An ANFIS approach to transmembrane protein prediction

Hassan B. Kazemian *SMIEEE* and Syed A. Yusuf

***Abstract*—This paper is concerned with transmembrane prediction analysis. Most of novel drug design requires the use of Membrane proteins. Transmembrane protein structure allows pharmaceutical industry to design new drugs based on structural layout. However, laboratory experimental structure determination by X-ray crystallography is difficult to be achieved as the hydrophobic molecules do not crystalize easily. Moreover, the sheer number of proteins demands a computational solution to transmembrane regions identifications. This research therefore presents a novel Adaptive Neural Fuzzy Inference System (ANFIS) approach to predict and analyze of membrane helices in amino acid sequences. The ANFIS technique is implemented to predict membrane helices using sliding window data capturing. The paper uses hydrophobicity and propensity to encode the datasets using the conventional one letter symbol of amino acid residues. The computer simulation results show that the offered ANFIS methodology predicts transmembrane regions with high accuracy for randomly selected proteins.**

I. Introduction

Every cell, whether it is prokaryotic or eukaryotic is surrounded in a thin covering coat named membrane. Membrane proteins are large sets of biological macromolecules and they play a central role in working of the cell. Cellular functions include communication between cells, communication between organelles and cytosol, ion transport, receptors, nutrient transport and links to the extracellular matrix to name a few [1-2]. Most membrane proteins are connected to cell membranes through the transmembrane domain that passes via the membrane lipid/fat bilayer. Transmembrane proteins perform several key tasks such as cell signaling, cell to cell communication, cell recognition and cell adhesion or bond. Their transport ability from outside to inside of cellular organisms makes them a focus of more than half of the drug based research. The parts of proteins which are in contact with the membranes tend to be made of fat-loving (hydrophobic) amino acids, since the membranes of cells are mainly made of fat. It is therefore vital to find transmembrane protein structure and transmembrane regions to be able to design novel drugs. Human Genome project demonstrated that around 30% of proteins are transmembrane and most of the drugs act on transmembrane proteins [3-4].

Initial research in amino acid sequence helices started in 1980's with some improvements in hydrophobicity prediction [5]. The research has gathered momentum since then and the applications of artificial intelligence (AI) and statistical analysis have been conducted to transmembrane proteins. Rost *et al* [6] has applied neural Networks (NN) to prediction of the helical transmembrane proteins. The technique has been used in simple type classification with some improvements in the predictions. There have also been some statistical analysis methods with supervised machine-learning algorithms such as PhdHTM [6] and DAS [7] to prediction of transmembrane protein. For instance, research recently has been carried out in the applications of Support Vector Machine (SVM) and other techniques to transmembrane protein prediction [8]. Hidden Markov Model has also significantly been applied on many occasions to transmembrane protein prediction analysis [9]. The simulation results reveal that HMM outperforms earlier methods both when evaluated on low-resolution topology data and on high-resolution 3D structures. The authors claim that the topology could be correctly predicted for approximately two-thirds of all membrane proteins using HMM. Kazemian *et al* [10] has taken the research further by applying a dual SVM – Genetic Algorithm (GA) schemes to prediction of membrane alpha-helices. The computer simulation results show that the SVM-GA algorithm performs better than most conventional techniques for randomly selected proteins containing single and multiple transmembrane regions. This research takes the membrane alpha helices prediction further by applying a novel ANFIS technique, a method which has never been used in transmembrane protein prediction.

II. An ANFIS Technique

Neural adaptive learning techniques can be utilized to provide a method for a Fuzzy Inference System (FIS) to learn information about a complex dataset of protein sequences. The technique can be used to compute the Member Function (MF) parameters that most optimally suited for the associated FIS to track the given input/output data. The technique constructs an FIS whose MF weights are tuned using a NN based back propagation algorithm. The relevant MF parameters change throughout the training process and the adjustments of these are facilitated by gradient vectors [11-12]. Although fuzzy logic has widely been used in conjunction with neural networks to solve a range of real world problems, the area of transmembrane protein prediction still remains largely unexplored. In the case of multi-variable amino acid chain assessments, pure fuzzy logic based systems may not offer a feasible solution. This is because of high number of inter-dependent variables such as

H. B. Kazemian is with Intelligent Systems Research Centre, School of Computing, London Metropolitan University, Holloway Road, London N7 8DB, UK (e-mail: h.kazemian@londonmet.ac.uk).

S. A. Yusuf is with Institute of Industrial Research (IIR), Faculty of Creative and Cultural Industries, Portsmouth University, UK.

This work was supported by the Emerald Fund under Grant 2009-2S-017. This work was carried out at Intelligent Systems Research Centre, Faculty of Life Sciences and Computing, London Metropolitan University.

978-1-4799-2072-3/14/$31.00 ©2014 IEEE

propensity and hydrophobicity, make the construction of a manual expert guided rule-based system that robustly maps input/output relationship almost an impossible task. In order to overcome the shortcomings of manual knowledge acquisition to create such rule based model, NNs are applied to automatically extract fuzzy rules from the numerical data [13].

This paper uses the hydrophobicity and the propensity schemes to encode the protein sequence datasets. The hydrophobic term describes the likelihood of amino acid residues to exist in a transmembrane domain. The possibility of a given amino acid to be in a transmembrane region can be calculated using the propensity values [14]. These two encoding methods of the hydrophobicity and the propensity are used as a result of an extensive research carried out by Bose et al [14-15] using neural networks and statistical data analysis.

TABLE I
RMS Error Values for a Single Input Variable after one Epoch using ANFIS

```
==================================================
9 ANFIS models, each with one input selected from nine candidates.
__________________________________________________

ANFIS model 1: Steric Parameter --> trn=0.3987, chk=0.4019
ANFIS model 2: Polarisibility --> trn=0.3845, chk=0.4073
ANFIS model 3: Volume--> trn=0.3893, chk=0.4011
ANFIS model 4: Hydrophobicity--> trn=0.3882, chk=0.3739
ANFIS model 5: Isoelectric Point --> trn=0.3819, chk=0.3870
ANFIS model 6: Helix Probability --> trn=0.3929, chk=0.4103
ANFIS model 7: Sheet Probability --> trn=0.3968, chk=0.3906
ANFIS model 8: Propensity --> trn=0.3828, chk=0.3784
ANFIS model 9: Polarity --> trn=0.3969, chk=0.3953
==================================================
```

To ascertain that the two encoding systems are also good enough to produce optimal results using ANFIS, the research was initially carried out by using exhaustive search criteria. For this purpose, other encoding schemes to represent amino acid sequence datasets, such as, steric parameter, polarisability, volume, isoelectric point, helix probability, sheet probability, polarity, and of course hydrophobicity and propensity [15] were individually used. A simple exhaustive search was performed within the available nine inputs to select the inputs that best influence the transmembrane protein segments in amino acid chains. The search mechanism built an ANFIS model for each input variable which was then trained

for one epoch and then the reported error was recorded for the training/checking dataset pairs. The sample data for this search was taken from alpha-helix protein sequences for randomly selected protein chains bearing the nine encoding schemes. The outcome of ANFIS Root Mean Square (RMS) error for one variable is shown in Table 1; the lowest two are hydrophobicity and propensity. Therefore a fixed protein sequence dataset is used in this research to ascertain the prediction accuracies of two of the physic-chemical encodings, propensity and hydrophobicity using ANFIS technique.

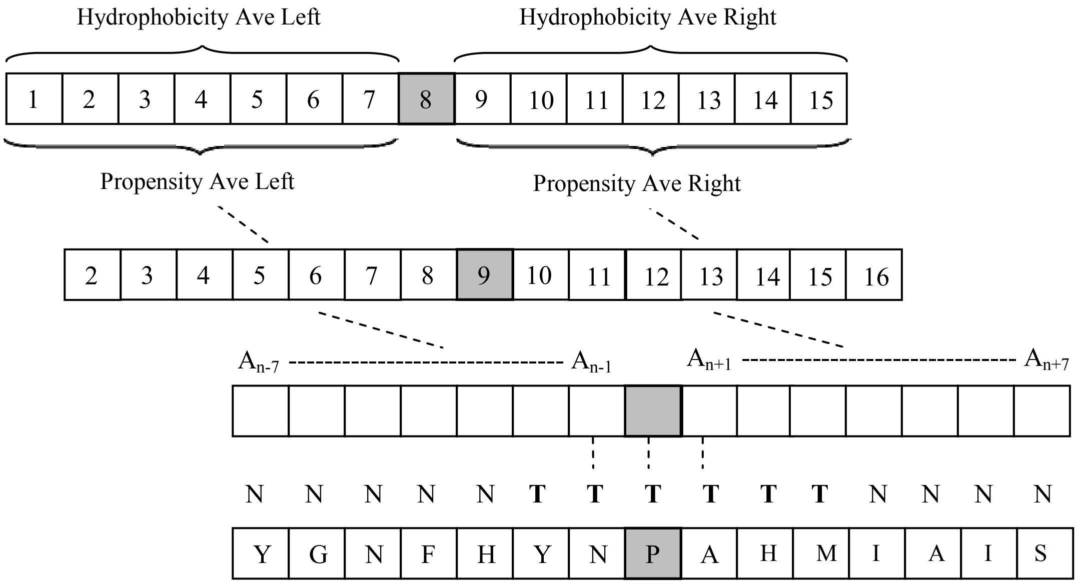


Fig. 1: An example of sliding window based extraction for alpha-helix amino acid sequences 1jgy_L.

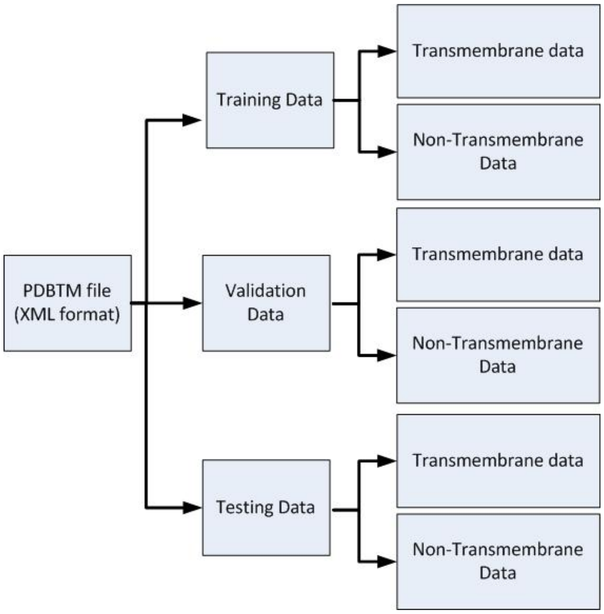


Fig. 2. Data distribution for ANFIS network.

An ANFIS model was setup to forecast alpha-helices by looking at individual amino-acid residue data sets based on a set number of previous trained parametric residues information. The fuzzy rule is provided below:

If H_{avg} is LOW and P_{avg} is AVERAGE and V_{max} is HIGH, then P_i is > HIGH (1)

where H_{avg} and P_{avg} are the average hydrophobicities and propensities, and V_{max} is maximum volume of a set range of amino acid residues immediately before the classified residue P_i at position i, as shown in Fig. 1. Fig. 1 utilizes a prescribed number of moving amino acid residues, in here 15, known as the 'sliding window' to analyze the protein sequences 1jgy_L for the purpose of protein segment identification, where the alpha-helices are continuously found for the middle residue based on the hydrophobic values on both sides of the sliding window [10, 16]. The four variables for hydrophobicity and propensity are 'Left_Hydrophobicity', 'Right_Hydrophobicity', 'Left_Propensity' and 'Right_Propensity'. The ANFIS technique considers the biological and statistical characteristics of the proteins and presents a method that differentiates single and multiple transmembrane segments in amino acid sequences.

III. Data Preparation

The training, checking (validation) and testing of datasets should include a diverse assortment of cases with each spanning a range of input variables. This is necessary to robustly train the fuzzy inference rule based and membership function (MF) weights. The decision of what and how many variables to use are important factors in training and performance of the ANFIS model for prediction of unknown amino acid sequences. Furthermore, dispersion of the data values plays a crucial role in the performance of the network. Incorrect data and extreme outliers within a dataset may induce imprecision on the prediction model. In this project it was decided to categorize the data as shown in Fig. 2. The testing data was selected in a 70-30% manner with the test amino acid chains comprised of the bottom 30% amines from the downloaded global database files. This research selects primary protein sequences in order to model the prediction and classification of transmembrane (TM) sequences. The structural databases such as Protein Data Bank (PDB) are required to train these models to predict membrane spanning regions. 1024 amino acid chains were selected from PDBTM website and 720 amino acid sequences were used for training and 304 were utilized for validation and testing.

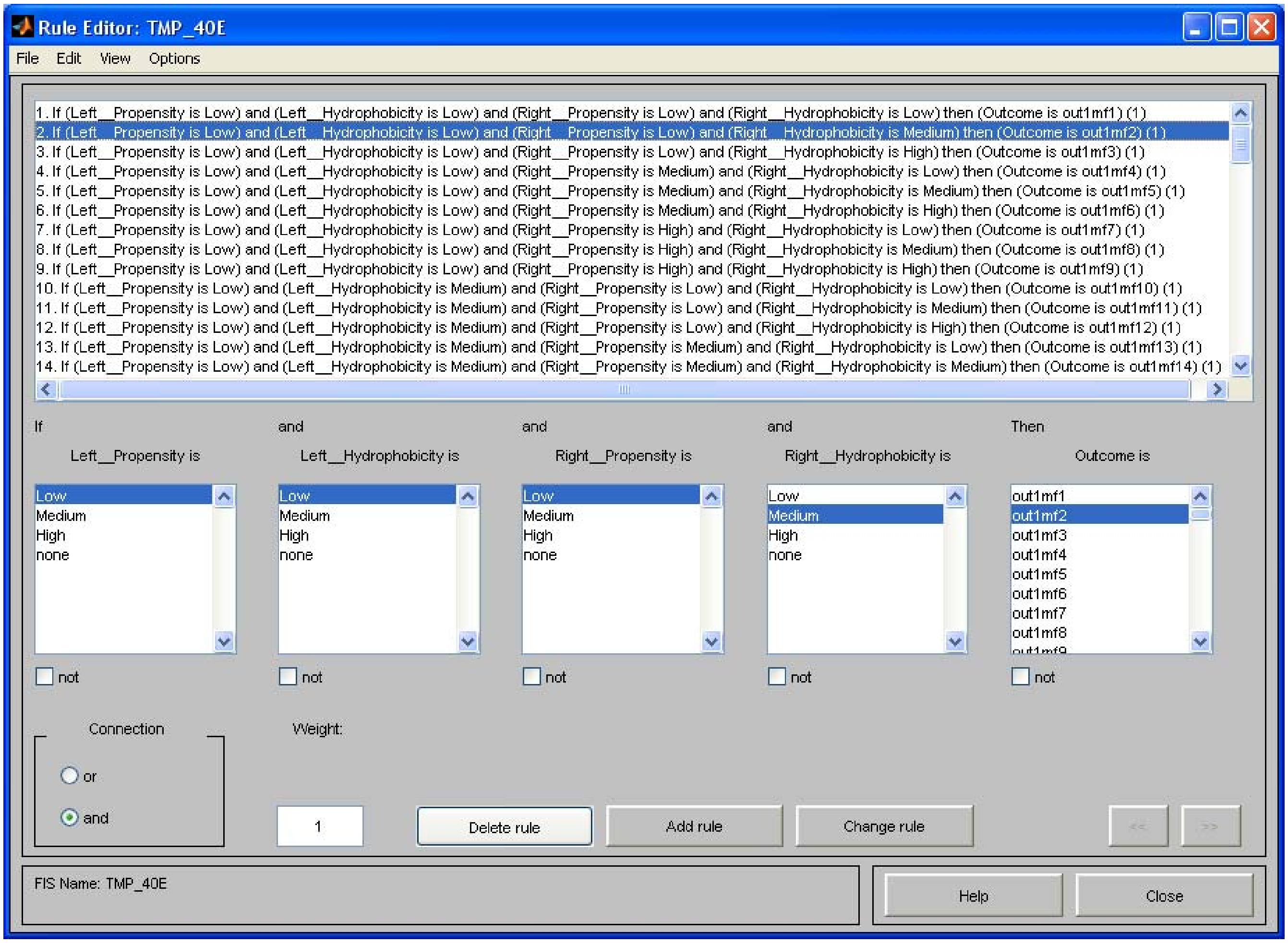
IV. Computer Simulation Results

The data for training and testing purposes of the ANFIS model was loaded using the standard 'ANFISEDIT' command in Matlab [13]. The entire system was then mapped into fuzzy MFs of average hydrophobicity H_{avg}, average propensity P_{avg} and maximum volume V_{max} using a Takagi-Sugeno-Kang fuzzy inference system with neural weight adjustments. Takagi–Sugeno–Kang method is known to complement optimization and adaptive techniques [17].

A. An ANFIS based Transmembrane Prediction

The four variables 'Left_Hydrophobicity', 'Right_Hydrophobicity', 'Left_Propensity' and 'Right_Propensity' shown in Fig. 1 finally mapped to the NN in the ANFIS model. The role of the NN is to tune the corresponding membership functions in order to model the underlying feature space of each middle amino acid sequence with respect to its immediate neighborhood residue as demonstrated in Fig. 1. The proposed ANFIS network produces a comprehensive computer generated fuzzy logic rules using equation (1). Fig. 3 demonstrates a rule based FIS which is obtained using the NN based training to generate the rules. In this scenario the ANFIS technique generates 81 rules, 14 of which are outlined in Fig. 3. Two examples of MFs are further elaborated in Fig. 4. In Fig. 4 (a) the propensity MFs greater than 1 predict a high affinity to be placed in a transmembrane region. The positive values for the hydrophobicity MFs demonstrate an increased possibility for the specific residue to appear in a transmembrane region. The differentiation between transmembrane and non-transmembrane regions are further demonstrated using amino acid sequence 2ZXW Chain D in Fig. 5. The figure outlines 145 amino acid sequence protein and the residues from 76 to 99 fall within transmembrane regions. In Fig. 5, the residues in the positive waveform area imply hydrophobicity and the residues greater than 1 signify propensity. The transmembrane region is called 'AA Chain' in the figure.

The mutual effect of each of these variables can be further analyzed with surface mapping of the input variables in Figs. 6 and 7. The surface plot of the rule-based fuzzy in Fig. 6 from Left-to-Right comparison of the propensity input variables, show a gradual increase of amino acid residues from low propensity to high propensity moving from non-transmembrane to transmembrane region. This is because the propensity MFs greater than 1 has likelihood to be in the transmembrane region. The surface plot of the rule-based fuzzy in Fig. 7 shows the output variables of hydrophobicity from negative to positive values, demonstrating that amino acid sequence residues changing from non-transmembrane to transmembrane segments. An ascent from 'Right_Hydrophobicity' to 'Left_Hydrophobicity' indicates that the amino acid residues are moving from non-transmembrane to transmembrane regions.



1. If (Left__Propensity is Low) and (Left__Hydrophobicity is Low) and (Right__Propensity is Low) and (Right__Hydrophobicity is Low) then (Outcome is out1mf1) (1)
2. If (Left__Propensity is Low) and (Left__Hydrophobicity is Low) and (Right__Propensity is Low) and (Right__Hydrophobicity is Medium) then (Outcome is out1mf2) (1)
3. If (Left__Propensity is Low) and (Left__Hydrophobicity is Low) and (Right__Propensity is Low) and (Right__Hydrophobicity is High) then (Outcome is out1mf3) (1)
4. If (Left__Propensity is Low) and (Left__Hydrophobicity is Low) and (Right__Propensity is Medium) and (Right__Hydrophobicity is Low) then (Outcome is out1mf4) (1)
5. If (Left__Propensity is Low) and (Left__Hydrophobicity is Low) and (Right__Propensity is Medium) and (Right__Hydrophobicity is Medium) then (Outcome is out1mf5) (1)
6. If (Left__Propensity is Low) and (Left__Hydrophobicity is Low) and (Right__Propensity is Medium) and (Right__Hydrophobicity is High) then (Outcome is out1mf6) (1)
7. If (Left__Propensity is Low) and (Left__Hydrophobicity is Low) and (Right__Propensity is High) and (Right__Hydrophobicity is Low) then (Outcome is out1mf7) (1)
8. If (Left__Propensity is Low) and (Left__Hydrophobicity is Low) and (Right__Propensity is High) and (Right__Hydrophobicity is Medium) then (Outcome is out1mf8) (1)
9. If (Left__Propensity is Low) and (Left__Hydrophobicity is Low) and (Right__Propensity is High) and (Right__Hydrophobicity is High) then (Outcome is out1mf9) (1)
10. If (Left__Propensity is Low) and (Left__Hydrophobicity is Medium) and (Right__Propensity is Low) and (Right__Hydrophobicity is Low) then (Outcome is out1mf10) (1)
11. If (Left__Propensity is Low) and (Left__Hydrophobicity is Medium) and (Right__Propensity is Low) and (Right__Hydrophobicity is Medium) then (Outcome is out1mf11) (1)
12. If (Left__Propensity is Low) and (Left__Hydrophobicity is Medium) and (Right__Propensity is Low) and (Right__Hydrophobicity is High) then (Outcome is out1mf12) (1)
13. If (Left__Propensity is Low) and (Left__Hydrophobicity is Medium) and (Right__Propensity is Medium) and (Right__Hydrophobicity is Low) then (Outcome is out1mf13) (1)
14. If (Left__Propensity is Low) and (Left__Hydrophobicity is Medium) and (Right__Propensity is Medium) and (Right__Hydrophobicity is Medium) then (Outcome is out1mf14) (1)

If	and	and	and	Then
Left__Propensity is	Left__Hydrophobicity is	Right__Propensity is	Right__Hydrophobicity is	Outcome is
Low Medium High none	**Low** Medium High none	**Low** Medium High none	Low **Medium** High none	out1mf1 **out1mf2** out1mf3 out1mf4 out1mf5 out1mf6 out1mf7 out1mf8 out1mf9
□ not	□ not	□ not	□ not	□ not

Connection
○ or
● and

Weight: 1

FIS Name: TMP_40E

Fig. 3. Rule-based Fuzzy Inference System obtained as a result of NN based training.

Table 2 outlines 18 randomly selected proteins from PDBTM datasets. The prediction rates of these proteins are shown for non-transmembrane, transmembrane and the overall accuracy using the ANFIS algorithm. The results are based on average encoding schemes of hydrophobicity and propensity. The average transmembrane protein prediction accuracy is 87.28% and the average non-transmembrane protein forecast is 64.05%. The overall average prediction accuracy is 83.21%. For transmembrane prediction, the lowest performance is for the amino acid sequence 3a0h (Chain J) of 46.67%, whereas the highest prediction accuracy of 100% is obtained for the amino acid sequences 3abk (Chain D) and 3abk (Chain G). The ANFIS method therefore presents the best outcome for transmembrane prediction (87.28%) for the helical sequences. The results are further shown in Fig. 8 using bar chart. Fig. 8 demonstrates transmembrane domain predictions for the lowest performance of the amino acid sequence 3a0h (Chain J) and the highest performance of the amino acid sequences 3abk (Chain D) and 3abk (Chain G). The overall average forecast accuracy is again 83.21% in Fig. 8.

V. Conclusion

This paper discusses the applications of a novel ANFIS technique to prediction and analysis of membrane alpha-helices. The proposed model applies a neural adaptive learning technique to provide a method for a Fuzzy Inference System (FIS) to analyze a complex dataset of protein sequences. Single and multiple transmembrane regions in amino acid sequences are differentiated by the ANFIS technique by considering the biological and statistical characteristics of the proteins. Training and testing dataset of 1024 global single and multiple transmembrane regions were selected from Protein Data Bank (PDBTM) database and utilized to optimize the proposed ANFIS model and the computer simulation has been carried out using a customized training and testing database of alpha-helix transmembrane regions from a diverse range of organisms. The computer simulation results using 18 randomly selected proteins reveal that the transmembrane protein accuracy of 87.28% can be obtained using the proposed ANFIS technique. The ANFIS methodology matches one of the best performing algorithms and presenting an alternative approach to transmembrane helix prediction. Further research is required to compare these results with other conventional techniques, such as Support Vector Machine and Hidden Markov Model using other datasets like Swiss-Prot and UniProt.

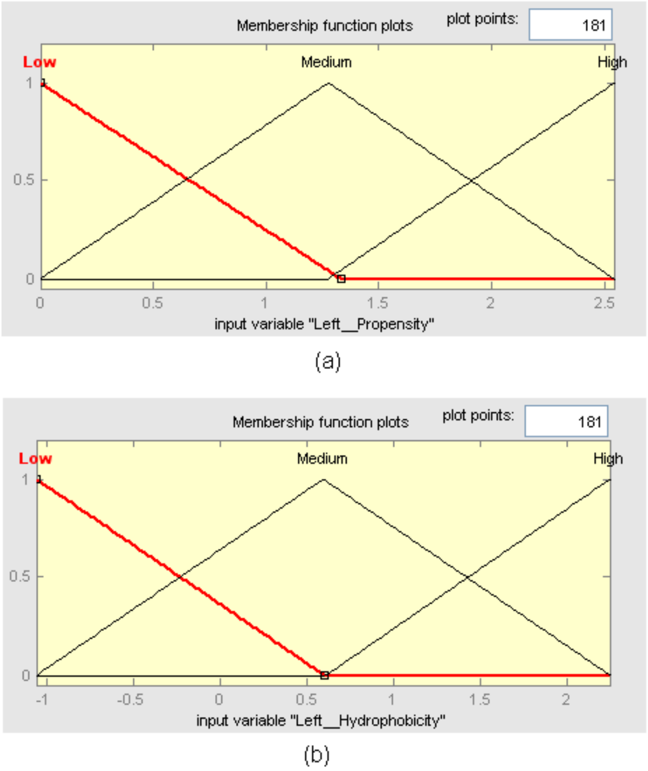


Fig. 4. (a) and (b) are two triangular MFs tuned using the NN based training.

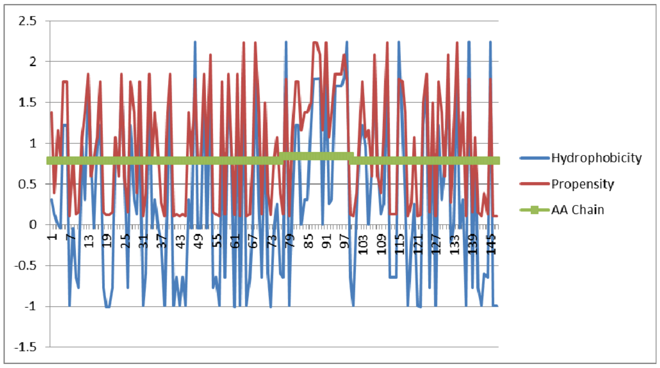


Fig. 5. To differentiate between transmembrane and non-transmembrane regions by hydrophobicity and propensity encodings.

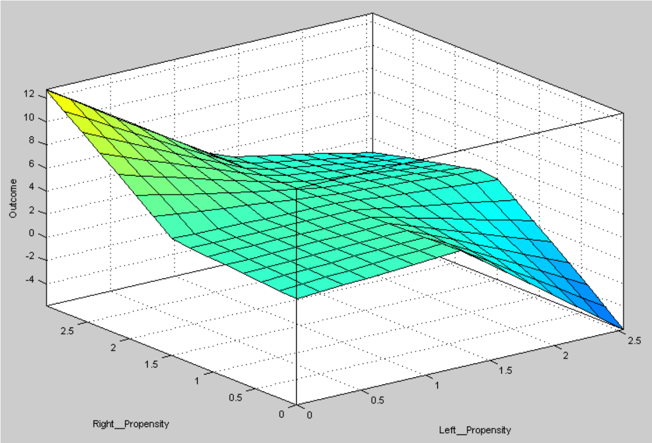


Figure 6: Ascent from 'Left_Propensity' to 'Right_Propensity' indicates the amino acid residues moving from non-transmembrane to transmembrane regions.

ACKNOWLEDGMENT

This work was part funded by the Emerald Fund UK under grant reference number 2009-2S-017. This work was carried out at Intelligent Systems Research Centre, Faculty of Life Sciences and Computing, London Metropolitan University, UK.

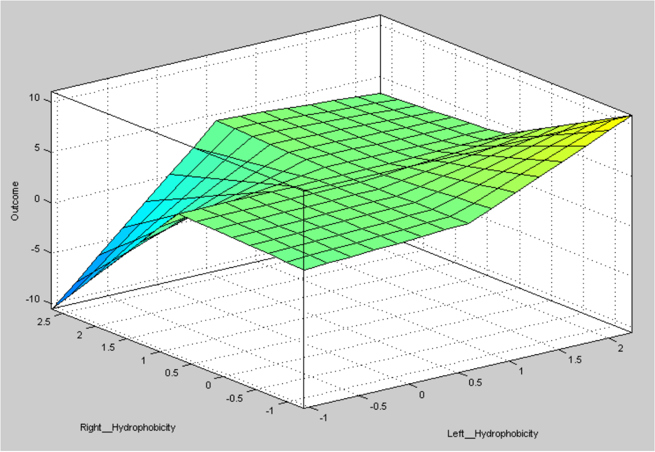


Fig. 7. Ascent from 'Right_Hydrophobicity' to 'Left_Hydrophobicity' indicates that the amino acid residues moving from non-transmembrane to transmembrane regions.

TABLE II
RESULTS OF 18 RANDOMLY SELECTED HELICAL SEQUENCES USING ANFIS TECHNIQUE

Protein	Algorithm	Overall Accuracy	Non-TM prediction accuracy	TM prediction accuracy
2a06_C	ANFIS	73.4417	60.2273	85.4922
2a06_E	ANFIS	86.5591	25	95.679
2a06_G	ANFIS	84.507	52.381	98
3a0b_A	ANFIS	75.4491	52.439	82.9365
3a0b_B	ANFIS	85.9833	67.0455	90.2564
3a0b_E	ANFIS	80.8219	64.2857	84.7458
3a0h_F	ANFIS	58.8235	52.9412	64.7059
3a0h_H	ANFIS	90.7407	92.8571	90
3a0h_J	ANFIS	70	93.3333	46.6667
2a79_B	ANFIS	81.1861	57.3529	85.0356
2a79_J	ANFIS	81.1861	57.3529	85.0356
2a79_N	ANFIS	81.1861	57.3529	85.0356
3abk_D	ANFIS	97.0803	77.7778	100
3abk_B	ANFIS	90.3226	69.697	94.0217
3abk_G	ANFIS	86.6667	44.4444	100
3a3y_A	ANFIS	83.4971	52.9703	91.0539
3a3y_B	ANFIS	93.5593	80.9524	94.5255
3a3y_G	ANFIS	96.875	94.4444	97.8261
Average		83.21587	64.04751	87.27869

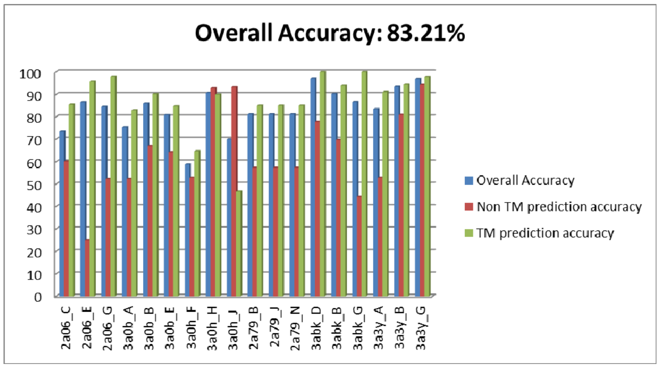


Fig. 8. Bar chart results for 18 randomly selected helical sequences using ANFIS network.

References

[1] J-J. Lacapère, *Membrane Protein Structure Determination: Methods and Protocols: 654 (Methods in Molecular Biology)*, Pub: Humana Press; 2010 ed., ISBN-10: 1607617617, ISBN-13: 978-1607617617, Aug. 2010.

[2] D. Frishman, *Structural Bioinformatics of Membrane Proteins*, ISBN-10: 3709100445, ISBN-13: 978-3709100448, Springer, 2010 ed., Jun 29th 2010.

[3] E. Wallin and G. von Heijne, "Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms," *Protein Sci.* in press, vol. 7, no. 4, pp. 1029-1038, 1998.

[4] G. C. Terstappen and A. Reggiani, "In silico research in drug discovery," *Trends Pharmacol Sci,* vol. 22, no. 1, pp. 23-26, 2001.

[5] J. Kyte and R. F. Doolittle, "A Simple Method for Displaying the Hydropathic Character of a Protein," *Journal of Molecular Biology*, vol. 157, pp. 105-132, 1982.

[6] B. Rost, P. Fariselli, and R. Casadio, "Topology prediction for helical transmembrane proteins at 86% accuracy," *Protein Science*, vol. 5, pp. 1704–1718, 1996.

[7] M. Cserzo, E. Wallin, I. Simon, G. von Heijne, and A. Elofsson, "Prediction of transmembrane alpha-helices in prokaryotic membrane proteins: The dense alignment surface method," *Protein Engineering*, vol. 10, pp. 673–676, 1997.

[8] D. Frishman, *Structural bioinformatics of membrane proteins*, Wien: Springer, 2010.

[9] H. Viklund, A. Elofsson, "Best α-helical transmembrane protein topology predictions are achieved using hidden Markov models and evolutionary information," DOI: 10.1110/ps.04625404, *Protein Science* [Online], vol. 13, issue 7, Jan 2009.

[10] H. B. Kazemian, K. White, D. Palmer-Brown, and S. A. Yusuf, "Applications of Evolutionary SVM to Prediction of Membrane Alpha-Helices," *Expert Systems With Applications*, Elsevier, DOI: 10.1016/j.eswa.2012.12.049, vol. 40, issue. 9, pp. 3412–3420, July 2013.

[11] L. H. Tsoukalas, R. E. Uhrig, L. A. Zadeh, *Fuzzy and Neural Approaches in Engineering (Adaptive and Learning Systems for Signal Processing, Communications and Control Series)*, Pub: Wiley-Blackwell,ISBN-10: 0471160032, ISBN-13: 978-0471160038, Feb 17th 1997.

[12] M. A. M. Basri, *Medical Image Classification and Symptoms Detection Using Neuro Fuzzy: ANFIS Based Classification and Tumor Detection of The Brain Medical Image*, Pub: LAP LAMBERT Academic Publishing, ISBN-10: 3846583545, ISBN-13: 978-3846583548, Jan 31st 2012.

[13] ANFIS Matlab toolbox. Available: http://www.mathworks.com.

[14] S. K. Bose, A. Browne, H. B. Kazemian, and K. White, *Classifying Membrane Proteins in the Proteome by Using Artificial Neural Networks Based on the Preferential Parameters of Amino Acids*, J. A. Tenreiro Machado, B. Patkai & I. J. Rudas (Eds.), Intelligent Engineering Systems and Computational Cybernetics, Springer, pp. 63-71, 2009.

[15] S. K. Bose, *The use of neural networks to identify and analyze membrane proteins in the proteome*, PhD thesis, London Metropolitan University, 2006.

[16] H. B. Kazemian, S. A. Yusuf, K. White, "Signal peptide discrimination and cleavage site identification using SVM and NN," *Computers in Biology and Medicine*, Elsevier, DOI: 10.1016/j.compbiomed.2013.11.017, vol. 45, pp. 98–110, Feb 1st 2014.

[17] H. B. Kazemian and K. Ouazzane, "Neuro-fuzzy approach to video transmission over ZigBee," *Neurocomputing Journal*, DOI: http://dx.doi.org/10.1016/j.neucom.2012.10.006, vol. 104, pp. 127-137, 15 March 15th 2013.