NN approach and its comparison with NN-SVM to beta-barrel prediction

Hassan Kazemian* (h.kazemian@londonmet.ac.uk), Syed Adnan Yusuf (adnan.yusuf@gmail.com), Kenneth White* (kenneth.white@londonmet.ac.uk), Cedric Maxime Grimaldi* (rx8879@gmail.com)**

London Metropolitan University*
STS Defence Limited**

Abstract:

This paper is concerned with applications of a dual Neural Network (NN) and Support Vector Machine (SVM) to prediction and analysis of beta barrel transmembrane proteins. The prediction and analysis of beta barrel proteins usually offer a host of challenges to the research community, because of their low presence in genomes. Current beta barrel prediction methodologies present intermittent misclassifications resulting in mismatch in the number of membrane spanning regions within amino-acid sequences.

To address the problem, this research embarks upon a NN technique and its comparison with hybrid-two-level NN-SVM methodology to classify inter-class and intra-class transitions to predict the number and range of beta membrane spanning regions. The methodology utilizes a sliding-window-based feature extraction to train two different class transitions entitled symmetric and asymmetric models. In symmetric modelling, the NN and SVM frameworks train for sliding window over the same intra-class areas such as inner-to-inner, membrane(beta)-to-membrane and outer-to-outer. In contrast, the asymmetric transition trains a NN-SVM classifier for inter-class transition such as outer-to-membrane (beta) and membrane (beta)-to-inner, inner-to-membrane and membrane-to-outer. For the NN and NN-SVM to generate robust outcomes, the prediction methodologies are analysed by jack-knife tests and single protein tests. The computer simulation results demonstrate a significant impact and a superior performance of NN-SVM tests with a 5 residue overlap for signal protein over NN with and without redundant proteins for prediction of transmembrane beta barrel spanning regions.

Keywords: Beta barrel prediction, protein analysis, Neural Networks, Support Vector Machine.

1 Introduction

Integral membrane proteins are a type of transmembrane proteins that are permanently attached to the membrane. According to their tertiary structure, they can be divided into alpha-helical and beta-barrel proteins (Rangwala and Karypis, 2010). The structural properties of alpha helix transmembrane domains are well established including the amino acid composition, spatial interaction between helical spaces and helical bundle packing (Russell and Cohn, 2012). For transmembrane beta-barrels, the integral protein segments are known to occur in outer membranes of bacteria, mitochondria and chloroplasts (Neupert and Lill, 1992). Beta barrel structures are known to be characterised by the number of anti-parallel beta strands and by the number representing the angle of inclination of the beta strands against the barrel axis. The structural information of beta-barrel membrane proteins lag behind alpha helical structures. For instance, in helical membrane proteins, the folding and assembly of transmembrane segments can be identified via computational analysis of sequential motifs (Kall, Krogh and Sonnhammer, 2004). For beta barrels, very little information is available for sequence motifs or their role in protein stability maintenance and function (Stillman, 1995; Arai, Ikeda and Shimizu, 2003).

Recent research has revealed the versatility and ever-presence of beta barrel proteins with the distribution spanning on many families of bacteria as well as eukaryotes. The outer membrane proteins of bacteria make transmembrane beta-barrels with evenly occurring beta strands ranging from 8 to 22 and a sheer number of 8 to 24 existing as monomers or oligomers (Neupert and Lill, 1992; Bagos, Liakopoulos, Spyropoulos and Hamodrakas, 2004; Bagos, Liakopoulos and Hamodrakas, 2005). There is a wide range of functions of beta barrel membrane proteins including transport of active ions, passive intake of nutrients, membrane anchoring, membrane-bound enzyme identification as well as providing defence against attacking proteins. Furthermore, research is concentrated on the folding process in membranes, underlying crystallisation, limited structural diversity observation and various channel engineering options.

Various types of stable transmembrane beta sheets are limited by the non-polar core of the bi-layer membrane. All proteins form structural folds that are almost cylindrical around the bi-layer and expose the mainly non-polar side chains to the transmembrane. This criterion is satisfied only by the beta sheet structure in which the beta transmembrane strands are laterally hydrogen bonded in a cylindrical fashion (Fariselli, Finelli, Rossi, Amico, Zauli, Martelli and Casadio, 2005).

Due to the underlying infrastructure, a beta barrel protein consists of a beta sheet that twists and surrounds in a closed coil form where the first strand forms a hydrogen bond to the last. An estimated 2 – 3 % of gram-negative bacteria proteins contain beta barrel encoding. Currently, there are less than 20 known 3D beta barrel structures whereas the genomic databases contain thousands of beta proteins belonging to dozens of beta families. Schulz (2000 and 2002) states a set of rules to identify the structural features of all known beta barrel structures.

Despite having structural differences, the composition of the lipid-exposed surface in beta barrel proteins is similar to its helical counterparts. The transmembrane bi-layers generally contain an abundance of phenylalanine, tyrosine, tryptophan, valine and leucine whereas the polar and charged residues are predominantly excluded. The bi-layer interface contains a large presence of aromatic residues which makes about 40% of the lipid-exposed amino-acids.

According to Protein Data bank of Transmembrane Proteins (PDBTM), as of 2nd Oct 2015, the database of protein structures currently contains 113251 proteins in the Protein Data Bank (PDB) web portals. Out of this, only 2599 are transmembrane domains with just 321 containing beta barrel sheets (the Institute of Enzymology, 2015). The low presence of beta-barrels presents a substantial challenge to currently existing artificial intelligence prediction algorithms such as NNs. Lack of non-redundant data further reduces the size of data available for training by reducing the number of non-redundant proteins obtained by Cluster Database at High Identity with Tolerance (CD-HIT) algorithm at 40% cut-off to be only 54. CD-HIT takes a protein sequence database as input and produces a set of 'non-redundant' representative sequences as output (Li and Godzik, 2006).

A number of researchers have developed beta-barrel identification algorithms using either composition or graph theory or evolutionary couplings (Wimley, 2001; Tran, Chassignet, Sheikh and Steyaert, 2012; Hayat, Sander, Elofsson and Marks, 2014; Hayat, Sander, Marks and Elofsson, 2015). Other researchers used various machine learning techniques, such as Hidden Markov Model (HMM), Bayesian networks, Genetic Algorithm (GA) and SVM (Bigelow, Petrey, Liu, Przybylski and Rost, 2004; Taylor, Toseland, Attwood and Flower, 2006; Zou, Wang, Wang and Hu, 2010; Singh, Goodman, Walter, Helms and Hayat, 2011; Hayat and Elofsson, 2012) to prediction analysis of beta barrel. The prediction methodologies lack in terms of intermittent misclassifications that result in mismatch in the number of membrane spanning regions within amino-acid sequences and therefore the prediction of beta barrel proteins generally poses a range of challenges to the research community.

One of the most encouraging results that has been obtained in applications of machine learning techniques to transmembrane proteins was, the application of SVM-GA to alpha helices where the overall outcomes were published in 2013 (Kazemian, White, Palmer-Brown and Yusuf, 2013). Through a future research, a hybrid NN and fuzzy logic technique entitled Adaptive

Neural Fuzzy Inference System was also applied to predict and analyse membrane helices in amino acid sequences which produced a comparable results to using SVM-GA (Kazemian and Yusuf, 2014). In general, SVM is known to model problems with a smaller sample size. This makes the SVM an appropriate technique for beta-barrel prediction problems where the modelling is undermined by problems of a smaller database. Furthermore, Levenberg-Marquardt algorithm is perceived as one of the most effective method for training NN. The Levenberg-Marquardt training algorithm is fast, but it is generally more demanding in terms of memory. This paper consequently applies a new hybrid-two-level cascaded NN-SVM methodology to beta barrel prediction. The methodology utilizes generic feed-forward NNs and SVM with a sliding-window-based feature extraction to train two different class transitions namely termed as symmetric and asymmetric models. Levenberg-Marquardt algorithm is utilised as a training method for NN. The methodology initially modelled two unique aspects of the sliding window operation in a hybrid operation including the modelling of same class modelling such as inside-to-inside as well as the inter-class transition when the middle residue of the sliding window moved from one class to another. The overall performance was evaluated over single protein sequences by pasting and feeding whole sequences to the modelling routine as well as via jack-knife-based testing where each protein was evaluated against a model trained over the remaining dataset. The underlying notion is to generate two different neural outputs to be combined using a weighted classification integration technique. The scores δ_A^i and δ_S^i thus obtained for sliding window instance i for the asymmetric and symmetric models respectively are combined based upon their weighted proximity to the respective outputs to obtain a combined transmembrane score for each amino-acid residue to belong to membrane or non-transmembrane classes.

Section 2 describes beta barrel feature extraction and two examples of intra-class and inter-class sliding windows. Section 3 outlines a dual model neural architecture for the transmembrane and non-transmembrane classifications, Levenberg-Marquardt sliding window training, and NN prediction results analysis for intra-class and inter-class for amino-acid prediction. Section 4 explains the SVM technique for the dual transmembrane and non-transmembrane classifications. This section continues discussing the NN results for comparison with NN-SVM results for jack-knife and signal protein tests. Finally, section 5 concludes the overall outcome for the prediction of beta barrel transmembrane protein.

2 Prediction of beta-barrel using intra-class and inter-class sliding windows

The beta-barrel prediction problem is generally divided into three unique domains (Reynolds and Kall, Riffle, Bilmes, 2008):

1. Discrimination of a protein sequence to actually contain transmembrane (TM) segments
2. The prediction of TM beta-barrel segments
3. Detection of number of membrane spanning regions

The first problem addresses a dual classification scenario, which is not addressed in this paper. The second point states a problem where prediction is to ascertain TM beta-barrel segments present within an amino-acid sequence. The final part models an algorithm's ability to accurately identify number of membrane spanning regions detected in part 2.

2.1 Feature extraction

The feature extraction technique for both NN and SVM-based classification differs according to the objectives of classification. This document reports two different feature extraction techniques to model two aspects of sliding window operation on a membrane protein sequence. Figure 1 shows an example for the symmetrical sliding window using SVM feature extraction technique with a window size of 30 residue, with the residues distributed evenly both upstream and downstream of the sliding window. The extraction technique models same-state transitions so that the window moves in only outer (O), Transmembrane (B) or inner (I) protein segments. For example, the technique extracts a sequence of features with each sample belonging to +15 to +15 length sliding window from left to right in outer (O) protein segments. The figure demonstrates a single training instance extraction from the amino-acid sequence of the *E.coli* outer membrane enzyme PagP (PDB accession no. 1mm4) shown in Figure 3.

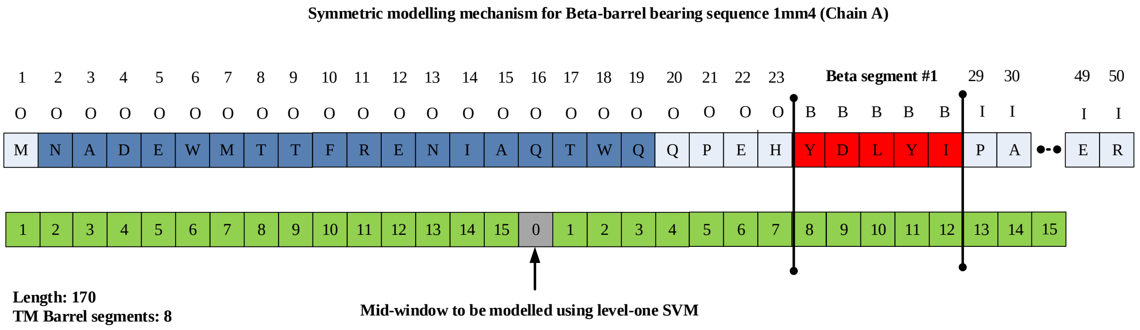

Symmetric modelling mechanism for Beta-barrel bearing sequence 1mm4 (Chain A)
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
Beta segment #1
29 30
49 50
O O O O O O O O O O O O O O O O O O O O O O O
B B B B B
I I
I I
M N A D E W M T T F R E N I A Q T W Q Q P E H Y D L Y I P A E R
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Length: 170
TM Barrel segments: 8
Mid-window to be modelled using level-one SVM

Figure 1: SVM-based feature extraction using symmetrical intra-class sliding window.

Figure 2 shows an asymmetric sliding window of feature extraction based on propensity-based values. For example, the technique extracts a sequence of features with each sample belonging to -3 to +15 length sliding window from left to right. In another example, the technique extracts a sequence of features with each sample from -15 to +3 length sliding window. Figure 2 demonstrates a single training instance extraction from the PagP amino-acid sequence shown in Figure 3. The feature values outlined in Figure 1 and Figure 2 are fed to a regular kernel-based system to classify each test sliding window to assign a classification of 0 to 1 for non-transmembrane and beta-barrel segments.

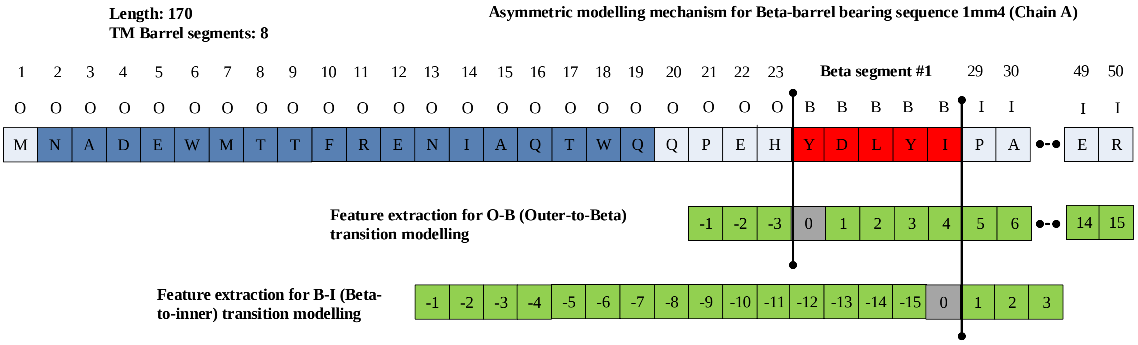

Length: 170
TM Barrel segments: 8
Asymmetric modelling mechanism for Beta-barrel bearing sequence 1mm4 (Chain A)
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
Beta segment #1
29 30 49 50
O O O O O O O O O O O O O O O O O O O O O O O B B B B B I I I I
M N A D E W M T T F R E N I A Q T W Q Q P E H Y D L Y I P A E R
Feature extraction for O-B (Outer-to-Beta) transition modelling
-1 -2 -3 0 1 2 3 4 5 6 14 15
Feature extraction for B-I (Beta-to-inner) transition modelling
-1 -2 -3 -4 -5 -6 -7 -8 -9 -10 -11 -12 -13 -14 -15 0 1 2 3

Figure 2: SVM-based feature extraction using asymmetrical inter-class sliding window.

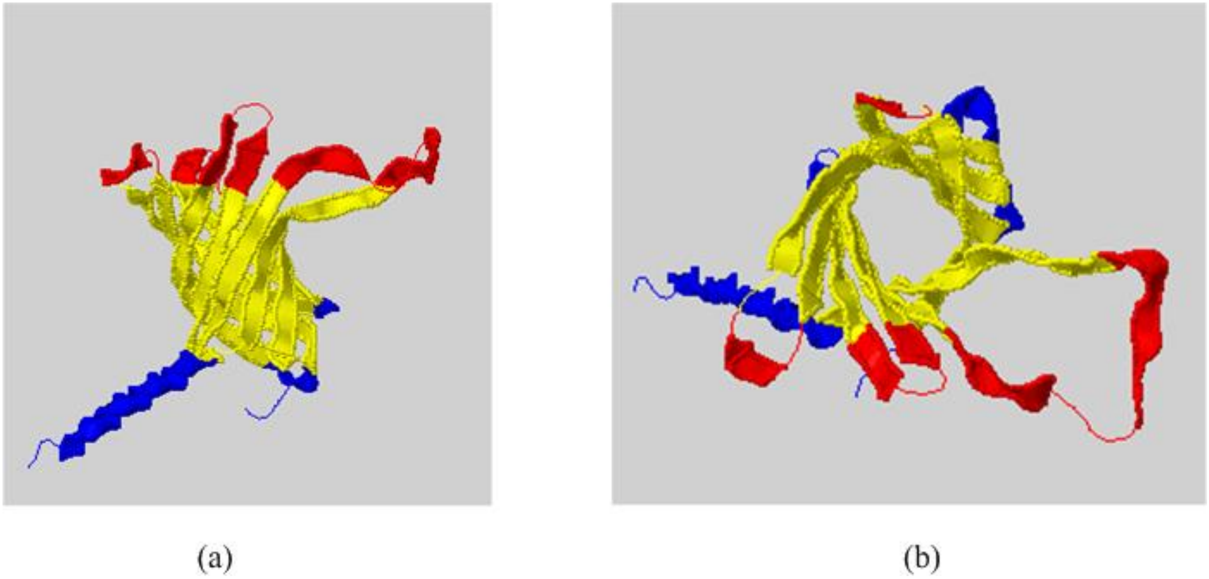


Figure 3: Structure of *E.coli* outer membrane enzyme PagP (PDB accession no. 1mm4), which comprises 10% helical (1 helix; 18 residues) and 44% beta sheet (10 strands; 76 residues) structural elements.

3 Dual model neural architecture for the TM-NTM classification

A conventional neural architecture with a back-propagation algorithm known as multi-layer feed-forward network is used in this research. In a multi-layer architecture, each neuron is fed with R inputs and each input is weighted with an appropriate *w*. The sum of weighted inputs and bias acts as the input to the transfer function where the most commonly used transfer function is the “Log Sigmoid” function (Hagan, Demuth, Beale and De Jesús, 2014). As the feed-forward networks tend to have one or more layers of sigmoid neurons, this study's architecture initially contains a single hidden layer architecture followed by an output layer of linear neurons representing different topological classes in a protein sequence. The two-layer feed-forward architecture takes multiple instances of a sliding window as an input feature fed to the input neurons against a set of target vectors with each output value representing the amino-acid present at the middle of the sliding window. The objective is to predict each amino-acid residue within an amino-acid protein sequence to belong to various topological classes including beta-barrel transmembrane segments, inside and outside sequence locations.

3.1 Levenberg-Marquardt sliding window training for a dual-layer NN in amino-acid protein prediction

The methodology employs two unique training algorithms termed as the "scaled conjugate" and the "Levenberg-Marquardt" algorithms. The initial method is relatively slow but memory efficient and can therefore be used to analyse complex functions. The later algorithm of Levenberg-Marquardt is faster, but it is generally more demanding in terms of memory. However, the Levenberg-Marquardt is used as the ultimate method to evaluate the efficiency of the underlying neural models (Yu and Wilamowski, 2010).

The Levenberg-Marquardt method employs search direction solution of a linear set of equations as follows:

$$(J(x_k)^T.J(x_k) + \lambda_k I)d_k = \left(-J(x_k)^T.F(x_k)\right) \qquad (1)$$

Where J stands for Jacobian matrix, function of x_k the input vector, and I is the identity matrix. In the above equation, the direction d_k is similar to Gauss-Newton method if $\lambda_k = 0$. For $\lambda_k \to \infty$, d_k approaches the steepest descent direction with the magnitude approaching zero implying the principle that for a reasonably large λ_k, the term $F(x_k + d_k) < F(x_k)$ is true. The variable λ_k can therefore be regulated to ensure decline of the second order terms are encountered that generally tend to restrict the efficiency of Gauss-Newton method. The Levenberg-Marquardt equation therefore utilizes a hybrid search direction methodology between Gauss-Newton and steepest descent direction.

Levenberg-Marquardt algorithm is considered the most efficient method for training artificial NN though its computational complexity and difficulty is marred by difficulties in calculating Hessian matrices, matrix inversion and region computation (Yu and Wilamowski, 2010). Therefore, similar to quasi-Newton methods, the Levenberg-Marquardt algorithm achieves a second order training speed without the computation of Hessian matrix. The matrix is later approximated once the performance function forms the sum of squares typical to regular feed-forward network training as follows:

$$H = J^T . J \tag{2}$$

The gradient can be obtained as:

$$g = J^T . e \tag{3}$$

The Jacobian matrix contains the first derivatives of neural feedback errors from the network weights and biases, whereas e is a network error vector. The approximation used to calculate the Hessian matrix is given as follows:

$$y_{i+1} = y_i - [J^T J + \mu I]^{-1} J^T e \tag{4}$$

Where for $\mu \rightarrow \infty$ it becomes a gradient-descent equation with a small step size. The value of μ is increased or decreased based upon after each successful or tentative step in order to reduce or increase the performance function respectively. The application of Levenberg-Marquardt training compared to Scaled Conjugate method is presented in the next section.

3.2 Intra-class and Inter-class sliding window for a dual layer NN in beta barrel regions prediction.

For the current objective of prediction of beta-barrel proteins, the defined NN has a single input, hidden and output layer. The feature extraction methodology elaborated in Section 2 exhibits a feature sliding window of 30 residue length that encodes each input amino acid sequence and the identification is performed on the structural state of the central residue in the window. Each residue position is encoded via standard propensity-based encoding. The input layer consists of two unique modelling representations termed as intra-class and inter-class modelling given below:

- Intra-class: Those instances of the sliding window are stored in a sliding window training matrix where the current middle window residue i and $i+1$ position belong to the same class i.e. outside (1), Membrane (B) or inside (2); please refer to Figure 1. The output layer therefore consists of three units represented by 11, BB and 22 for the three different intra-class transitions respectively. The output coding is a digit value of 10 for 'class 11', 20 for 'class BB' and 30 for 'class 22'.
- Inter-class: Only those instances are stored in a sliding window training matrix where the current middle window residue i and $i+1$ position belong to different classes; please refer to Figure 2. The output layer therefore represents units as 1B, B1, 2B and B2 for four different inter-class transitions, encoded via 10, 20, 30 and 40 digital values respectively.

Thus, the input layer R = 30 X 1 input units representing 30 residues in the sliding window with each containing a single propensity encoding. Once the input and target matrices are obtained for the non-redundant dataset, the encoding is fed into the network shown in Figure 4 for training. As discussed earlier, the problem of secondary protein structure prediction can be taken as a pattern recognition problem. The network is trained to recognise the associated class of the central residue based upon the residues observed in the particular sliding window instance.

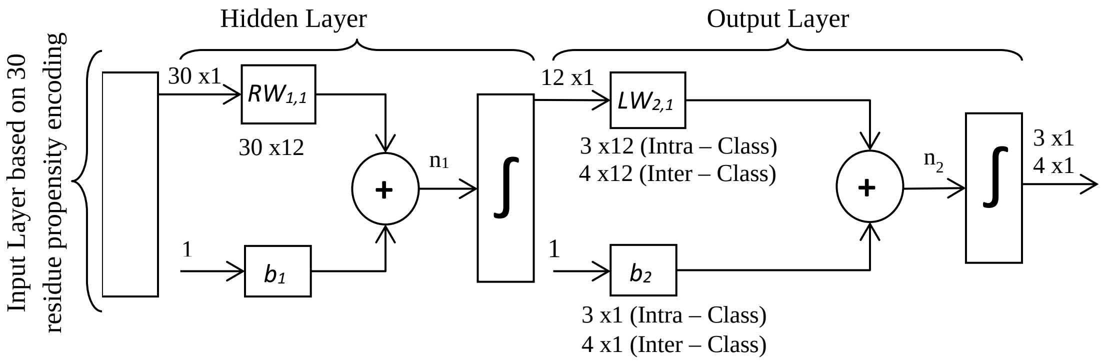


Figure 4: Architectural framework of a 30 input single hidden layer neural network for dual-model (intra-class and inter-class) beta barrel classification.

The NN to be trained for this problem uses the Levenberg-Marquardt training algorithm as discussed earlier. In each training cycle, the training sequences are presence via the sliding window operation discussed above that is one residue at a time. Each unit in the hidden layer transforms the signal value obtained from the input layer via a log-sigmoid transfer function in order to produce an output signal between zero and one, thus simulating neuron firing. During the process, the weights are adjusted in order to adjust/minimise the error values between each unit’s observed output and the actual output specified in the target matrix. In order to prevent an over-fitting of the problem and to generalise to new circumstances, an early stopping method is used to divide the data into training for gradient computation, testing and validation datasets.

3.3 Neural Network results analysis

As discussed, the feature sets for two training cases in neural architecture were trained using a dual layer feed-forward NN based on back propagation algorithm. The data used was divided into 70, 15, 15 percent sets for training, testing and validation respectively over data obtained from a sliding window size of 30 residues with 15 residues distributed evenly on both the sides of the window. Before simulating the proposed NNs system, tuning of each of these neural models are considered, such as:

1. Increasing the number of training feature vectors: With beta barrel class containing a comparatively lower number of instances, this step can only be performed with alpha helical databases.

2. The number of input values can be increased by adjusting the size of the sliding window. Computer simulation results demonstrate that a sliding window size of 30 residues with 15 residues distributed evenly on both the sides of the window provide best outcomes.

3. Increasing the number of neurons can also result in an increased accuracy though this also increases the risk of data over-fitting.

Table 1 shows the training, testing and validation analysis using a Levenberg-Marquardt training using randomised data division and a mean square error (MSE) based performance comparison. MSE is a risk function, corresponding to the anticipated value of the squared error loss. The sampling data are 6566 samples for training, 1407 samples for validation and 1407 samples for testing, a total of 9380 samples. The lower MSE values are generally due to the fact that inner-inner and outer-outer sliding window feature sets are substantially similar to each other. This is because the amino acid residues are either in the outside or inside or within membrane layers of protein sequences. The lower number of asymmetric test sets is primarily due to the lesser number of inter-class transitions within any training dataset, which produces a slightly higher MSE for asymmetric inter-class than symmetric intra-class, as there are not many beta-barrel proteins. It is also worth mentioning that NNs usually work better and produce smaller errors using more datasets.

Table 1: Data samples and relevant mean square values obtained for each sliding window-based feature input using a dual layer feed-forward network for both asymmetric and symmetric models.

	Symmetric *Hidden neurons:*	**MSE**	**Asymmetric** *Hidden neurons:*	**MSE**
Training	**6566**	**0.452**	**496**	**0.776**
Validation	**1407**	**0.471**	**106**	**0.699**
Testing	**1407**	**0.493**	**106**	**0.875**

The regression outcome for the symmetric and asymmetric datasets for NN training is shown in Figure 5 and Figure 6 respectively. In Figures 5 and 6, the variables that are predicting are referred to as Outputs and the variables that the predictions are based on are called Targets. Linear regression consists of finding the best-fitting straight regression line through the points. The vertical data (lines) from the points to the regression line represent the errors of predictions. The closer each data point to the regression line is, its error of prediction is smaller. In equation 5, the Output y-axis constitutes a gradient multiply by Target x-axis plus the point of intersection of the regression line with the y-axis (Montgomery, Peck and Vining 2012).

$$\text{Output} = \text{Gradient} * \text{Target} + \text{Intersection point with y-axis} \tag{5}$$

In Figure 5 the regression outcome for symmetric NNs training produces smaller errors than Figure 6 for asymmetrical NNs. The results of Table 1 MSE and Figures 5 & 6 are all consistent and produce very similar outcomes, reinforcing the prediction accuracies of NNs in prediction of beta-barrel amino acid proteins.

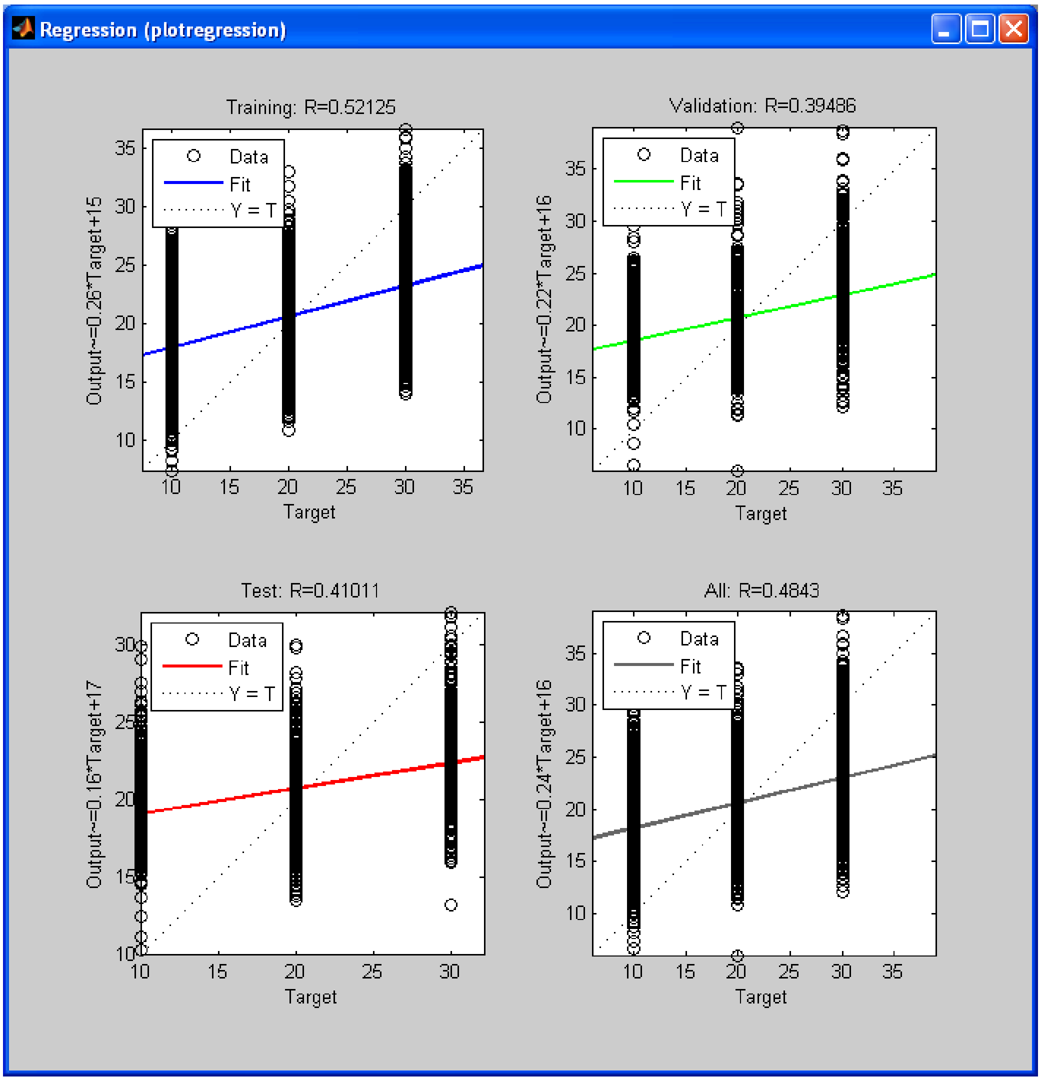


Figure 5: Regression outcome for symmetric NN training.

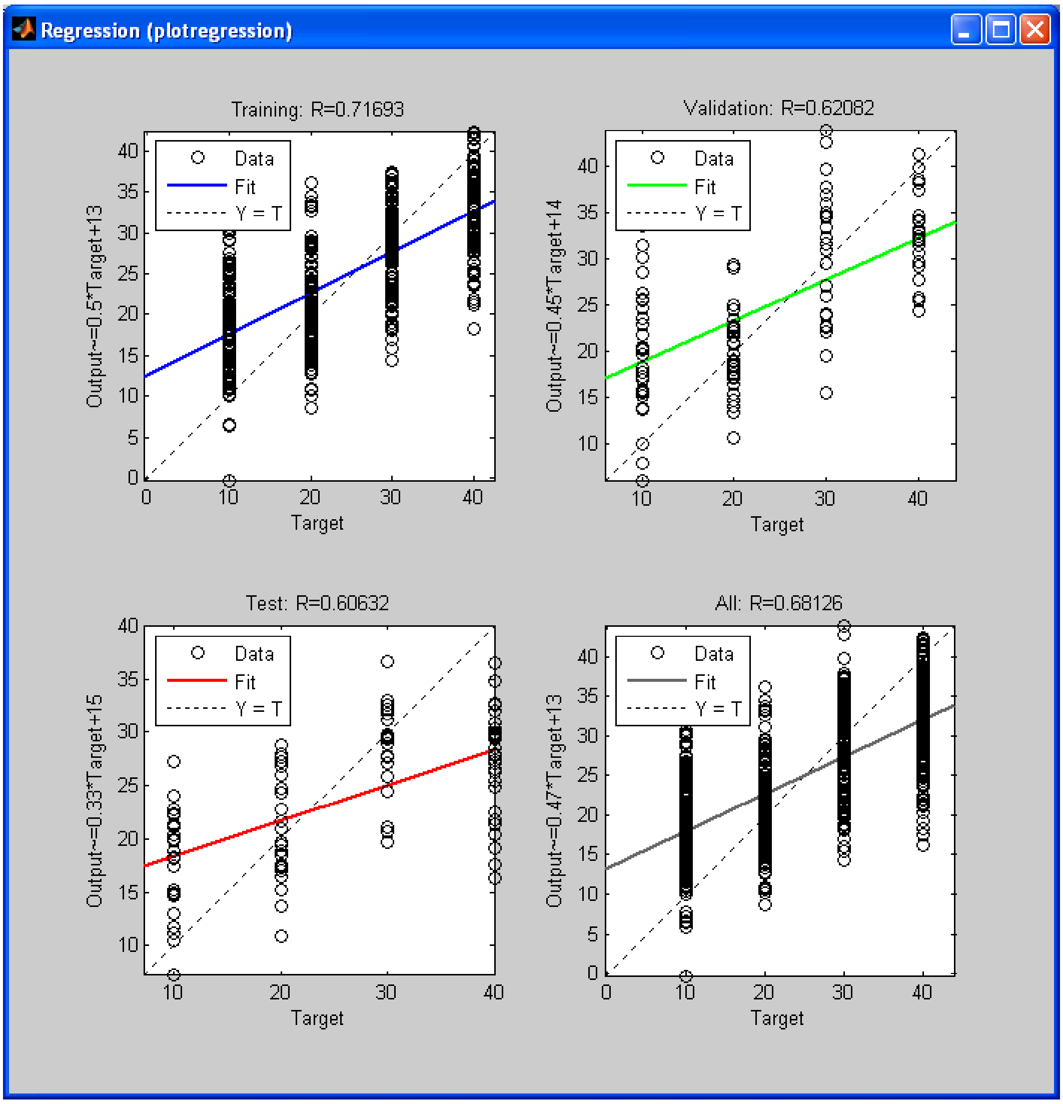
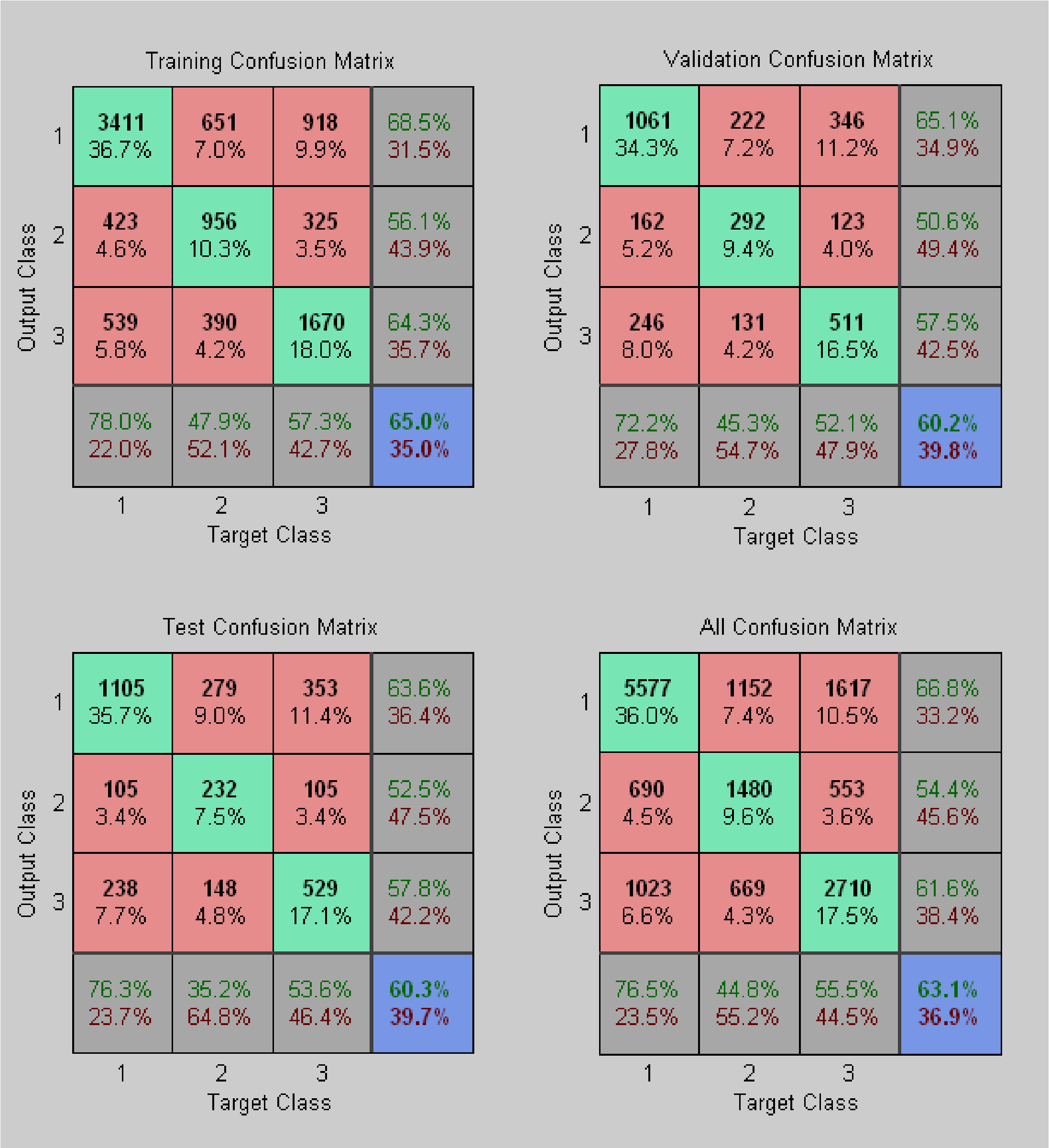


Figure 6: Regression outcome for asymmetric NN training.

This research further uses confusion matrices to analyse results of the proposed NNs technique. In machine learning, a confusion matrix is a table layout that facilitates visualization of the performance of an algorithm. A confusion matrix makes it easy to ascertain if the proposed model is confusing two classes that is mislabelling one as another. The following confusion matrix was obtained with a separate set of data compared to that shown in Table 1. The data size obtained was larger (9823 samples) than 9380 samples evaluated in Table 1. In Figure 7, the diagonal cells show the number of correctly identified residues for each structural class, outside, beta-barrel and inner. The off-diagonal cells represent misclassification for each group, for example, beta-barrel segment identified as inner non-transmembrane segment. Figure 7 shows overall low miss-classification rates shown as Targets – Outputs axes 2-1, 3-1, 1-2, 3-2,1-3 and 2-3. The overall accuracy shown in the lower right corner value with a highest overall value of validation data at 39.8%. The figure 60.2% in the validation confusion matrix and other training and testing confusion matrices are the sum of all diagonally correctly identified residues for each structural class, outside, beta-barrel and inner. It must be noted that the accuracy shown here only models a single (inter-class) model within the methodology, as identification of beta-barrel transmembrane protein is the overall objective of this research.

The overall accuracies are shown in Figures 8 and 9 with a weighted sum obtained over a logical AND operation performed over δ_A and δ_S scores. The scores δ_A and δ_S thus obtained for the sliding window for the asymmetric and symmetric models respectively and they are combined based upon their weighted proximity to the respective outputs to obtain an overall transmembrane score for each amino-acid residue to belong to membrane or non-transmembrane classes.



Training Confusion Matrix

Output Class \ Target Class	1	2	3	
1	3411 36.7%	651 7.0%	918 9.9%	68.5% 31.5%
2	423 4.6%	956 10.3%	325 3.5%	56.1% 43.9%
3	539 5.8%	390 4.2%	1670 18.0%	64.3% 35.7%
	78.0% 22.0%	47.9% 52.1%	57.3% 42.7%	65.0% 35.0%

Validation Confusion Matrix

Output Class \ Target Class	1	2	3	
1	1061 34.3%	222 7.2%	346 11.2%	65.1% 34.9%
2	162 5.2%	292 9.4%	123 4.0%	50.6% 49.4%
3	246 8.0%	131 4.2%	511 16.5%	57.5% 42.5%
	72.2% 27.8%	45.3% 54.7%	52.1% 47.9%	60.2% 39.8%

Test Confusion Matrix

Output Class \ Target Class	1	2	3	
1	1105 35.7%	279 9.0%	353 11.4%	63.6% 36.4%
2	105 3.4%	232 7.5%	105 3.4%	52.5% 47.5%
3	238 7.7%	148 4.8%	529 17.1%	57.8% 42.2%
	76.3% 23.7%	35.2% 64.8%	53.6% 46.4%	60.3% 39.7%

All Confusion Matrix

Output Class \ Target Class	1	2	3	
1	5577 36.0%	1152 7.4%	1617 10.5%	66.8% 33.2%
2	690 4.5%	1480 9.6%	553 3.6%	54.4% 45.6%
3	1023 6.6%	669 4.3%	2710 17.5%	61.6% 38.4%
	76.5% 23.5%	44.8% 55.2%	55.5% 44.5%	63.1% 36.9%

Figure 7: Confusion matrices of asymmetric dataset for training, test and validation datasets with low false positive and false negative rates.

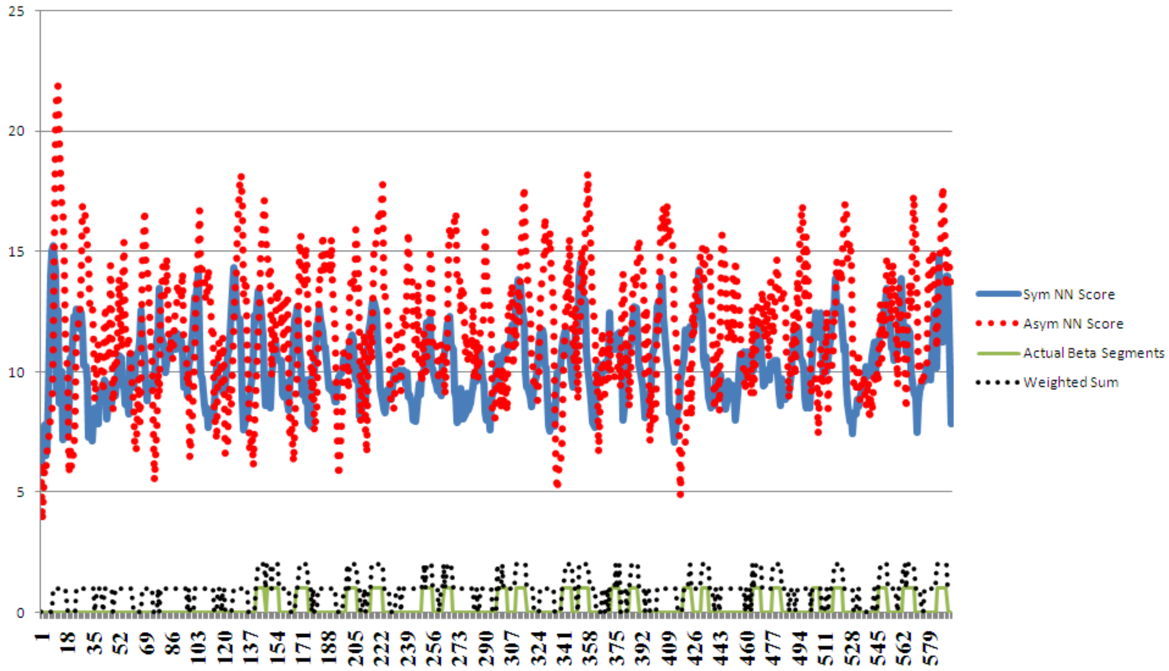


Figure 8: Transmembrane asymmetric δ_A red or broken dotted waveform and symmetric δ_S blue waveform scores with a weighted classification perform a normalised AND operation for δ_A and δ_S scores which are shown with thin violet line and the actual beta-barrel outcome in green colour underneath. The test was performed over a beta-barrel amino acid sequence of BtuB from *E.coli* (PDB accession no. 1ujw, chain A).

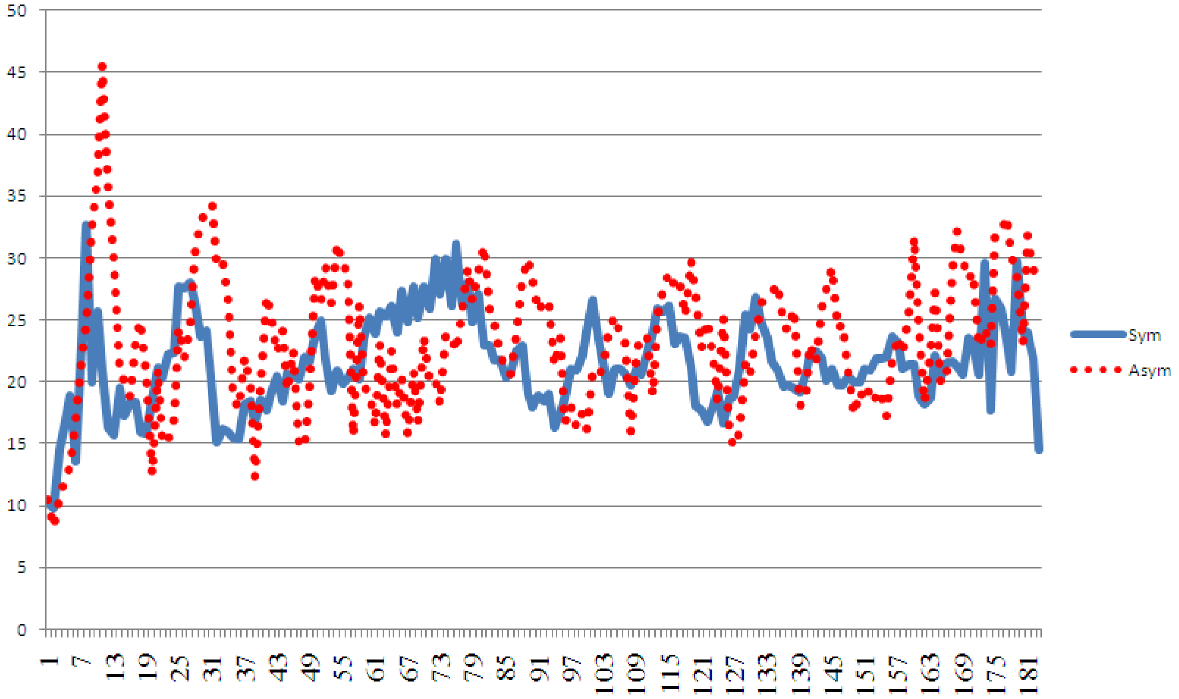


Figure 9: Non-transmembrane symmetric and asymmetric (δ_A and δ_S) scores tests performed over a negative amino acid sequence of colicin E2, a helical protein that binds to BtuB (PDB accession no. 1ujw, chain B).

4 Comparison of NN with NN-SVM for transmembrane and non-transmembrane calcifications.

4.1 Support Vector Machine technique

As discussed earlier, this research focuses on a cascaded (two-stage NN-SVM) to identify beta barrel segments within amino acid proteins. SVM has previously been used for the prediction of transmembrane segments to predict the presence of helical segments in transmembrane proteins (Kazemian, White, Palmer-Brown and Yusuf, 2013). The application of SVM in SP topological prediction has largely been focussed on differentiating N-terminal SPs from alpha-helical segments (Kazemian, Yusuf and White, 2014) primarily due to the fact that both the segments share a number of traits including their underlying hydrophobic nature.

In general, SVM are known to outperform or match identification accuracy for most benchmarking problems (Yuan, Mattick and Teasdale, 2004). Furthermore, contrary to NNs, SVM are known to excellently model problems with a smaller sample size. This makes this methodology an ideal technique for beta-barrel and signal peptide prediction problems where the modelling is undermined by problems of a smaller database or uneven class sample distribution, respectively (Yuan, Davis, Zhang and Teasdale, 2003).

As explained earlier, SVM is used to address the issue of beta-barrel identification. However, the modelling is largely limited to dual outcomes due to the underlying nature of the SVM classifier. The problem in transmembrane prediction is generally addressed as a two-class problem to differentiate protein segments containing beta barrel (or alpha-helices) from regular non-transmembrane segments. This topological identification plays a crucial role in helical and barrel protein prediction and a correction classification at this stage is very likely to improve the overall accuracy of transmembrane region prediction. To supplement and extend the ability of SVM for larger number of classes, this section reports a cascaded SVM methodology where the technique utilises two different feature sets based on symmetric (intra-class) and asymmetric (inter-class) sliding window operations, as discussed above.

The extracted features are shown in Figure 1 and Figure 2 where the transmembrane class feature vector is extracted under two sliding window sizes for symmetric and asymmetric modelling. For symmetric modelling the window size is -15/+15, refer to Figure 1. For the asymmetric modelling with Outer (O) to Beta (B) transition, the sliding window size is -3/+15 whereas for Beta to Inner (I) case the window size is -15/+3 as shown in Figure 2. The outcomes of these two models are later hybridised to obtain a more accurate topological prediction. The samples for this two class problem are described by the feature vector x_i where $i = 1, 2, \ldots, N$ where N is +30 with corresponding labels $y_i \epsilon \{+1, 0\}, X_i \epsilon\ R^d$. This study represents transmembrane protein (TM) class as +1 and non-transmembrane (NTM) class as 0. To predict the two class representations, SVM scheme maps and constructs a hyperplane in a high-dimensional space, which can be used to separate NTM from TM. A good separation generally is attained by the hyperplane that has the biggest margin, since the higher the margin the lower the generalization error of the classifier (Kazemian, Yusuf and White, 2014).

Supposing the hyperplane separates the two class representations of NTM and TM. The points X lying on the hyperplane satisfy $g.X + b = 0$, where g is regarded as the normal to the hyperplane and ${|b|}/{||g||}$ is the perpendicular distance from the line to the origin, $||g||$ is regarded as the Euclidean norm of g. Let's assume f^+ and f^- to be the shortest distances separating the two class samples $\{+1\ and\ 0\}$ respectively. For this problem, a margin is formulated as a separation and the largest margin is required by the support vector algorithm. The equations (6) and (7) are utilised for the training data with the following constraints:

$$\boldsymbol{X_i.g + b \geq 1\ for\ y_i = 1\ (f^+; TM)} \tag{6}$$

$$X_i.g + b \leq 0\ for\ y_i = 0\ (f^-; NTM) \tag{7}$$

Considering (6) and (7), the points $H_1: X_i. g + b = 1$ and $H_2: X_i. g + b = 0$ lie on the hyperplane with normal g and the perpendicular distance from the origin to be $|-1 - b|/2. \|g\|$. Therefore, $f^+ = f^- = {}^1\!/_{2.\,\|g\|}$ with a margin to be ${}^1\!/_{\|g\|}$, where b is zero. Consequently, based on the above equations, the objective is to ascertain a hyperplane for a two-class problem by maximising the margin and minimising $\|g\|$ (Steinwart and Christmann, 2008).

4.2 Analysis of NN and NN-SVM Results

In order to generate robust outcomes, a prediction methodology is evaluated by either of these conventional techniques, re-substitution test, independent data set test, jack-knife test, cross-validation or self-consistency based test. Jack-knife based testing is deemed most objective for any prediction methodologies involving large datasets with possible outliers (Abdi and Williams, 2010; Sawyer, 2005). The jack-knife or 'leave one out' testing is a cross-validation technique initially proposed by Quenouille (1949) to estimate the bias of an estimator. The jack-knife testing is an iterative process. To begin with, the parameter is estimated from the sampled dataset. Then each fold is in turn left out from the sample and the parameter of interest is estimated from the reminder of the dataset and finally the average of these tests is calculated. Therefore this methodology was used where each protein was set aside for testing and the model was trained using the remaining proteins belonging to the non-redundant protein set obtained from PDBTM website. This research uses jack-knife as a benchmark to compare the results with single protein tests. The underlying XML files contain all entries from the PDB database, TM as well as NTM proteins. The non-redundant protein set is obtained by the PDBTM server using the CD-HIT algorithm (Li and Godzik, 2006) with word size 2, 40% similarity measure and protein sequence length longer than 30. CD-HIT is an extensively used program for clustering protein secondary structures. CD-HIT is fast and enables reducing the computational processes in many sequence analysis tasks.

Table 2 demonstrates four unique outcomes of beta-barrel prediction for membrane spanning region (Jack knife), membrane spanning region (single protein tests), membrane prediction with 5 residue overlap (jack-knife), membrane prediction with 5 residue overlap (single protein tests) using NN. Membrane spanning region for single protein tests performs better than membrane spanning region for Jack knife tests with overall averaging results of 94.1625% and 85.904% respectively using NN. Furthermore, membrane prediction with 5 residue overlap for single protein tests produces better overall results than membrane prediction with 5 residue overlap for jack-knife tests with average of 94.4735% and 92.635% respectively.

Table 2: A combination of 20 beta barrel non-redundant outcomes.

	Type	Protein	Membrane Spanning Region (Jack knife) NN	Membrane Spanning Region (Single protein tests) NN	Membrane prediction (5 residue overlap) (Jack-knife) NN	Membrane prediction (5 residue overlap) (Single protein tests) NN
1	Beta	1a0s_P	91.92	89.94	93.75	92.99
2	Beta	3a2r_X	79.73	95.86	95.8	96.04
3	Beta	3aeh_A	79.91	92.93	91.82	92.15
4	Beta	3bry_A	75.89	95.85	91.73	95.99
5	Beta	3csl_A	86.82	96.73	91.78	94.02
6	Beta	3dwn_A	95.76	91.81	86.95	91.96
7	Beta	3dwo_X	93.94	93.77	89.97	94.04
8	Beta	1e54_A	85.75	91.86	88.99	93.15
9	Beta	3efm_A	92.8	95.9	95.84	96.95
10	Beta	3emn_X	89.93	93.81	90.78	90.98
11	Beta	3emo_C	88.94	93.83	95.9	97.98
12	Beta	2erv_A	88.96	92.9	94.79	98.02
13	Beta	2f1c_X	73.92	96.81	95.76	92.95
14	Beta	1fep_A	74.71	95.74	86.72	92.00
15	Beta	3fhh_A	93.78	90.86	93.88	91.01
16	Beta	3fid_A	84.94	94.95	96.83	93.06
17	Beta	1fw2_A	83.91	95.98	94.9	98.16
18	Beta	2grx_A	85.74	90.89	95.73	93.12
19	Beta	2guf_A	84.78	95.84	88.93	96.94
20	Beta	1h6s_1	85.95	96.99	91.85	97.96
	Average		85.904	94.1625	92.635	94.4735

The results of Table 2 using NN are compared with the proposed NN-SVM technique in Table 3. NN-SVM based topological prediction accuracy is shown in Table 3 with jack-knife demonstrating a comparatively reliable overall accuracy. Jack-knife testing was used as a benchmark to compare with single protein testing and the results are outlined in Table 3. Table 3 provides results for 194 redundant and 54 non-redundant proteins for NN and hybrid NN-SVM simulations for non-residue overlap and 5 residue overlap. In all cases outlined in Table 3, hybrid NN-SVM performs better than NN applications for prediction of beta barrel. The results reveal that single protein tests perform better than jack-knife. Furthermore, with 5 residue overlap the results of the computer simulations are better. In general, the best results are obtained for NN-SVM with 5 residue overlap for signal protein tests with and without redundant proteins. This demonstrates that the overall performance for single protein sequences is higher as compared with jack-knife-based testing which result in a more reliable measure of protein evaluation.

Table 3: Global proteins TM prediction accuracy with relevant outcomes based on Jack-knife based testing (Total proteins: 194).

Type (Percentage accuracy)	Membrane Spanning Region (Jack knife) NN / NN-SVM		Membrane Spanning Region (Single protein tests) NN / NN-SVM		Membrane prediction (5 residue overlap) (Jack-knife) NN / NN-SVM		Membrane prediction (5 residue overlap) (Single protein tests) NN / NN-SVM	
Non redundant (Total proteins: 54)	85.90	86.65	94.16	94.51	92.63	93.55	94.47	94.58
Redundant (Total proteins: 194)	88.76	89.45	95.28	97.85	94.65	96.5	96.31	97.96

The hybridised outcomes are further elaborated in Figure 10 for *E.coli* OmpF porin (PDB accession no. 1gfm, chain A) with a length of 340 amino acids, which shows the comparison of the asymmetric SVM model outcome (red —) and asymmetric NN score (blue ▬) with the actual membrane segments shown by yellow (• • • •) segments. The cross-point of both SVM and NN demonstrate the strong potential of inter-class transitions when used to predict transmembrane regions. As stated before inter-class or asymmetric transitions such as outer-to-membrane (beta), membrane (beta)-to-inner, inner-to-membrane and membrane-to-outer are, when the middle residue of the sliding window moved from one class to another predicting beta barrel regions. SVM and NN graphs of Figure 10 are further superimposed to graphically show the hybrid scoring of NN-SVM in Figure 11. Moreover, an instant of a sliding window is outlined in Figure 12 to demonstrate how NN-SVM outcomes are selected for prediction of transmembrane protein. Figure 12 is an instant of inter-class and intra-class and its comparison with the actual beta barrel segment using NN-SVM techniques.

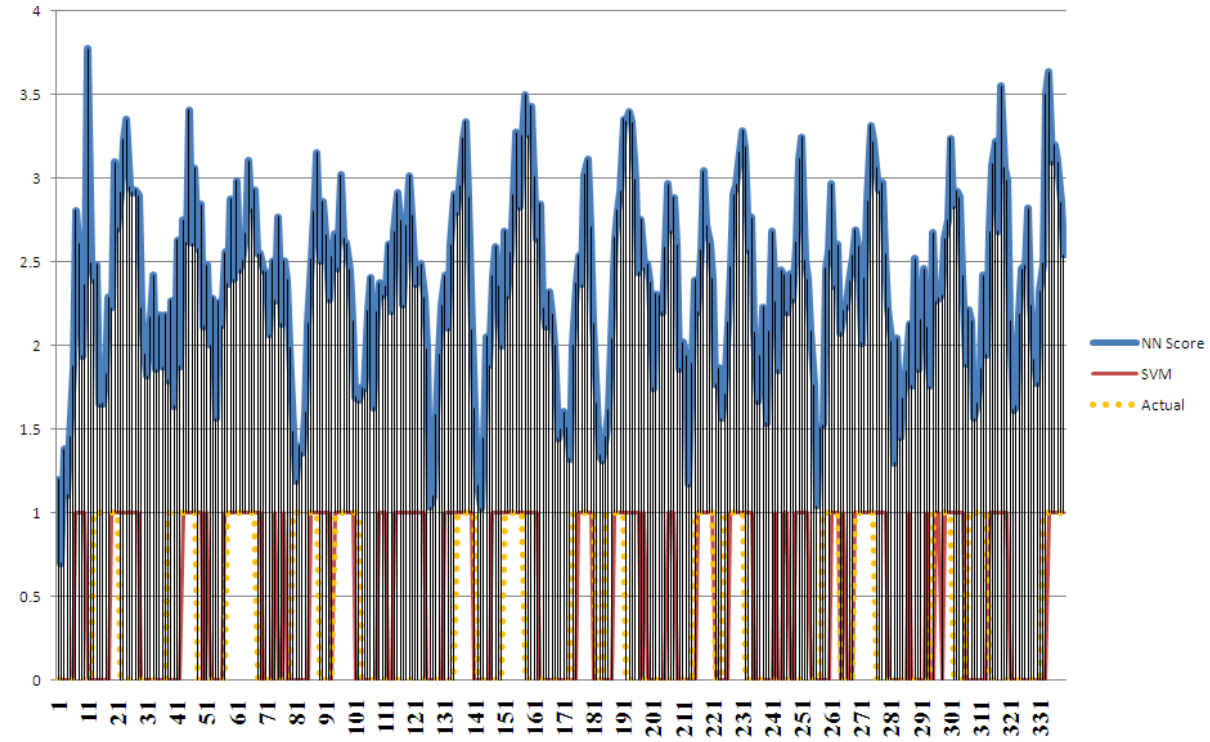


Figure 10: Protein 1gfm (Chain A) - Comparison of asymmetric SVM model outcome (red —) and asymmetric NN score (blue ▬) with the actual membrane segments shown by yellow (••••) segments. The hybrid NN-SVM score is shown in Figure 11.

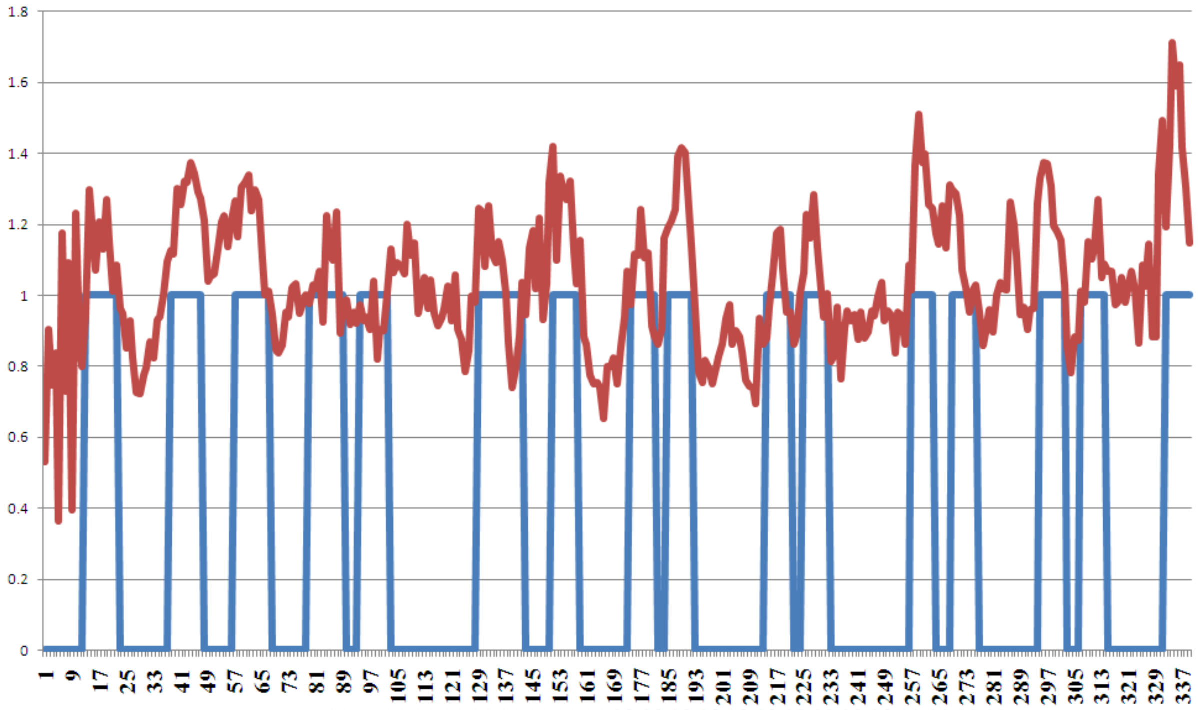


Figure 11: Asymmetric hybrid NN-SVM score is shown in red and actual output is shown in blue.

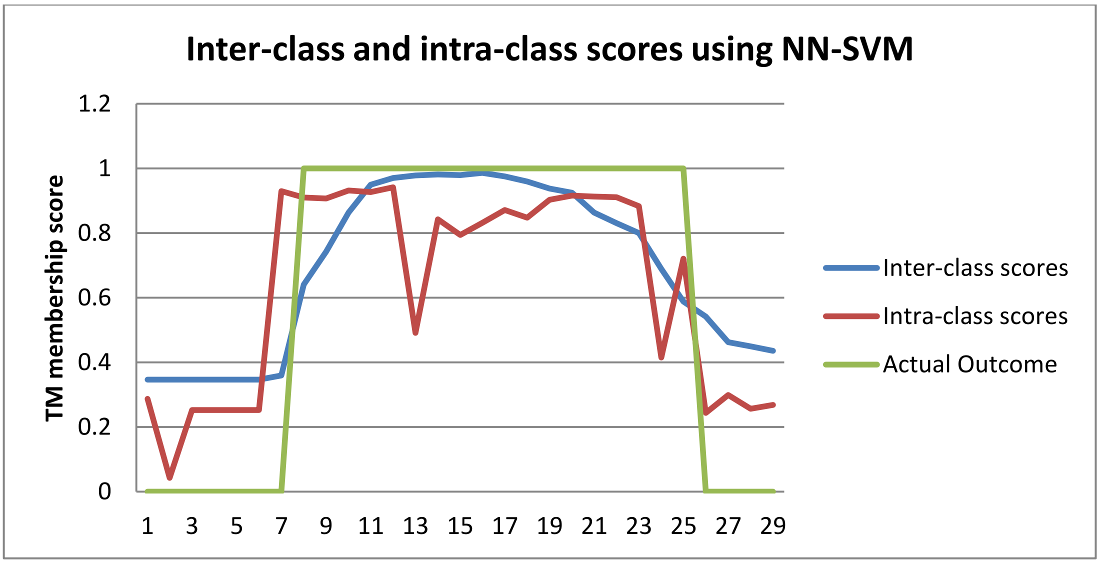


Figure 12 – An instant of inter-class and intra-class and its comparison with actual outcome.

The introduction section presents the review of historical background and compares many machine learning techniques for beta barrel prediction research and discusses why NN-SVM methodology is a next step forward in this endeavour. To be able to discuss the strengths and weaknesses of the proposed research, one needs to fundamentally analyse what are the requirements to develop a useful prediction technique. To develop a useful prediction method for a biological system (Chou, 2011), one needs to propose a robust algorithm for the prediction, select a valid benchmark dataset to train and test the model, and use appropriate cross-validation tests to critically appraise the expected accuracy of the prediction model. The strengths of the research is that the above criteria are fully implemented for the prediction of beta barrel transmembrane proteins with very encouraging results. This research proposes new requirements criteria using a sliding-window feature extraction to train two different class transitions called symmetric and asymmetric models to classify intra-class and inter-class transitions for the prediction of number and range of beta membrane spanning regions. As described throughout the paper, the research proposes NN and NN-SVM two robust machine learning algorithms and the well-known jack-knife testing as a benchmark to compare the results with single protein testing to critically evaluate the accuracy of the prediction models. The weakness of the paper is that the research in this area is not complete and the prediction accuracies may be further improved by using other techniques. For example, the research could be taken further in two different ways. Firstly, other machine learning techniques could be utilised to increase the prediction accuracy as outlined in the conclusion section below, and secondly, the prediction analysis of the protein topology, such as, intra-cellular, membrane spanning and extra-cellular could be researched upon to predict beta barrel topologies in amino acid sequences.

5 Conclusion

Research in beta barrel prediction analysis has many medical applications such as, production of novel drugs, addressing nutritional disorders for iron deficiency anaemia, cancer resistance genes, developing new antibiotics, study of neurotransmitter in the central and peripheral nervous system, and metabolic diseases associated with obesity, like diabetes, heart disease and cholesterol. This research utilises two new approaches NN and NN-SVM to prediction analysis of beta barrel transmembrane protein. The overall performance was evaluated over single protein sequences by feeding and pasting the whole sequences to the modelling routine and also utilising jack-knife-based testing. The accuracy were later combined and it was observed that the overall accuracy was higher for individual proteins as compared with jack-knife-based testing which reflected a more reliable measurement of protein evaluation.

For 54 non-redundant protein data, NN-SVM technique performed better than NN for prediction of membrane spanning regions. For single protein tests NN-SVM produced an overall accuracy of 94.51% and for jack-knife tests the average accuracy was 86.65% for prediction of membrane spanning regions. For 194 redundant protein data, NN-SVM also produced better results than NN for prediction of membrane spanning regions and the result were higher for single protein tests as opposed to jack-knife tests with 97.85% and 89.45% accuracies respectively. Membrane prediction with 5 residue overlap using NN-SVM technique resulted in improved outcome than NN and generated better overall accuracies for single protein tests than jack-knife tests with 94.11% and 93.55% respectively for non-redundant protein. Furthermore, the overall accuracy of NN-SVM technique with 5 residue overlap for single protein tests for membrane prediction was high at 97.96% for 194 redundant protein data, which is an outstanding achievement over the comparable research by TMM-HMM and PredTMR.

Further research will need to be carried out for the prediction accuracy of beta barrel transmembrane proteins, by using many other beta barrel amino acid sequences and some other machine learning techniques such as spiking NNs and deep learning. The prediction analysis of the protein topology, such as intra-cellular, membrane spanning and extra-cellular are understudied and also require improvements. Therefore, various machine learning techniques such as SVM, NNs, spiking NNs and deep learning could also be applied to prediction of the protein topologies.

Acknowledgement

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

References

Abdi, H. and Williams, L. J. (2010). Jackknife, In Neil Salkind (Ed.), *Encyclopedia of Research Design.* Thousand Oaks, CA: Sage Publications, 1-10.

Arai, M., Ikeda, M. and Shimizu, T. (2003). Comprehensive analysis of transmembrane topologies in prokaryotic genomes. *Gene*, 304, 77-86.

Bagos, P. G., Liakopoulos, T. D. and Hamodrakas, S. J. (2005). Evaluation of methods for predicting the topology of beta-barrel outer membrane proteins and a consensus prediction method, *Bmc Bioinformatics*, 6:7, 1-13.

Bagos, P. G., Liakopoulos, T. D., Spyropoulos, J.C. and Hamodrakas, S. J. (2004). PRED-TMBB: a web server for predicting the topology of b-barrel outer membrane proteins, *Nucleic Acids Research*, 32, W400–W404. Doi:10.1093/nar/gkh417.

Bigelow, H. R., Petrey, D. S., Liu, J., Przybylski, D. and Rost, B. (2004). Predicting trans-membrane beta-barrels in proteomes, *Nucleic Acids Res*earch, 32(8), 2566–2577.

Chou, K.C. (2011). Some remarks on protein attribute prediction and pseudo amino acid composition, *J. Theor. Biol*., 273 (1), 236–247.

Fariselli, P., Finelli, M., Rossi, I., Amico, M., Zauli, A., Martelli, P. L. and Casadio, R. (2005). TRAMPLE: the transmembrane protein labelling environment. *Nucleic Acids Research*, 33, W198-W201.

Hagan, M. T., Demuth, H. B, Beale, M. H. and De Jesús, O. (Sep 2014). Neural network design (2nd Ed), *Pub: Martin Hagan*, ISBN-10: 0971732116, ISBN-13: 978-0971732117.

Hayat, S. and Elofsson, A. (2012). BOCTOPUS: improved topology prediction of transmembrane β barrel proteins, *Bioinformatics* (First published online), doi: 10.1093/bioinformatics/btr710, 28 (4), 516-522.

Hayat, S., Sander, C., Elofsson, A. and Marks, D. A. (2014). Accurate Prediction of Transmembrane B-Barrel Proteins from Sequences. doi: http://dx.doi.org/10.1101/006577, *bioRxiv:006577.*

Hayat, S., Sander, C., Marks, D. S. and Elofsson, A. (2015). All-Atom 3D Structure Prediction of Transmembrane B-Barrel Proteins from Sequences. *Proceedings of the National Academy of Sciences of the United States of America*, 112 (17), 5413–18.

Kall, L., Krogh, A. and Sonnhammer, E. L. L. (2004). A combined transmembrane topology and signal peptide prediction method. *Journal of Molecular Biology*, 338(5), 1027-1036.

Kazemian, H. B., Yusuf, S. A. and White, K. (2014). 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, 45, 98–110.

Kazemian, H. B. and Yusuf, S. A. (6th – 11th July 2014), An ANFIS approach to transmembrane protein prediction. *IEEE World Congress on Computational Intelligence (IEEE WCCI 2014), IEEE International Conference on Fuzzy Systems*, Beijing China, 1360-1365.

Kazemian, H. B., White, K., Palmer-Brown, D. and Yusuf, S. A. (2013). Applications of evolutionary SVM to prediction of membrane alpha-helices, *Expert Systems with Applications, Elsevier*, DOI: 10.1016/j.eswa.2012.12.049, 40(9), 3412–3420.

Institute of Enzymology. (2015). PDBTM: Protein Data Bank of Transmembrane Proteins, *PDBTM version: 2015-10-02,* http://pdbtm.enzim.hu/?_=/statistics/growth, accessed 2nd Oct 2015.

Li, W. and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics,* 22(13), 1658-1659.

Montgomery, D. C., Peck, E. A. and Vining, G. G. (2012). Introduction to linear regression analysis (Wiley Series in Probability and Statistics), *Pub: Wiley-Blackwell; 5th Edition*, ISBN-10: 0470542810, ISBN-13: 978-0470542811.

Neupert, W. and Lill, R. (1992). Membrane biogenesis and protein targeting. *ISBN: 978-0-444-89638-4, Elsevier*.

Quenouille, M. H. (1949). Problems in plane sampling. *The Annals of Mathematical Statistics,* 20(3), 355–375.

Rangwala, H. and Karypis, G. (2010). Introduction to Protein Structure Prediction: Methods and Algorithms, *Wiley Series on Bioinformatics*.

Reynolds, S. M., Kall, L., Riffle, M. E., Bilmes, J. A. and Noble, W. S. (2008). Transmembrane topology and signal peptide prediction using dynamic Bayesian networks. *Plos Computational Biology*, 4(11), 1-14.

Russell, J. and Cohn, R. (2012). Alpha helix, *Bookvika Publishing,* ISBN-13: 9785510831610.

Sawyer, S. (2005). Resampling Data: Using a Statistical Jack-knife. *Washington University*, 1-5.

Schulz, G. E. (2000). Beta barrel membrane proteins, *Current Opinion in Structural Biology, Elsevier,* 10(4), 443 – 447.

Schulz, G. E. (2002). The structure of bacterial outer membrane proteins, *Biochimica et Biophysica Acta (BBA) - Biomembranes*, 1565(2), 308–317.

Singh, N. K., Goodman, A., Walter, P., Helms, V. and Hayat, S. (2011). TMBHMM: A Frequency Profile Based HMM for Predicting the Topology of Transmembrane Beta Barrel Proteins and the Exposure Status of Transmembrane Residues. *Biochimica et Biophysica Acta (BBA)* - Proteins and Proteomics, doi: 10.1016/j.bbapap.2011.03.004, 1814 (5), 664–70.

Steinwart, I. and Christmann, A. (2008). Support Vector Machines [electronic resource]. *Dordrecht, Springer,* ISBN: 978-0-387-77242-4.

Stillman, B. (1995). Protein kinesis: the dynamics of protein trafficking and stability, *60th Cold Spring Harbor Symposium on Quantitative Biology*. N.Y.: Cold Spring Harbor Laboratory Press.

Taylor, P. D., Toseland, C. P., Attwood, T. K. and Flower, D. R. (2006). Beta Barrel Trans-Membrane Proteins: Enhanced Prediction Using a Bayesian approach, *Bioinformation by Biomedical Informatics Publishing Group*, ISSN 0973-2063, 1(6), 231-233.

Tran, V. D. T., Chassignet, P., Sheikh, S. and Steyaert, J. M. (2012). A Graph-Theoretic Approach for Classification and Structure Prediction of Transmembrane Β-Barrel Proteins. *BMC Genomics*, doi:10.1186/1471-2164-13-S2-S5, 13 (Suppl 2): S5.

Wimley, W. C. (2001). Toward genomic identification of b-barrel membrane proteins: composition and architecture of known structures. *Protein Science*, 11, 301-312.

Yu, H. and Wilamowski, B. M. (2010). Levenberg–Marquardt training, *Auburn University*, K10149_C012.indd, 1- 16.

Yuan, Z., Mattick, J. S. and Teasdale, R. D. (2004). SVMtm: Support vector machines to predict transmembrane segments. *Journal of Computational Chemistry*, 25(5), 632-636.

Yuan, Z., Davis, M. J., Zhang, F. and Teasdale, R. D. (2003). Computational differentiation of N-terminal signal peptides and transmembrane helices. *Biochemical and Biophysical Research Communications*, 312(4), 1278-1283.

Zou, L., Wang, Z., Wang, Y. and Hu, F. (2010). Combined Prediction of Transmembrane Topology and Signal Peptide of Beta-Barrel Proteins: Using a Hidden Markov Model and Genetic Algorithms. *Computers in Biology and Medicine*, 40(7), 621–28.

Point-to-Point Responses to Reviewers

Reviewer #1:
I think the paper is well written and cover an interesting topic. I have just some comments to give to the authors and I hope these can help them to improve it.
Response to reviewer #1: Thank you.

Reviewer #1: *Note: (1p1L) stands for 1st paragraph-1st Line, (2p3L) 2nd paragraph-2nd Line and so on.
- ABSTRACT: I think the abstract is too long. Authors should consider writing it in a more concise way especially as they describe the methodology. However, this is just a suggestion.

Response to reviewer #1:
We have tried to reduce the abstract. Unfortunately, it was not possible without compromising the main objectives of the paper and to make sure that the abstract is concise and the overall methodology is clear.
However, the paragraphs have been slightly rearranged and some new words have been added and taken out to make sure that the abstract reads better. The second paragraph provides a clear description of the methodology.

Reviewer #1: - SECTION 1 (Introduction):
(1p1L) I would clarify that: Integral-membrane proteins are a type of transmembrane proteins that are permanently attached to the membrane. According to their tertiary structure, they can be divided into alpha-helical and beta-barrels (reference is needed). N.B. this is a technical journal; readers may not be familiar with these concepts. I think one line in the introduction would help readers to contextualize them.
Response to reviewer #1:
The following two sentences and a new reference have been added to 1p1L. Thank you.
"Integral membrane proteins are a type of transmembrane proteins that are permanently attached to the membrane. According to their tertiary structure, they can be divided into alpha-helical and beta-barrel proteins (Rangwala and Karypis, 2010)."

Reviewer #1: (1p4L) "Beta-barrel membrane proteins generally occur in special membranes" - what do you mean by "special membrane"?
Response to reviewer #1:
'Special' is a mistake. I have now re-written the sentence. The sentence is also outlined below:
"For transmembrane beta-barrels, the integral protein segments are known to occur in outer membranes of bacteria, mitochondria and chloroplasts (Neupert and Lill, 1992)". Thank you.

Reviewer #1: (5p6L) In my opinion, authors should first define the acronym PDBTM as the Protein Data Bank of Transmembrane Proteins (PDBTM); and PDB as Protein Data Bank (PDB). As they appear for the first time in the text.
Response to reviewer #1:
Both PDBTM and PDB acronyms have been defined. Thank you.

Reviewer #1: (6p5L) again, acronyms are not defined when they first appear in the text….. "HMM" …."GA". I have also noticed that there is a confusing use of acronyms, as already defined acronyms are then not used in the text anymore (e.g. Section 3.3-1p5L- where NN now becomes Neural Network again). I just find this a bit confusing. Although, action is not required from the authors.
Response to reviewer #1:

Both HMM and GA acronyms have been defined. Thank you.

Reviewer #1: - SECTION 2:
(1p3L) please define "TM" as "transmembrane (TM) segments" or simply write "transmembrane segments" and reduce the number of acronyms in the text.
Response to reviewer #1:
TM acronym has been defined. Thank you.

Reviewer #1: - SECTION 3: Subsection 3.1
(2p3L): Every parameter in Equation (1) (and the following equations) needs to be clearly identified (i.e. what is J? what is xk? etc.)
Response to reviewer #1:
J, x_k and *I* parameters have been defined. Thank you.

Reviewer #1: - SECTION 4: Subsection 4.1
(4p9L) In the text is stated: "This study represents non-transmembrane protein (TM) class as +1 and transmembrane (NTM) class as 0." I think it should be: "represents non-transmembrane protein (NTM) and transmembrane (TM)".
Response to reviewer #1:
The mistake has been rectified and it now reads: “This study represents transmembrane protein (TM) class as +1 and non-transmembrane (NTM) class as 0”. Thank you.

Reviewer #1: Subsection 4.2: (1p16L) CD-Hit was already defined in the introduction (p5).
Response to reviewer #1:
The definition has been taken out. Thank you.

Reviewer #1: (2p4L) NN was already defined in the abstract.
Response to reviewer #1:
The definition has been taken out. Thank you.

Reviewer #1: - SECTION 5:
Did you think about future work to be done in this research? Have you considered using spiking neural networks?
Response to reviewer #1:
The answer is yes and the following paragraph has been newly added to section 5.
“Further research will need to be carried out for the prediction accuracy of beta barrel transmembrane proteins, by using many other beta barrel amino acid sequences and some other machine learning techniques such as spiking NNs and deep learning. The prediction analysis of the protein topology, such as intra-cellular, membrane spanning and extra-cellular are understudied and also require improvements. Therefore, various machine learning techniques such as SVM, NNs, spiking NNs and deep learning could also be applied to prediction of the protein topologies.”.

Reviewer #2:
- Even though the article is interesting in its current format, some aspects should be improved for possible publication and for a better understanding by the readers.
Response to reviewer #2:
Thank you.

Reviewer #2: - The authors should give the readers some concrete information to get them excited about their work. The current abstract only describes the general purposes of the article. It should also include the article's main (1) impact and (2) significance on expert and intelligent systems.

Response to reviewer #2:

To add excitement to the abstract, the paragraphs have been rearranged. The overall four objectives of the two paragraphs are:

(i) To introduce the research area. This is in the first paragraph.

(ii) To explain the problem in this area of research needs addressing. This is in the first paragraph.

(iii) To outline the solution that is proposed for this research problem. In other words, the precise methodology that is used. This is in the second paragraph.

(iv) To present the overall computer simulation results. This is in the second paragraph and new words such as significant impact and superior performance have been used to mention to the reader that SVM-NN technique provides a significant impact on the beta barrel prediction analysis.

Reviewer #2: - Please give a frank account of the strengths and weaknesses of the proposed research method. This should include theoretical comparison to other approaches in the field.

Response to reviewer #2:

A comprehensive theoretical comparison to other approaches in this field has been discussed in detail in the Introduction section, which describes the historical evolution of the beta barrel prediction using various machine learning techniques.

A frank account of the strengths and weaknesses of the proposed research has been outlined in a new paragraph at the bottom of section 4.2. The paragraph is also outlined here:

“The introduction section presents the review of historical background and compares many machine learning techniques for beta barrel prediction research and discusses why NN-SVM methodology is a next step forward in this endeavour. To be able to discuss the strengths and weaknesses of the proposed research, one needs to fundamentally analyse what are the requirements to develop a useful prediction technique. To develop a useful prediction method for a biological system (Chou, 2011), one needs to propose a robust algorithm for the prediction, select a valid benchmark dataset to train and test the model, and use appropriate cross-validation tests to critically appraise the expected accuracy of the prediction model. The strengths of the research is that the above criteria are fully implemented for the prediction of beta barrel transmembrane proteins with very encouraging results. This research proposes new requirements criteria using a sliding-window feature extraction to train two different class transitions called symmetric and asymmetric models to classify intra-class and inter-class transitions for the prediction of number and range of beta membrane spanning regions. As described throughout the paper, the research proposes NN and NN-SVM two robust machine learning algorithms and the well-known jack-knife testing as a benchmark to compare the results with single protein testing to critically evaluate the accuracy of the prediction models. The weakness of the paper is that the research in this area is not complete and the prediction accuracies may be further improved by using other techniques. For example, the research could be taken further in two different ways. Firstly, other machine learning techniques could be utilised to increase the prediction accuracy as outlined in the Conclusion section below, and secondly, the prediction analysis of the protein topology, such as, intra-cellular, membrane spanning and extra-cellular could be researched upon to predict beta barrel topologies in amino acid sequences.”.

Reviewer #2: - Moreover, I believe that it will make this paper stronger if the authors present managerial insights based on their experimental outcomes.

Response to reviewer #2:

At Intelligent Systems Research Centre, we have been applying AI techniques to transmembrane proteins prediction since 2003. There have been many methodologies that we have been pursuing to improve the prediction results of transmembrane proteins, which will be out of the scope of this paper, if we try to outline those. However, In the Introduction section, in the penultimate paragraph,

the starting paragraph has been expanded to provide some insights to the choice of NN-SVM technique based on the experimental results of the last three papers in this area. The added sentences and the whole paragraph will provide a good back ground to managerial insides based on the experimental outcomes. The added sentences to the paragraph are also outlined below:
"One of the most encouraging results that has been obtained in applications of machine learning techniques to transmembrane proteins was, the application of SVM-GA to alpha helices where the overall outcomes were published in 2013 (Kazemian, White, Palmer-Brown and Yusuf, 2013). Through a future research, a hybrid NN and fuzzy logic technique entitled Adaptive Neural Fuzzy Inference System was also applied to predict and analyse membrane helices in amino acid sequences which produced a comparable results to using SVM-GA (Kazemian and Yusuf, 2014). In general, SVM is known to model problems with a smaller sample size. This makes the SVM an appropriate technique for beta-barrel prediction problems where the modelling is undermined by problems of a smaller database. Furthermore, Levenberg-Marquardt algorithm is perceived as one of the most effective method for training NN. The Levenberg-Marquardt training algorithm is fast, but it is generally more demanding in terms of memory. ....".

Reviewer #2: - Finally, There are no real insightful conclusions drawn from the study and no suggestions for practical use of the results. Therefore, the conclusion section should be totally rewritten in order to:
a) discuss research contributions in Expert and Intelligent Systems and indicate practical advantages (in at least one separate paragraph),

Response to reviewer #2:
a) The practical advantages to biology are highlighted at the beginning of the first paragraph of the Conclusion. The research contributions in Expert and Intelligent Systems are discussed in the first paragraph and continued into the second paragraph for in depth analysis.

Reviewer #2: b) discuss research limitations (at least one separate paragraph), and

Response to reviewer #2:
b) Research limitations are outlined in paragraph two of the Conclusion by discussing the percentage accuracies of prediction of beta barrels in amino acid sequences. Needless to say that since the accuracies are not 100%, then, there are limitations in the research.

Reviewer #2: c) supply 4-5 solid and insightful future research suggestions in Expert and Intelligent Systems (in at least one separate paragraph) for the ESWA community. No bullets should be used in your conclusion section.

Response to reviewer #2:
c) Paragraph three in the Conclusion outlines some suggestions for further research for the ESWA community. Initially, it recommends two other machine learning techniques spiking NNs and deep learning to be applied to beta barrel transmembrane protein to increase the prediction accuracy. Then, the paragraph mentions that a very close area related to this research called 'protein topology' such as intra-cellular, membrane spanning and extra-cellular is understudied and could be further researched. Finally, it recommends that machines learning techniques such as SVM, NNs, spiking NNs and deep learning could be applied to protein topology.

Reviewer #2: - If the paper is resubmitted as a significantly reworked piece of work, offering a proper view with clear Point-to-Point responses on what is the novelty and significantly improving the evaluation, then I can imagine a more positive second evaluation.

Response to reviewer #2:

The paper is resubmitted as a significantly revised piece of work addressing all the recommendations made by both reviewers, providing clear Point-to-Point answers as outlined above and emphasising on the novelty and evaluation of the research. Please refer to the Point-to-Point responses to the reviewers' comments above and the paper itself.