments of the training set, (c) the HOC scatter plot of (D_4,D_3) for the TM_{ts} (stars) and NTM_{ts} (circles) segments of the 'high length' testing set.

The coordinates of the cluster centers, estimated by the FCM technique for the TM and NTM groups, as well as the corresponding distances between the centers of the clusters are the same as the ones mentioned above for the 'low-length' dataset, whereas in the case of the 'high-length' dataset the coordinates (2769.8, 2333.9) and (101.7, 90.0) and a distance of 3486.2 were estimated, respectively.

To further evaluate the performance of the proposed algorithm, two sets of randomly chosen proteins (not included in the training sets) were used. The results are shown in Figs. 4(c) and 5(c) for the 'low-length' and 'high-length' test datasets, respectively. In both subfigures, the members of each group fall within the cluster defined by the corresponding training set. In the upper left corner of Fig. 5(c), a detailed segment of the diagram is presented, illustrating the group of TM segments within the predefined cluster. It is apparent that the increase in the protein segment length has increased the distance between the two clusters. A decrease in the optimum HOC order from 8 to 4 and a smaller 'opened eye' is noticed comparing Figs. 4(a) and 5(a). Nevertheless, the distance between the NTM and TM HOC sequences in both 'opened eyes' is adequate, providing a linear discrimination between the two groups. Furthermore, the method is independent of the TM segment length, ranging from 10 to 30 residues in the abovementioned datasets, as it classifies correctly all the protein segments.

V. Conclusion

A linear discrimination method of TM from NTM segments has been proposed. The method has exhibited prominent performance (100% accuracy) when applied on human protein sequences with single TM segments. In addition, it is simple, accurate and easy to implement. Although it was tested for the case of single TMs, it could be applied on other protein sets including proteins with multiple TM segments, as well as proteins from organisms other than human. In this way, a more extended tool for the classification and identification of TM protein segments could be establish.

VI. References

[1] C. P. Chen, A. Kernytsky and B. Rost, "Transmembrane helix predictions revisited," *Protein Sci.*, vol. 11, pp. 2774–2791, Nov. 2002.

[2] J. M. Cuthbertson, D. A. Doyle1 and M. S. P. Sansom, "Transmembrane helix prediction: a comparative evaluation and analysis," *Protein Eng. Des. Sel.*, vol. 18, pp. 295-308, June 2005.

[3] B. Rost, P. Farisselli and R. Casadi, "Topology prediction for helical transmembrane proteins at 86% accuracy," *Protein Sci.*, vol. 5, pp. 1704–1718, Aug. 1996.

[4] A. L. B. Krogh, G. von Heijne and E. L. L. Sonnhammer, "Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes," *J. Mol. Biol.*, vol. 305, pp. 567–580, 2001.

[5] C. Pasquier and S.J. Hamodrakas, "An hierarchical artificial neural network system for the classification of transmembrane proteins," *Prot. Eng.*, vol. 12, pp. 631-634, Aug. 1999.

[6] G. E. Tusnady, Z. Dosztanyi and I. Simon, "Transmembrane proteins in the Protein Data Bank: identification and classification," *Bioinformatics*, vol. 20, pp. 2964-2972, Nov. 2004.

[7] D. Kihara, T. Shimizu and M. Kanehisa, "Prediction of membrane proteins based on classification of transmembrane segments," *Prot. Eng.*, vol. 11, pp. 961–970, Nov. 1998.

[8] J. Kyte and R. F. Doolittle, "A simple method for displaying the hydropathic character of a protein," *J. Mol. Biol.*, vol. 157, pp. 105–132, 1982.

[9] B. Kedem, *Time series analysis by higher-order crossings*. Piscataway, N.J.: IEEE Press, 1994.

[10] B. Boeckmann, A. Bairoch,, R. Apweiler, M. C. Blatter, A. Estreicher, E. Gasteiger, M. J. Martin, K. Michoud, C. O'Donovan, I. Phan, S. Pilbout and M. Schneider, *Nucl. Acids Res.*, vol. 31, pp. 365-370, 2003, Available: ftp://ftp.ebi.ac.uk/pub/databases/swissprot/special selections/human.seq.gz)

[11] J. C. Bezdek, *Pattern recognition with fuzzy objective function algorithms*. New York: Plenum Press, 1981.

[12] S. Needleman and C. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," J Mol Biol., vol. 48, no. 3, pp. 443-453, Mar. 1970.