Computers in Biology and Medicine 45 (2014) 98–110

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/cbm

Signal peptide discrimination and cleavage site identification using SVM and NN

H.B. Kazemian [a,*], S.A. Yusuf [b], K. White [a]

[a] London Metropolitan University, UK
[b] Portsmouth University, UK

ARTICLE INFO

Article history:
Received 31 July 2013
Accepted 22 November 2013

Keywords:
Signal peptide discrimination
Support vector machine
Cleavage site prediction
Neural networks.

ABSTRACT

About 15% of all proteins in a genome contain a signal peptide (SP) sequence, at the N-terminus, that targets the protein to intracellular secretory pathways. Once the protein is targeted correctly in the cell, the SP is cleaved, releasing the mature protein. Accurate prediction of the presence of these short amino-acid SP chains is crucial for modelling the topology of membrane proteins, since SP sequences can be confused with transmembrane domains due to similar composition of hydrophobic amino acids. This paper presents a cascaded Support Vector Machine (SVM)-Neural Network (NN) classification methodology for SP discrimination and cleavage site identification.

The proposed method utilises a dual phase classification approach using SVM as a primary classifier to discriminate SP sequences from Non-SP. The methodology further employs NNs to predict the most suitable cleavage site candidates. In phase one, a SVM classification utilises hydrophobic propensities as a primary feature vector extraction using symmetric sliding window amino-acid sequence analysis for discrimination of SP and Non-SP. In phase two, a NN classification uses asymmetric sliding window sequence analysis for prediction of cleavage site identification.

The proposed SVM-NN method was tested using Uni-Prot non-redundant datasets of eukaryotic and prokaryotic proteins with SP and Non-SP N-termini. Computer simulation results demonstrate an overall accuracy of 0.90 for SP and Non-SP discrimination based on Matthews Correlation Coefficient (MCC) tests using SVM. For SP cleavage site prediction, the overall accuracy is 91.5% based on cross-validation tests using the novel SVM-NN model.

1. Introduction

The prediction of a protein topology starts with the process of predicting if it contains a signal peptide (SP) in the N-terminus and hence whether the protein accesses many of the secretory pathways of the cell, in both eukaryotes or prokaryotes [1]. SPs are short N-terminal peptides that are cleaved off after the protein has been correctly inserted into a secretory pathway. The remaining protein is regarded as the mature protein, and the delivery of these proteins to the correct cellular location must be made accurately. Mistakes or mutation in the signal peptide cleavage position may result in the protein being delivered to the wrong cellular location and causing disease [2]. Almost 15% of human proteins contain SPs [3] and such proteins are either secreted or inserted into membranes as type I membrane proteins. Signal peptide prediction is an important step in predicting membrane protein topology, because signal peptide and transmembrane sequences are handled by the same mechanism of membrane insertion, involving the translocon. Similar to transmembrane (TM) segments, SPs also contain a hydrophobic alpha-helix region [4]. The SP region is however shorter, 7–15 residues approximately, compared with a TM helical segment. The SP is structured as such that at the N-terminus a positively charged n-region is located. The length of residues in the n-region varies from 1 to 12 residues and is followed by the hydrophobic region, ‘h-region’, of 7–15 residues. After the h-region another 3–8 residues long polar and uncharged c-region is positioned, where the cleavage point is located. Prediction of SPs therefore requires two processes, one to correctly identify the SP in the N-terminus of the sequence and the second to map accurately the position of the SP cleavage site. The proposed method in this paper is for predicting the N-terminal SPs and their cleavage sites, as most SPs are of N-terminal, excluding exceptionally long SPs [5]. However, SPs may also occur in the middle of a protein sequence or it's C-terminal [6]. Many researchers have investigated SP discriminations and cleavage sites identifications in human, plant, animal, eukaryotic, Gram-positive and Gram-negative protein sequences [7–10], some of which are discussed below.

Arai et al. [11] carried out a comprehensive analysis of TM topologies of over 50 prokaryotic genomes which was mainly

[*] Corresponding author.
E-mail address: h.kazemian@londonmet.ac.uk (H.B. Kazemian).

0010-4825/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.compbiomed.2013.11.017

intended to evaluate the evolutionary mechanism of TM topology, and found that 13% of TM proteins have signal peptides; this is further supported by the SP database (db) that contains 2584 verified signal peptide entries. Prediction accuracy of TM proteins is considerably enhanced if the N terminal and signal peptides are taken into account in the model input [12,3]. It should be noted that NN and SVM based methodologies generally tend to confuse signal peptide segments as TM regions due to their underlying hydrophobic properties. The incorrect position of signal peptide cleavage sites in a database stems from trivial database errors and from peptide sequencing, where it may be difficult to control the level of post-processing of the protein by other peptidases after the signal peptidase I has found its initial cleavage. In relation to the true signal peptidase I cleavage site position, this type of post-processing may lead to cleavage site shifted downstream [13]. Gomi et al. [14] used clustering based signal sequences and signal peptide scoring indices to discriminate signal peptides from globular proteins and predict their cleavage scores.

Support Vector Machines (SVMs) are also generally used to address and implement fast prediction algorithms for cleavage site prediction. Cai et al. [15] used 20-bit binary based feature vectors to predict signal peptide cleavage sites and proposed the use of SVM as a complementary approach. Kahsay et al. utilised a SVM-Fisher discrimination method to enhance SVM performance and improved mis-prediction of signal peptides by 30% [16]. Hybridisation/ensembling of classifiers to predict transmembrane topology including signal peptide prediction have become the focus of attention of researchers since 2003. Martelli et al. [17] used three machine learning methods on a single neural to two HMM-based classifiers to implement a voting mechanism that resulted in substantial improvement in individually performing classifiers. Kall et al. [3] applied a HMM approach to increase the prediction accuracy using pre-training of length models for TM and SP regions. Kall et al. conclude that since the evaluation methodologies are variable, some of the classifications could be regarded as overestimations. Considering this issue, Melen et al. presented reliability scores for five widely used topology prediction methods, such as, TMHMM, HMMTOP, MEMSAT, PHD and TopPred [18]. Clote [19] evaluated SVM, HMM, stochastic context free grammars and neural networks with weight matrices to predict 70% detection of true positives compared to 10% false positives. Fariselli et al. [20] developed a neural network-based predictor for four sets namely, Gram-positive prokaryotes, Gram-negative prokaryotes, eukaryotes and *Escherichia coli*. The result is a content management web-server for personalised user research. Furthermore, Hawkins and Boden [21] utilised recurrent neural networks to predict data divided into plant and non-plant signal peptides.

Recently, there has been substantial work covering the integration of transmembrane topological prediction with signal peptide modelling. Plewczynski et al. [22] used neural networks to detect signal peptides from the extracted Swiss-PROT protein database and obtained a combined accuracy of 73% for eukaryotes and prokaryotes. Reynolds [23] utilised dynamic Bayesian networks which resulted in achieving a relative accuracy of 13% over Phobius methodology with sensitivity and specificity of 0.96 in signal peptide detection. Sun and Wang [24] used SVM with a *K*-nearest classifier. They claim an overall achievement of 97% in signal peptide prediction rate. Nugent and Jones [25] developed a SVM based model to predict both signal peptide and re-entrant helices. The method achieved an accuracy of 93% in the prediction of signal peptide prediction. However due to lack of available data, Nugent and Jones's method limited the accuracy of re-entrant helices to only 44%. Zou et al. applied a hybrid HMM/Genetic algorithms model for signal peptide prediction and achieved an overall accuracy of 84.8%, which outperforms a number of previous techniques of SignalP 3.0 – NN/HMM and SignalP 2.0 – NN/HMM [26].

von Heijne used the initial concept of the weight matrix to predict signal peptide cleavage sites [27]. Folz and Gordon further utilised two different algorithms to predict signal cleavage locations [28]. Chou also [29] developed the sub-site coupling method using the sequence encoded algorithm and the scaled window approach [30,31]. However, in 1997 Nielsen et al. utilised an artificial intelligence (AI) based approach for the widely used SignalP algorithm [32]. Bendtsen et al. introduced SignalP 3.0 to improve the accuracies by adding supplementary network attributes [13]. Nielsen et al. and Bendtsen et al. applied neural networks and hidden Markov models (HMMs) to SignalP. The work was further extended to SignalP 4.0 by Petersen et al. [33]. The paper argues that SignalP 4.0 was not in all cases as good as SignalP 3.0 according to cleavage-site sensitivity or signal-peptide correlation when there are no transmembrane proteins present. Petersen et al. however concludes that SignalP 4.0 is an improvement over SignalP 3.0. SignalP 4.0 is the latest research in SP and Non-SP discrimination and cleavage site identification and both SignalP 3.0 and SignalP 4.0 are widely used as a benchmark. This paper therefore endeavours to compare the overall results of the computer simulation with SignalP 4.0. As demonstrated by a series of recent publications [34–36] and summarised in a review [37], to develop a useful prediction method for a biological system, we need to consider the following procedures: (i) construct or select a valid benchmark dataset to train and test the model or predictor; (ii) introduce a powerful algorithm to operate the prediction; (iv) properly perform cross-validation tests to objectively evaluate the anticipated accuracy of the predictor; and (v) subsequently establish a user-friendly web-server for the model or prediction method that is accessible to the public [38].

In this report, Section 2 outlines the current methodologies to predict signal peptides. The section describes feature extractions using symmetric and asymmetric sliding windows and further explanation on data extraction from the Swiss-Prot database. Section 3 describes the proposed novel support vector machine and neural network architectures for SP and Non-SP discrimination and cleavage site identification. Section 4 outlines the results of computer simulations for support vector machine and neural network methodologies for SP and Non-SP discrimination and cleavage site prediction respectively. Finally, Section 5 concludes the outcome of the research.

2. Prediction of signal peptides

Traditional transmembrane (TM) region prediction is based on a supervised learning algorithm using a sequence of amino acids. The training set consists of sequence of the form (t, l), where $t = t_1$, t_2, …, t_n are considered as a sequence of amino acids of known topology and $l = l_1, \ldots, l_n$ correspond to the training sequence t. n is 20 representing 20 amino acid types. The conventional TM domains are generally analysed and predicted using three main approaches, the weight matrix, neural networks and hidden Markov model (HMM). Each of these techniques has its own merits briefly described below.

Weight matrix is a matrix of score values that provides a weighted match to any given sub-sequence of fixed length, which has been used for amino acid sequences. Weight matrix attempts to specify the specific scores of cleavage position in the amino acid residues. The methodology was developed by von Heijne to classify the cleavage location of a signal peptide in a sequence of protein [39,27]. In order to predict, the matrix is matched against an unknown protein sequence to obtain a position with highest sum of weights denoted as the designated cleavage site [40]. Computational results demonstrate that weight matrix performance is inferior to neutral networks and HMM methodologies.

Wang et al. [41] further used a SVM to predict the cleavage sites of SPs from the protein sequences. They concluded that at small false positive ratios, the method outperforms the classical weight matrix method, indicating the proposed technique may serve as a complementary tool to other methods for predicting the SP cleavage sites.

Conventional AI techniques use a moving-segment known as the "sliding window" to analyse the amino acid sequences for the purpose of protein segment identification where the window continuously determines an encoding technique such as hydrophobicity [42,43]. Neural networks are used with sequence features extracted for sliding window operation. A variable or a fixed sized window is selected with amino acid residues on both sides to move from left to right of a given protein sequence to identify a transmembrane region or signal peptide. For a signal peptide, each position of the sliding window presents a numerically encoded feature vector to detect if the window's centre actually contains a cleavage site. The technique employs dual window configurations called symmetric and asymmetric using 20 amino acid residues and a distributive encoding technique. We have adopted this approach in the present study, based on previous work [42,43].

In a hidden Markov model (HMM), the state is not directly visible, but the output is visible and depends on the state [44]. Each residue in a protein sequence is based on a probabilistic distribution of previous states and each state has a probability of showing a set of observable features. In a protein sequence, these observation probabilities can represent the three regions of a signal peptide relative to the cleavage site. The most likely cleavage location is therefore obtained by the probabilistic transitional path within the amino acid residues aligning with the cleavage site node. HMM can also be used to incorporate biological knowledge into the signal peptide or transmembrane modelling [13]. Neural networks usually require numerical encodings. Since the HMM inputs are symbolic, the numerical encodings similar to neural networks are not required. There is a wide range of encodings available and the prediction accuracy predominantly depends upon the type of encoding used [45,46].

2.1. A dual phase SVM-NN methodology to predict signal peptide topology

In this paper, a signal peptide topology prediction could be divided into two distinct problems:

- The discrimination of signal peptide from globular proteins using SVM;
- The prediction of signal peptide cleavage site using NNs.

A SVM method will initially differentiate whether a protein sequence has a SP region or not using symmetric sliding window. If the protein sequence has SP, a NN method will predict the location of cleavage site where the mature protein begins using asymmetric sliding window. SVM-NN combination approach has been never used in the research community to discriminate SP and predict cleavage site.

2.1.1. Feature extraction

Fig. 1 shows a symmetric sliding window of feature extraction based on propensity features, used for discrimination of SP from mature protein or Non-SP protein. The technique extracts a sequence of features from each sample which slides from left to right, +1 position at a time. Each test sliding window is assigned a score ranging from 0 to 1 to ascertain a potential candidate as being closest to the cleavage site. For example in

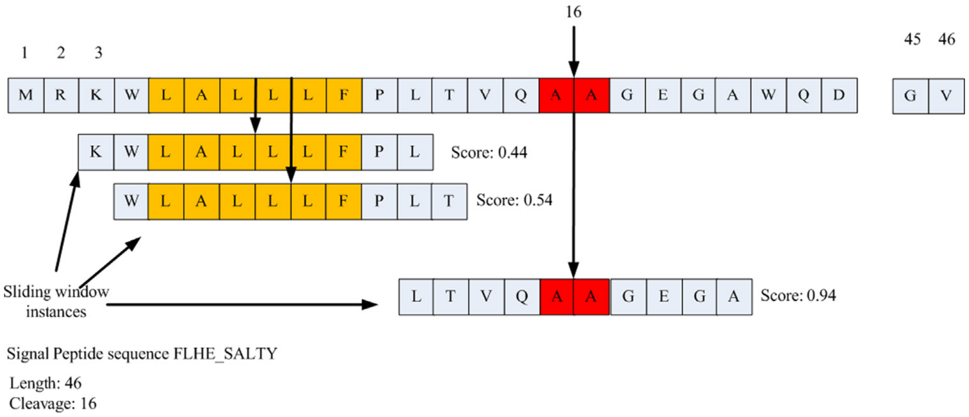


Fig. 1. Feature extraction using symmetric sliding window size −13 to +13 (*Note*: Here −5 to +5 are shown) for discrimination of SP and Non-SP from a single training instance of the flagellar protein of *S. typhimurium* (FLHE_SALTY) amino-acid sequence.

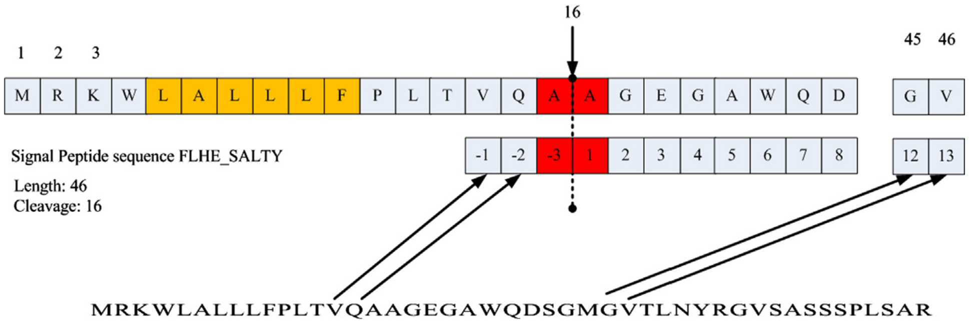

16
1 2 3
45 46
M R K W L A L L L F P L T V Q A A G E G A W Q D
G V
Signal Peptide sequence FLHE_SALTY
Length: 46
Cleavage: 16
-1 -2 -3 1 2 3 4 5 6 7 8
12 13
MRKWLALLLFPLTVQAAGEGAWQDSGMGVTLNYRGVSASSSPLSAR

Fig. 2. Feature extraction using asymmetric sliding window size −3 to +13 for cleavage site identification from a single training instance of FLHE_SALTY amino-acid sequence.

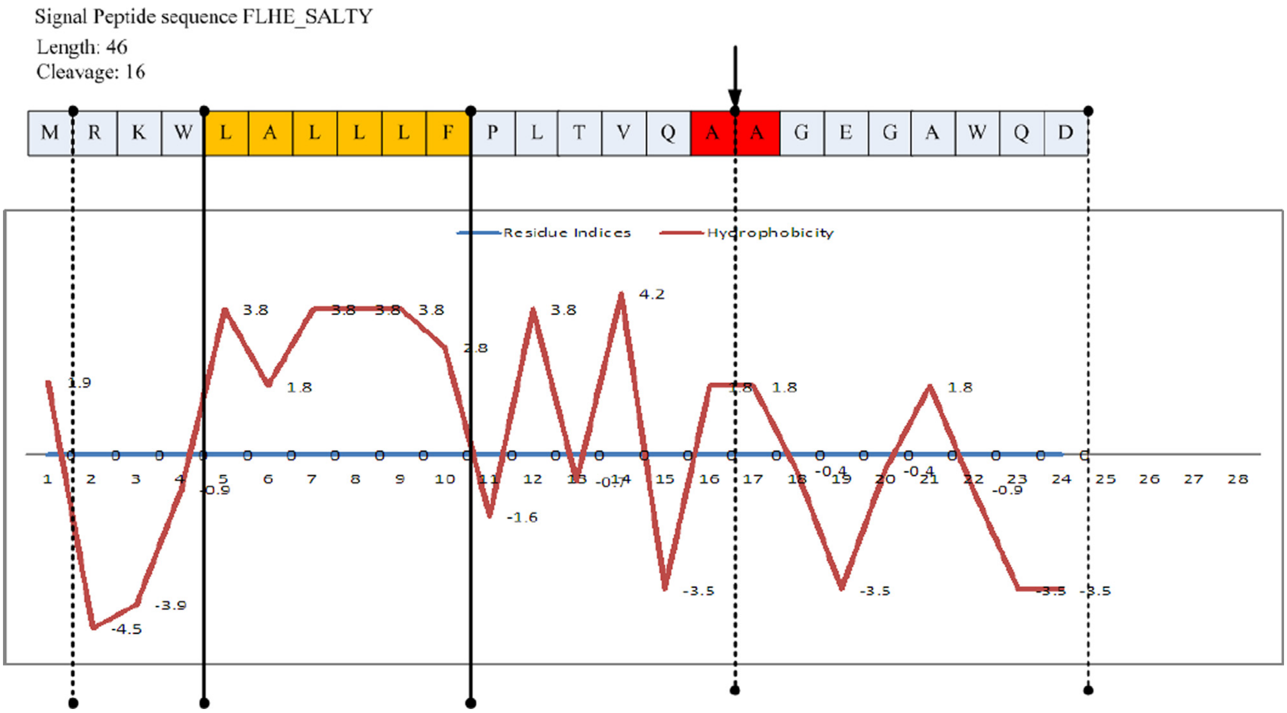


Fig. 3. Hydrophobicity and cleavage site position in the signal peptide of the flagellar protein of *S. typhimurium* (FLHE_SALTY).

Table 1
Data selection – symmetric and asymmetric sliding windows.

	Discrimination	**Cleavage prediction**
Data selection	Eukaryotic, Bacterial, Gram +ve, Gram −ve	Eukaryotic, Bacterial, Gram +ve, Gram −ve
Sliding window format/length	−13, +13	−3, +13

Fig. 1, the scoring accuracy increases from 0.44 to 0.54 up to the closest value to 1, 0.94, where the cleavage site is located. Hydrophobicity present in the signal peptide is highlighted from residue L up to residue F.

Fig. 2 shows the feature extraction technique using an asymmetric window size −3 to +13 to pull out features from the signal peptide sequence of FLHE_SALTY to identify the cleavage site.

Fig. 3 demonstrates the signal peptide region and the beginning of mature protein starting from the cleavage site A A 16th residue onwards for the flagellar protein of *Salmonella typhimurium* (FLHE_SALTY). SP is structured as such that after N-terminal a slightly positively charged n-region is located. In Fig. 3, M, R, K and W amino acids represent the n-region. The hydrophobic region in SP is called 'h-region' appears after n-region, which is represented by the positive shaded area of six amino acids in Fig. 3. After h-region an uncharged c-region is positioned, where the cleavage point is located. In Fig. 3, the positive peak A A shows the cleavage site location.

2.2. Separating cytoplasmic TM helical and signal peptide data

Conventional transmembrane prediction methods generally predict signal peptides as being transmembrane regions, due to the presence of hydrophobic segments within both SP and TM regions. TM topology classifiers identify SPs as helical segments, whereas the SP classifiers predict N-terminal helices as SPs. Because of this cross-prediction, the proposed research aims to predict the presence of signal peptide discrimination and cleavage site identification in protein sequences separately using symmetric and asymmetric sliding windows, from a variety of organisms, as shown in Table 1.

2.3. Extraction of data

The training and testing sequence information were extracted from the Swiss-Prot database (Uni-Prot release 2012_07). Uni-Prot is a comprehensive, high-quality and freely available database contains detailed information about the biological functions of proteins. The dataset was initially constructed with the keyword "signal sequence", using the advanced search option which generated 36,718 reviewed entries and 422,411 un-reviewed entries. The reviewed entries were further reduced to 5134 to eliminate sequence annotations such as confidence level of 'probable', 'potential' and 'by_similarity'. The reviewed entries containing more than one cleavage site were also removed, as were archeal or viral proteins, by including only eukaryota or bacteria in the organism classification line, leaving a total 4919 entries. The resulting entries were further grouped into two separate databases as follows:

- Group 1: Eukaryotic (Eukaryota in the organism classification line) – 3845 entries.
- Group 2: Prokaryotic (Bacteria in the organism classification line) – 1074 entries.

Eukaryotes are organisms where cells contain nucleus and prokaryotes are organisms whose cells do not have membrane-bound nucleus. Groups 1 and 2 were further classified as follows:

Eukaryotic

- All organelle proteins (sequence entries with organelle lines) were removed, leaving 3839 entries.
- Only the signal peptide sequences between 15 and 45 residue lengths were accepted (3812 entries).

- Only the proteins with A, G, S, C, T, P, L and Q at position '−1' were accepted [32]. This classification was achieved using computer programming.

Prokaryotic

- All the prokaryotic lipoproteins were removed.
- Only the signal peptide sequences between 15 and 50 residue lengths were accepted (1017 entries).
- Only the proteins with A, G, S, T, and V at position '−1' were accepted [32]. The classification was obtained using computer programming.

The dataset of negative entries was prepared by selecting the first 70 entries based upon the taxonomic lineage, stating the term 'eukaryota' and cellular component containing 'cytoplasm', 'cytosol' and 'nucleus' entries, eliminating fragments and sequences shorter than 70 amino-acid residue length, as follows:

- The N-terminal parts of eukaryotic cytoplasmic sequences with entries taken from the complete proteome set, was reduced by 50% to 3283.
- Nuclear eukaryotic proteins (2265 entries).
- Prokaryotic (bacterial) cytoplasmic proteins reduced to 987.

Table 2
UniProt query to extract model data from SwissProt database release.

```
Eukaryotic
keyword:"Signal sequence"
NOT annotation:(type:signal confidence:by_similarity)
NOT annotation:(type:signal confidence:potential)
NOT annotation:(type:signal confidence:probable)
AND (taxonomy:eukaryota)
AND annotation:(type:positional length:[15 TO 45])
Prokaryotic
keyword:"Signal sequence"
NOT annotation:(type:signal confidence:by_similarity)
NOT annotation:(type:signal confidence:potential)
NOT annotation:(type:signal confidence:probable)
AND (taxonomy:bacteria)
NOT keyword:"Lipoprotein [KW-0449]"
AND annotation:(type:positional length:[15 TO 50])
Negative samples (Eukaryotic-cytoplasmic)
uniprot:
((
NOT annotation:(type:signal AND length:[00000 TO 00070]
  AND confidence:probable)
NOT annotation:(type:signal AND length:[00000 TO 00070]
  AND confidence:potential)
NOT annotation:(type:signal AND length:[00000 TO 00070]
  AND confidence:by_similarity)
AND keyword:cytoplasmic
AND taxonomy:eukaryota
NOT fragment
NOT length:[00000 TO 00070]
AND reviewed:yes)
AND keyword:181)
identity:0.5
Negative samples (Prokaryotic-cytoplasmic)
uniprot:
((
NOT annotation:(type:signal AND length:[00000 TO 00070]
  AND confidence:probable)
NOT annotation:(type:signal AND length:[00000 TO 00070]
  AND confidence:potential)
NOT annotation:(type:signal AND length:[00000 TO 00070]
  AND confidence:by_similarity)
AND keyword:cytoplasmic
AND taxonomy:bacteria
NOT fragment
NOT length:[00000 TO 00070]
AND reviewed:yes)
AND keyword:181)
identity:0.5
```

Table 3
The reduced datasets used for SP discrimination and cleavage site predictions. +30 means the dataset taken for SP extends to 30 residues from cleavage position, where for Non-SP proteins, the first 70 residues of globular proteins are used.

	Signal peptide		Cytoplasmic proteins		Nuclear proteins	
	Number of sequences	τ	Number of sequences	τ	Number of sequences	τ
Eukaryotic	1784	+30	3283	70	2265	70
Prokaryotic	646	+30	987	70	N/A	N/A

Table 2 lists the UniProt query commands used to obtain the data.

For eukaryotes a dataset of 1784 SP entries was used, extending to +30 residues from the cleavage position. Furthermore for prokaryotic with bacteria, the dataset of 646 entries were used for SP extending to +30 residues from the cleavage position (Table 3). In order to remove the bias from negative and positive sets in neural network training, the datasets were reduced to an approximately equal size. The datasets were evaluated using Matthews Correlation Coefficient (MCC) and cross-validation tests where the data was put into amino-acid groups. The computer modelling was developed using two separate classification techniques where SVM was utilised for the topological discrimination and NNs were used to model the cleavage site location. The window types for SVM and NN were a symmetric 26 amino acid window and asymmetric 16 amino acid window respectively, extracting the propensity values based on the following formula [46,47]:

$$P_i = F_i^{TM}/F_i \tag{1}$$

In the above equation, P_i represent the propensity of each amino-acid i to appear in a transmembrane sequence. The F_i^{TM} and F_i are the frequencies of the ith residue to appear in transmembrane and non-transmembrane region respectively calculated using the recent Swiss-Prot database entries. P_i value greater than '1' suggests the inclination or propensity of residues to appear in transmembrane region and less than '1' indicates non-transmembrane region.

3. Support vector machine and neural network architectures

Broadly speaking, an AI technique will discriminate whether a protein sequence has a SP region or not using symmetric sliding window. If the protein sequence does not possess a SP region, then the sequence is a mature protein. If the protein sequence has SP, another AI technique will identify the exact location of cleavage site of the SP where the mature protein begins using asymmetric sliding window. In this paper, for the discrimination of SPs from Non-SPs proteins the SVM methodology is used and for cleavage location prediction, the neural networks methodology is utilised. The underlying reason for using different classification techniques is that ANNs are well-known to handle large datasets [48]. Therefore, a neural network technique could be more suitably modelled for the cleavage-site prediction with large datasets. In contrast, SVM is known to handle smaller datasets [16,24,25]. SVM models are known to perform well for a dual-class problem, which is one of the two goals of the proposed research. Moreover, hyper-plane fitting in SVM for a two-class separation with a large number of data points is computationally intensive and is therefore not recommended for large data cases, unless a technique to reduce the database size is implemented.

3.1. SVM architecture for dual SP and Non-SP classification

SVM has previously been used to predict the presence of helical segments in amino acid sequences [49,50]. The application of SVM in SP topological prediction has largely been focussed on differentiating N-terminal SPs from alpha-helical segments [51] primarily due to the fact that both segments share hydrophobic traits. Generally, in real-world applications, SVM is known to outperform or match identification accuracy for most benchmarking problems [52,53]. As discussed, contrary to neural networks, SVM is known to outstandingly model problems with smaller data size sample. This makes the SVM methodology an ideal technique for beta-barrel and signal peptide prediction problems, where the modelling is usually undermined by a smaller database or uneven class sample distribution problems respectively [54].

For SVM, the issue of SP/Non-SP discrimination is a two-class problem to categorise proteins with signal peptides from globular proteins with no signal peptides. The discrimination plays a crucial role in helical and barrel proteins predictions and a correct classification improves the overall accuracy of transmembrane domains predictions. One example of feature extraction is shown in Fig. 1, where the signal peptide class feature vector is extracted up to 30 residues into the mature protein, exceeding the signal peptide part.

The reduced datasets for discrimination and cleavage site predictions are shown in Table 3. τ is the index of +30 residues from cleavage site position onwards into the mature protein for discrimination of SP and Non-SP. For Non-SP proteins, τ is the first 70 residues of globular proteins. The samples for this two class problem are described by the feature vector x_i where $i=1, 2, \ldots, N$. N is +30 from cleavage point onwards and 70 is for negative sets Non-SP proteins with corresponding labels $y_i \epsilon \{+1, 0\}$. This study represents SP protein class as +1 and Non-SP class as 0. In order to predict the two class representations, SVM model learns by mapping the input space into a higher dimensional domain and then fitting a hyperplane to the domain. The hyperplane represented by $\alpha_i$is obtained by the following equation [55]:

$$\sum_{i=1}^{N} \alpha_i y_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \gamma_{i,j} \tag{2}$$

where $\gamma_{i,j} = (x_i,\ x_j)$ signifies a kernel function to map input classification to a higher dimensional feature space represented by a radial basis function $\exp\left(-\gamma \|x_i, x_j\|^2\right)$. A radial basis function is a real-value function whose value depends on the distance from the origin.

For N samples, the above equation is represented as

$$\sum_{i=1}^{N} \alpha_i y_i = 0 \quad (0 \le \alpha_i \le M) \tag{3}$$

where M is the argument controlling the relationship between the margin and classification error and y_i is either +1 or 0 indicating the SP and Non-SP classes to which the points x_i belong.

A hyperplane separates the two class representations of SP and Non-SP. A hyperplane is a set of points x_i which satisfies the equation $wx+c=0$, where w is regarded as the normal to the hyperplane and $|c|/\|w\|$ is the perpendicular distance from the line to the origin. w represents the Euclidean norm of w. The idea is to generate a maximum margin between Non-SP class and SP class using the support vector algorithm shown in Fig. 4. Suppose f^1 and f^2 are the distances separating the two class samples Non-SP and SP {0 and +1} respectively. Eqs. (4) and (5) for the training data in this research satisfy the following constraints:

$$x_i.w + c \ge 1 \text{ for } y_i = 1(f^1;\ \text{Non SP}) \tag{4}$$

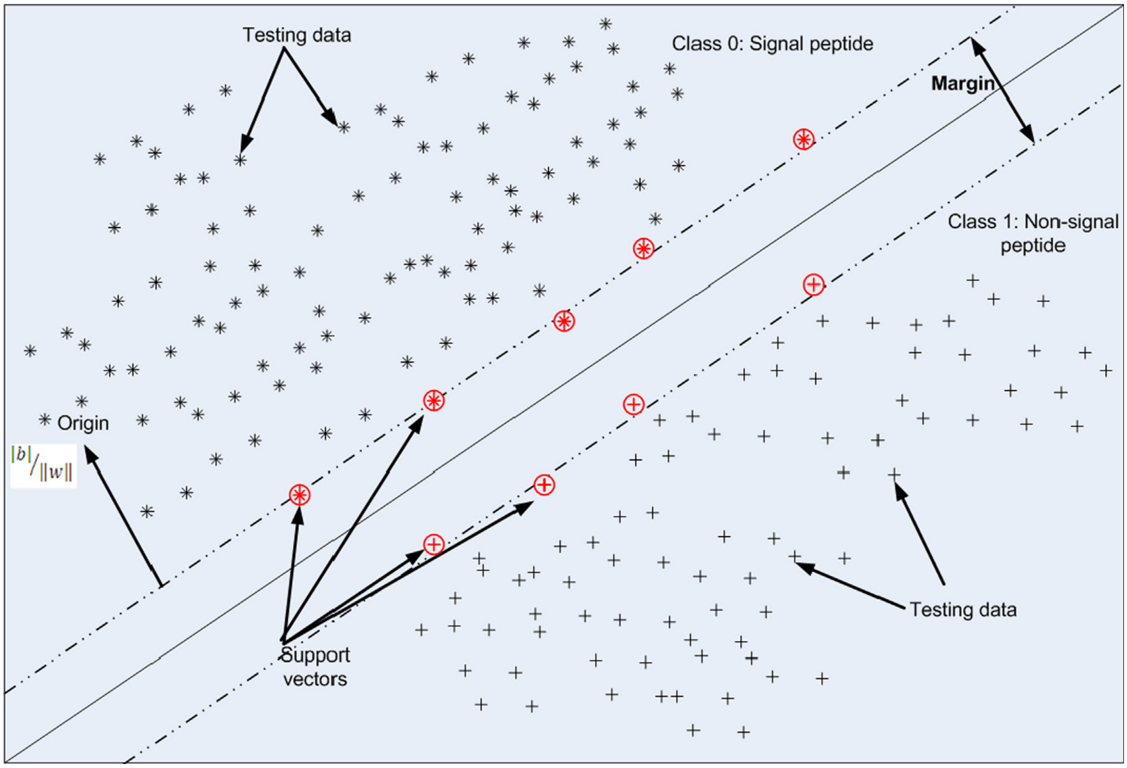


Fig. 4. Broken lines hyperplanes linearly separating SP and Non-SP. The circled stars and crosses show support vectors.

$$x_i w + c \le 0 \text{ for } y_i = 0\ (f^2;\ \text{SP}) \quad (5)$$

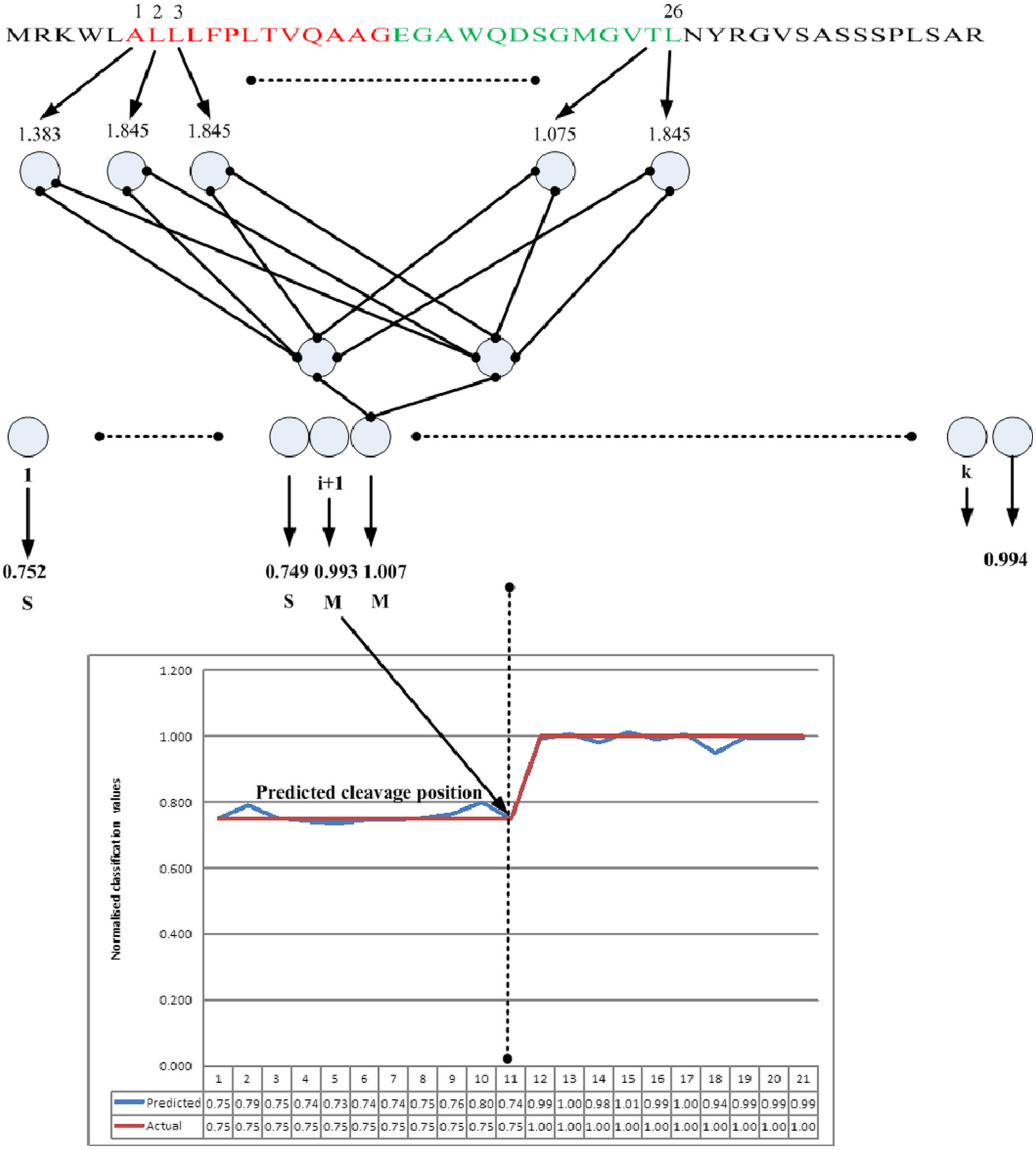
Considering (4) and (5), the points $H_1 : x_i w + c = 1$ and $H_2 : x_i w + c = 0$ lie on the margin hyperplane with normal w and the perpendicular distance from the origin to be $\frac{1}{2}\|w\|$' where c is zero. Therefore, $f^1 = f^2 = \frac{1}{2}\|w\|$ with a margin to be $1/\|w\|$. It must be noted that the hyperplanes shown in Fig. 4 are parallel with no testing points falling within the margin. Based on the above analysis, the objective is to obtain two hyperplanes for the two-class problem in order to maximise the margin by minimising $\|w\|$. In Fig. 4, those testing points that lean on the broken lines hyperplanes are called 'support vectors'.

3.2. Neural network architecture for cleavage site recognition

The proposed neural network model is dual-layer feed-forward architecture with an input layer, a hidden layer and an output layer. The network is regarded as a 'fully-connected' neural network. Each unit of a layer is connected to each unit in the next layer where each connection's strength is given by a weight w_{ij}, where i is the input layer and j is the hidden layer. In the network each inner layer node S_j is computed by the sigmoid function as follows [56]:

$$S_j = 1/1 + e^{-(w_{jo} + \Sigma_{i=1}^{n} w_{ij} s_i)} \quad (6)$$

The dataset comprised of a set of S_i asymmetric propensity values representing a single sliding window operation, where $i=16$ for -3 and $+13$ values on both sides of the network. w_{j0} is a bias from the states S_i of lower layers, where o is the output layer. The activation was performed using the equation given above and were fed forward via all the layers to the output. The process is further elaborated in Fig. 5. Numbers 1.383, 1.845, 1.845, 1.075 and 1.845 are propensity values, and 0.752, 0.749, 0.993, 1.007 and 0.994 are neural outcome scores. The graph in Fig. 5


1 2 3
26
MRKWLALLLFPLTVQAAGEGAWQDSGMGVTLNYRGVSASSSPLSAR
1.383
1.845
1.845
1.075
1.845
1
i+1
k
0.752
S
0.749 0.993 1.007
S M M
0.994
1.200
1.000
0.800
0.600
0.400
0.200
0.000
Normalised classification values
Predicted cleavage position

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
Predicted	0.75	0.79	0.75	0.74	0.73	0.74	0.74	0.75	0.76	0.80	0.74	0.99	1.00	0.98	1.01	0.99	1.00	0.94	0.99	0.99	0.99
Actual	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Fig. 5. A dual-layer feed-forward neural network training using sliding window based on propensity values. The output layer is represented by actual normalised decimal values of 0.75 for signal section and 1.00 for mature protein, and predicated values of 0.74 for signal peptide and 0.99 for mature protein.

demonstrates the predicated cleavage site by neural network and actual cleavage position at residue 11, S shows the signal peptide location and M shows the start of the mature protein. Pasquier and Hamodrakas [47] conducted a similar experiment and concluded that the optimal size of the hidden layer should be 30 neurons, whereas this study obtained the optimal results with 20 neurons. Table 4 demonstrates convergence with 20 neurons in the hidden layer and compares the results for regression and mean square error with other number of neurons. For regression analysis 20 neurons in the hidden layer provides validation of 0.9317 and for mean square error analysis it gives validation of 0.0367. It should be noted that the regression analysis was only carried-out at the initial model training phase using *K*-fold based testing, described in Section 4.1 below.

Fig. 6 further shows a dataset trained with a 20 neurons neural network generates an outstanding mean square error outcome, which is also outlined in Table 4. Fig. 6 shows that the error outcome for validation is substantially improved as compared to testing evaluation, although the overall training evaluation is better.

Table 4
Comparison of the optimal neural training convergence using 20 neurons with 10, 30, 80, 160 and 240 neurons.

Neurons	10	20	30	80	160	240
Regression						
Training	0.916	0.9545	0.959	0.94023	0.9774	0.8681
Validation	0.92	0.9317	0.924	0.91251	0.8579	0.8735
Test	0.903	0.9218	0.9225	0.91421	0.8537	0.8603
Mean square error						
Training	0.0401	0.0222	0.0206	0.029	0.0119	0.062
Validation	0.0383	0.0367	0.037	0.041	0.076	0.059
Test	0.046	0.0375	0.0377	0.041	0.075	0.657

4. Results of computer simulation and testing

In order to generate robust outcomes, prediction methodologies are generally evaluated by the re-substitution tests, independent dataset tests, MCC, jack-knife tests [50], cross-validation or self-consistency tests. Cross-validation based testing is considered to be the most objective for any prediction methodologies requiring accurate performance estimation [57]. Cross-validation reduces the computational time and is adopted by many investigators using SVM as a prediction engine. Cross-validation analysis involves testing each protein group in the dataset against the trained model, then putting the data back and removing another protein group and repeating the analysis. To evaluate a global performance of the system, the model is repeated for every protein group and the average is reported for the whole dataset. MCC is used in machine learning as a means of two-class binary classifications. The MCC is in essence a correlation coefficient between the observed and predicted binary classifications, which is very suitable for two-class binary discrimination.

4.1. SVM-NN based signal peptide discrimination and cleavage site prediction

As shown in Fig. 5, the cleavage site position is determined by a dual-layer feed-forward neural network utilising sliding windows technique based on propensity values. The output layer produces outstanding results, which is represented by actual normalised decimal values of 0.75 for SP section and 1.00 for mature protein, and the predicated values of 0.74 for SP and 0.99 for mature protein. The overall training procedure was based on a dual-level regression classification training process. In the first phase, 50% of all training datasets were passed through a regression process to perform a regression-based NN training. This half of the training datasets were split into 70%, 15%, 15% data segments based on

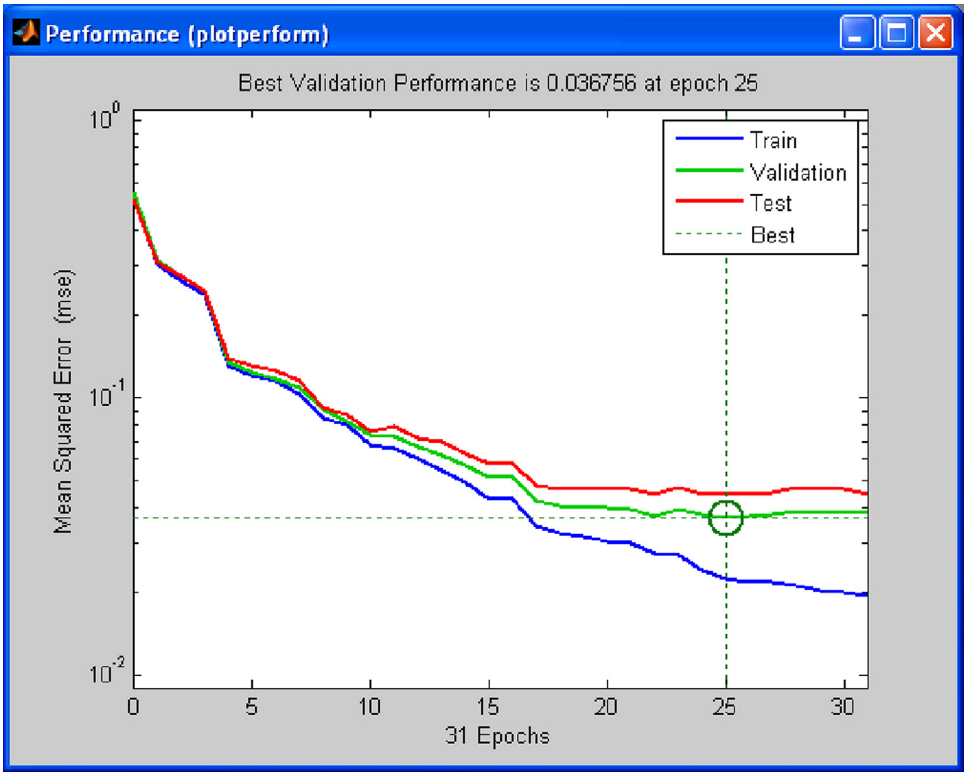


Fig. 6. Validation training performance based on 20 neurons and a back propagation feed-forward neural network.

random pointers. 70% of these training datasets were regressively trained for adjustment of the network error. 15% was used for network generalisation as validation datasets, where the training was stopped when the generalisation ceased improving for 5 consecutive epochs. 15% of the remaining data was used to test the accuracy of the trained model. The regression-based training stage was objectively used to evaluate the reliability of the trained model. The regression values were iteratively obtained by retraining and the process was stopped when a regression of greater than 0.9 was obtained demonstrating a high correlation between the target and outcome values. Once phase 1 was accomplished, the underlying NN model was saved. It was at this stage where the actual cleavage site identification was performed against the remaining 50% of all datasets by 5-fold data evaluation. The overall data was divided into cross-validation based five protein groups with each group was tested over a trained model on the remaining four groups. Cross-validation analysis was used where each protein group was set aside for testing and the model was trained using the remaining protein groups belonging to nuclear, cytoplasm, eukaryotic and prokaryotic. The cleavage site position accuracy was obtained by averaging the five cross-validation based

Table 5
Discrimination and cleavage site prediction of signal peptide based on Matthews correlation coefficient and cross-validation.

Type	Group	Discrimination (MCC)	Cleavage site (cross-validation)
Non-SP	Nuclear	0.92	–
	Cytoplasmic	0.88	–
SP	Eukaryotes	0.91	91%
	Prokaryote (Bacterial)	0.89	92%

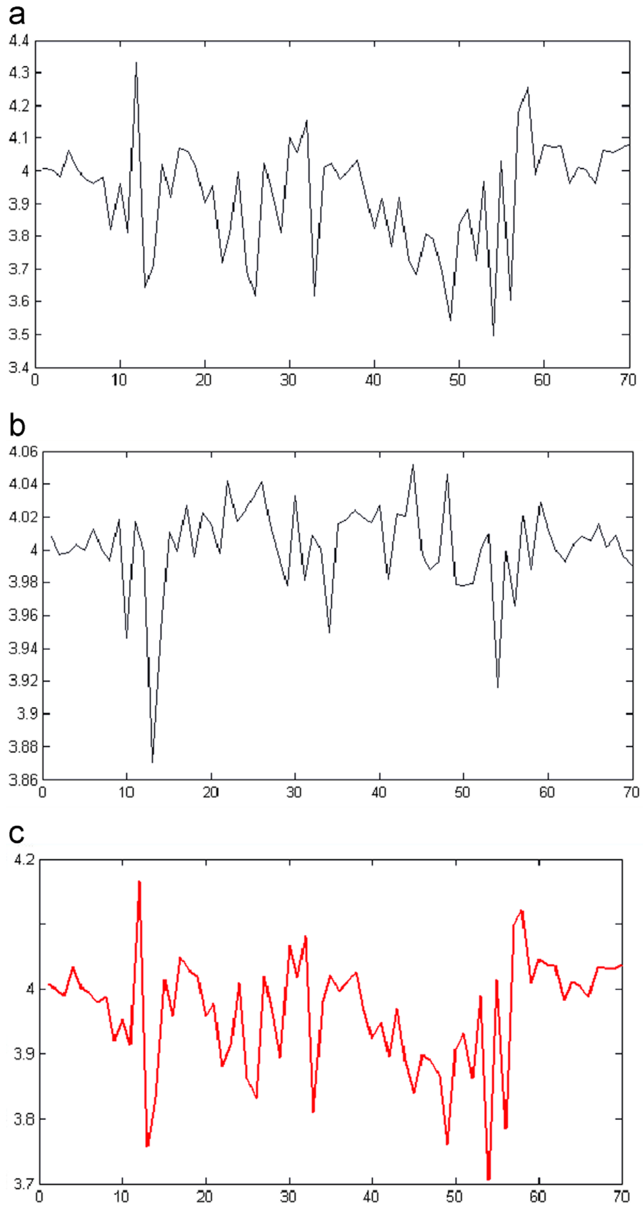


Fig. 7. Evaluation of a cytoplasmic protein mouse ataxin 2 (ATX2_MOUSE) using an ensemble of SVM-NN classifications. S-score is given in (a), C-score is shown in (b), and when combined in (c) the Y-score shows many peaks characteristic of a cytoplasmic protein with Non-SP.

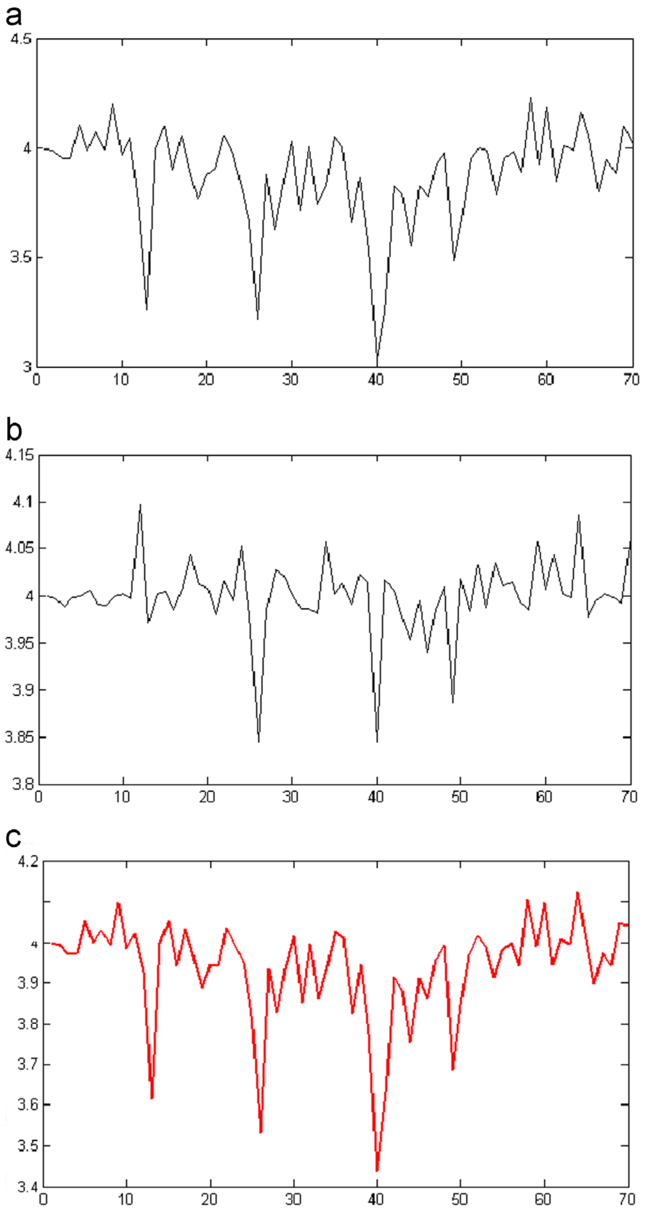


Fig. 8. Evaluation of a mouse nuclear protein (AKIP_MOUSE) using an ensemble of SVM-NN classifications. S-score is given in (a), C-score is shown in (b), and when combined in (c) the Y-score shows many peaks, characteristic of a Non-SP protein.

groups. The discrimination accuracy of SP and Non-SP was evaluated using MCC based ranking. Table 5 demonstrates discrimination and cleavage site prediction for signal peptide based on MCC tests and cross-validation tests for nuclear, cytoplasm, eukaryotic and prokaryotic (bacteria). The figures in Table 5 are rounded.

Computer simulations were carried out for the proposed SVM-NN model on unseen SP and globular proteins and the following results are obtained in Figs. 7–14 and Table 6. Fig. 7 shows the analysis of a mouse cytoplasmic protein ataxin 2 (ATX2_MOUSE) using an ensemble of SVM and NN classifications for SP/Non-SP discrimination and cleavage site prediction. Fig. 7 (a) shows the outcome of using SVM for discrimination of SP and Non-SP expressed as the S-score. Fig. 7(b) shows the use of NN for cleavage site prediction, named as the C-score. Fig. 7(c) is a combination of Fig. 7(a) and (b) used to ascertain the precise location of the cleavage site and it is called the Y-score. In Fig. 7, for cleavage site prediction one looks for one single peak and for SP discrimination one looks for many peaks. Since Fig. 7(c) has many peaks, it clearly demonstrates a Non-SP cytoplasmic protein. Fig. 8 is an example of a nuclear protein (AKIP_MOUSE) where again multiple inverted peaks demonstrate a globular Non-SP protein. Fig. 8(a) again demonstrates the discrimination of SP and Non-SP using SVM, represented by the S-score. Fig. 8(b) shows the use of NN for cleavage site prediction, denoted as the C-score. Fig. 8(c) Y-score is a combination of Fig. 8(a) and (b) utilised to determine the precise location of the cleavage site, but it provides many peaks representing a globular Non-SP protein. In contrast, Fig. 9 shows the cleavage site position of a eukaryotic type I membrane protein with an SP (HA2Q_HUMAN) protein represented by a single peak, which matches 100%. In Fig. 9(c), the Y-score clearly indicates a cleavage site at amino acid 31 within the 60 residue length, where the mature protein begins [32]. The dataset was further evaluated for general bacterial sequences, and bacterial sub-classes Gram −ve and Gram +ve; three samples are shown in Figs. 10–12, respectively. Figs. 10 and 11 show two examples, evaluations of a prokaryotic (bacterial) and of a Gram −ve sequence respectively, where the inverted peaks demonstrate Non-SP cytoplasmic proteins in both cases. Fig. 12 shows the cleavage site position of a Gram +ve bacterial sequence represented by a single peak, where the S-score and the C-score match 100% and both indicate a cleavage site at amino acid 27 within the 70 residue length.

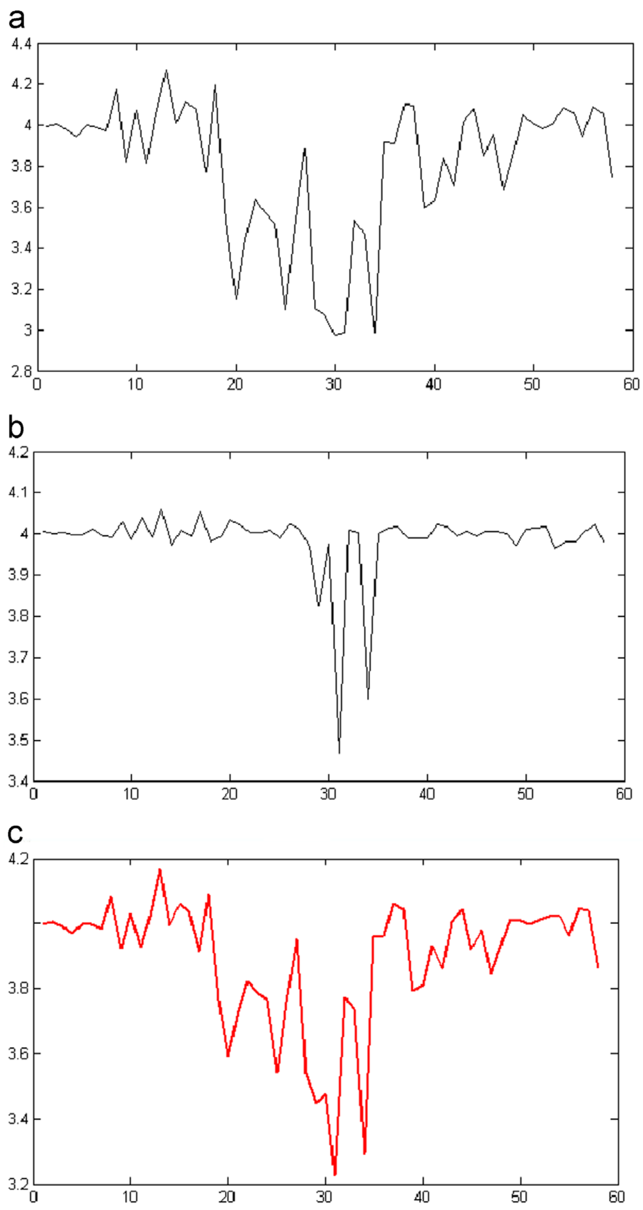


Fig. 9. Evaluation of a human type I membrane protein, a class I MHC antigen (HA2Q_HUMAN) which contains a signal peptide using an ensemble of SVM-NN classifications. A mapping of hydrophobic propensity transition in S-score (a) and crisp cleavage mapping in C-score (b) produces an accurate combined cleavage site Y-score in (c).

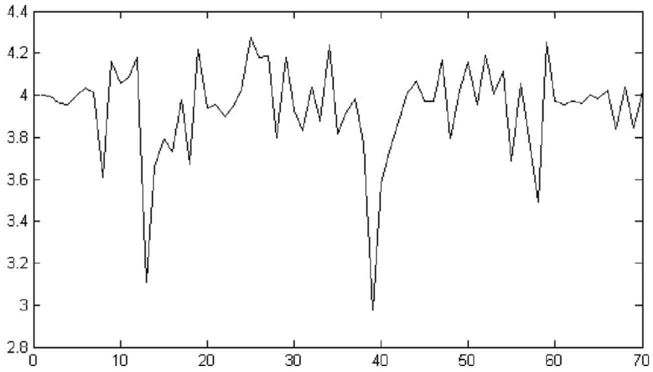


Fig. 10. Evaluation of a cytoplasmic bacterial protein sequence (CHEB3_LEPIC) using an ensemble of SVM-NN classifications. C-score is shown.

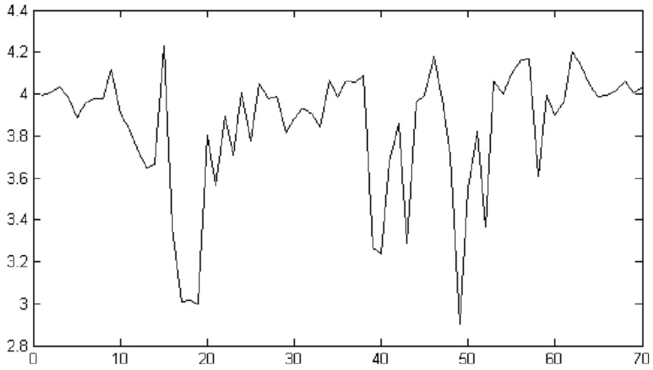


Fig. 11. Evaluation of a cytoplasmic protein sequence, asparaginase (ASPG_WOLSU), from the Gram −ve cytoplasmic bacterial database, using an ensemble of SVM-NN classifications. C-score is shown.

The research further used a various combination of SP proteins selected out of 646 non-redundant proteins for SP and Non-SP identification based on MCC tests using SVM, and cleavage site prediction based on cross-validation evaluation using SVM-NN technique. Table 6 shows SP and Non-SP discrimination and

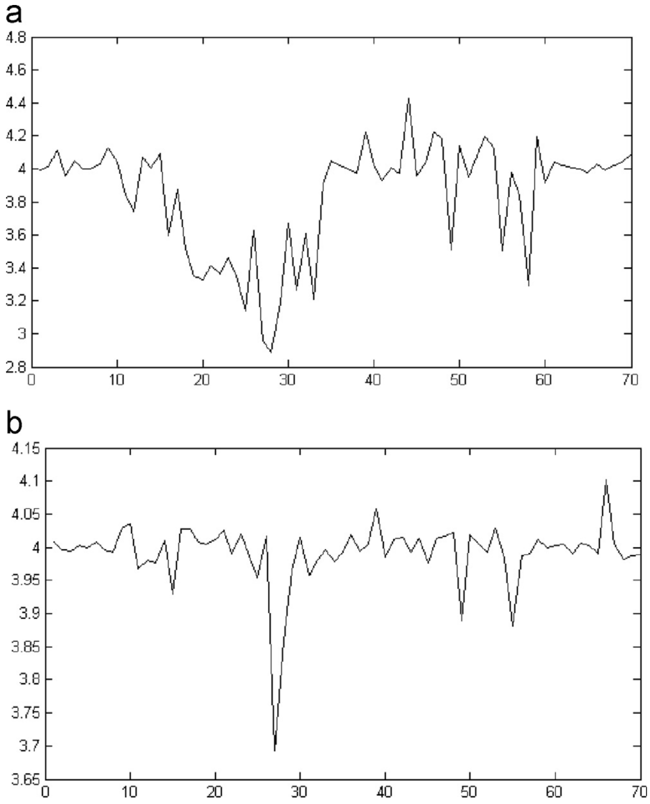


Fig. 12. Evaluation of a SP containing sequence (CDG2_BACMA) taken from the Gram +ve bacterial SP database using an ensemble of SVM-NN classifications. A mapping of the hydrophobic propensity transition in S-score (a) and crisp cleavage mapping in C-score (b) are shown.

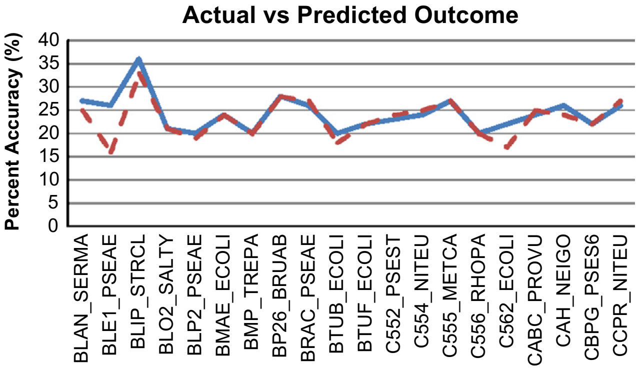


Fig. 13. Comparison of actual (continuous lines) and predicted (broken lines) cleavage site locations for 20 SP proteins.

cleavage site prediction for a combination of 20 SP proteins which are chosen out of 646 non-redundant proteins. Table 6 demonstrates the accuracies for discrimination and cleavage site prediction for each protein and indicates an overall accuracy of 90% for SP and Non-SP discrimination for MCC based evaluation using SVM. Furthermore, Table 6 shows an accuracy of 94.0% for cleavage site prediction for cross-validation based evaluation using SVM-NN. In order to remove the bias from negative and positive sets in neural network training, the datasets were reduced to an approximately equal size. Additionally, computer simulation results further confirm that a combination of 20 cytoplasmic proteins selected out of 646 non-redundant proteins for SP and non-SP discrimination produces an accuracy of 90% for MCC based evaluation using SVM. Fig. 13 further uses the same 20 SP proteins from Table 6 and compares the predicted and the actual outcomes for cleavage site prediction. Fig. 13 shows that the predicted cleavage site predictions closely match the actual ones, demonstrating the high accuracy of the proposed SVM-NN model. Fig. 14 demonstrates the percentage of cleavage site prediction accuracies for the same 20 SP proteins from Table 6. Apart from one protein,

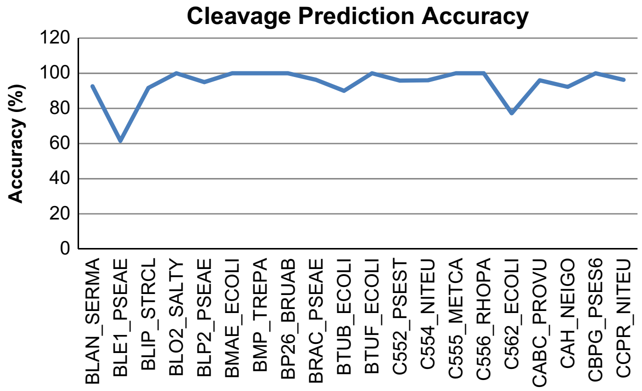


Fig. 14. Accuracy of cleavage site predictions for 20 bacterial SP proteins using SVM-NN.

Table 6
A combination of 20 SP proteins were selected out of 646 non-redundant proteins for cross-validation based evaluation.

	Type	Protein	Discrimination accuracy with MCC tests	Cleavage site accuracy with cross-validation tests
1	Signal Peptide	BLAN_SERMA	Y	92.6
2	Signal Peptide	BLE1_PSEAE	Y	61.5
3	Signal Peptide	BLIP_STRCL	N	91.7
4	Signal Peptide	BLO2_SALTY	Y	100
5	Signal Peptide	BLP2_PSEAE	Y	95
6	Signal Peptide	BMAE_ECOLI	Y	100
7	Signal Peptide	BMP_TREPA	Y	100
8	Signal Peptide	BP26_BRUAB	Y	100
9	Signal Peptide	BRAC_PSEAE	Y	96.3
10	Signal Peptide	BTUB_ECOLI	Y	90
11	Signal Peptide	BTUF_ECOLI	Y	100
12	Signal Peptide	C552_PSEST	Y	95.8
13	Signal Peptide	C554_NITEU	Y	96
14	Signal Peptide	C555_METCA	Y	100
15	Signal Peptide	C556_RHOPA	N	100
16	Signal Peptide	C562_ECOLI	Y	77.3
17	Signal Peptide	CABC_PROVU	Y	96
18	Signal Peptide	CAH_NEIGO	Y	92.3
19	Signal Peptide	CBPG_PSES6	Y	100
20	Signal Peptide	CCPR_NITEU	Y	96.3
			90	94.0

BLE1_PSEAE, the percentage of accuracy of the proposed SVM-NN prediction tool is very high.

5. Conclusion

The proposed methodology is based on the novel idea of applying SVM to SP and Non-SP discrimination and NN to SP cleavage site prediction, and the hybridisation of these two classification outcomes using an ensample-based SVM-NN classifier. The idea is to eliminate false cleavage site selections due to very high scores presented by the NN classifier because of the preceding hydrophobic segments. Therefore, in the presence of a dual SVM-NN classifier, higher false positive scores from both classifiers are rejected.

SignalP 4.0 with hidden Markov model method performs better than SignalP 3.0 for SP and Non-SP discrimination. The computer simulation results show that the proposed SVM technique outperforms SignalP 4.0 for SP and Non-SP discrimination using eukaryotic and prokaryotic proteins based on MCC tests with averaged accuracies of 0.90 for the SVM and 0.85 for SignalP 4.0. The SVM technique for MCC based testing also produces promising results for Non-SP detection for nuclear and cytoplasmic data with accuracies of 0.92 and 0.88 respectively. Furthermore, the proposed dual SVM-NN classifier performs better than SignalP 4.0 for cleavage site identification using eukaryotic and prokaryotic proteins based on cross-validation tests with accuracies of 91–92% for the SVM-NN and 66–83% for SignalP 4.0.

Finally, the simulation results demonstrate that a combination of 20 SP proteins selected out of 646 non-redundant proteins for SP and non-SP discrimination produces an accuracy of 90% for MCC based evaluation using the SVM and an accuracy of 94% for cleavage site prediction based on cross-validation evaluation using the proposed SVM-NN model.

Conflict of interest statement

None declared.

Acknowledgement

This work was funded by the Emerald Fund UK under Grant reference no. 2009-2S-017. This work was carried out at Intelligent Systems Research Centre, London Metropolitan University, UK.

References

[1] W. Neupert, R. Lill, Membrane Biogenesis and Protein Targeting, New Comprehensive Biochemistry, vol. 22, Elsevier, 1992.

[2] J-M. Chen, C. Ferec, Molecular basis of hereditary pancreatitis, Eur. J. Hum. Genet. 8 (2) (2000) 473–479.

[3] L. Kall, A. Krogh, E.L.L. Sonnhammer, A combined transmembrane topology and signal peptide prediction method, J. Mol. Biol. 338 (5) (2004) 1027–1036.

[4] A. Kessel, N. Ben-Tal, Introduction to Proteins: Structure, Function, and Motion, Chapman & Hall/CRC, Mathematical & Computational Biology, 1st Ed. CRC Press Taylor and Francis Group - A Chapman and Hall book, 2011.

[5] J.A. Hiss, G. Schneider, Architecture, function and prediction of long signal peptides, Brief. Bioinform. 10 (5) (2009) 569–578.

[6] K.C. Chou, Review: prediction of protein signal sequences, Curr. Protein Pept. Sci. 3 (6) (2002) 615–622.

[7] K.C. Chou, H.B. Shen, Signal-CF: a subsite-coupled and window-fusing approach for predicting signal peptides, Biochem. Biophys. Res. Commun. 357 (3) (2007) 633–640.

[8] H.B. Shen, K.C. Chou, Signal-3L: a 3-layer approach for predicting signal peptide, Biochem. Biophys. Res. Commun. 363 (2) (2007) 297–303.

[9] D.Q. Liu, H. Liu, H.B. Shen, J. Yang, K.C. Chou, Predicting secretory protein signal sequence cleavage sites by fusing the marks of global alignments, Amino Acids 32 (4) (2007) 493–496.

[10] H. Liu, J. Yang, J.G. Ling, K.C. Chou, Prediction of protein signal sequences and their cleavage sites by statistical rulers, Biochem. Biophys. Res. Commun. 338 (2005) 1005–1011.

[11] M. Arai, M. Ikeda, T. Shimizu, Comprehensive analysis of transmembrane topologies in prokaryotic genomes, Gene 304 (2003) 77–86.

[12] P.G. Bagos, T.D. Liakopoulos, T.D. Hamodrakas, Evaluation of methods for predicting the topology of beta-barrel outer membrane proteins and a consensus prediction method, BMC Bioinformatics 6 (7) (2005).

[13] J.D. Bendtsen, H. Nielsen, G. von Heijne, S. Brunak, Improved prediction of signal peptides: SignalP 3.0, J. Mol. Biol. 340 (4) (2004) 783–795.

[14] M. Gomi, M. Sonoyama, S. Mitaku, High performance system for signal peptide prediction: SOSUIsignal, ChemBio Inform. J. 4 (4) (2004) 142–147.

[15] Y.D. Cai, S.L. Lin, K.C. Chou, Support vector machines for prediction of protein signal sequences and their cleavage sites, Peptides 24 (1) (2003) 159–161.

[16] R.Y. Kahsay, G.R. Gao, L. Liao, Discriminating transmembrane proteins from signal peptides using SVM-Fisher approach, in: ICMLA 2005: Fourth International Conference on Machine Learning and Applications, Proceedings, 2005, pp. 151–155.

[17] P.L. Martelli, P. Fariselli, R. Casadio, An ENSEMBLE machine learning approach for the prediction of all-alpha membrane proteins, Bioinformatics 19 (2003) i205–i211.

[18] K. Melen, A. Krogh, G. von Heijne, Reliability measures for membrane protein topology prediction algorithms, J. Mol. Biol. 327 (3) (2003) 735–744.

[19] R. Clote, Performance, comparison of generalized PSSM in in signal peptide cleavage site and disulfide bond recognition, in: Third IEEE Symposium on Bioinformatics and Bioengineering – Bibe 2003, Proceedings, 2003, pp. 37–44.

[20] P. Fariselli, M. Finelli, I. Rossi, M. Amico, A. Zauli, P.L. Martelli, R. Casadio, TRAMPLE: the transmembrane protein labelling environment, Nucleic Acids Res. 33 (2005) W198–W201.

[21] J. Hawkins, M. Boden, The applicability of recurrent neural networks for biological sequence analysis, IEEE-ACM Trans. Comput. Biol. Bioinform. 2 (3) (2005) 243–253.

[22] D. Plewczynski, L. Slabinski, K. Ginalski, L. Rychlewski, Prediction of signal peptides in protein sequences by neural networks, Acta Biochim. Pol. 55 (2) (2008) 261–267.

[23] S.M. Reynolds, L. Kall, M.E. Riffle, J.A. Bilmes, W.S. Noble, Transmembrane topology and signal peptide prediction using dynamic Bayesian networks, PLoS Comput. Biol. 4 (11) (2008).

[24] J.J. Sun, L.P. Wang, Predicting signal peptides and their cleavage sites using support vector machines and improved position weight matrixes, in: Fourth International Conference on Natural Computation – ICNC 2008, 2008, pp. 95–99.

[25] T. Nugent, D.T. Jones, Transmembrane protein topology prediction using support vector machines, BMC Bioinformatics 10 (159) (2009).

[26] L. Zou, Z. Wang, Y. Wang, F. Hu, Combined prediction of transmembrane topology and signal peptide of beta-barrel proteins: using a hidden Markov model and genetic algorithms, Comput. Biol. Med. 40 (7) (2010) 621–628.

[27] G. von Heijne, A new method for predicting signal sequence cleavage sites, Nucleic Acids Res. 14 (1986) 4683–4690.

[28] R.J. Folz, J.I. Gordon, Computer-assisted predictions of signal peptidase processing sites, Biochem. Biophys. Res. Commun. 146 (2) (1987) 870–877.

[29] K.C. Chou, Using subsite coupling to predict signal peptides, Protein Eng. 14 (2) (2001) 75–79.

[30] K.C. Chou, Prediction of protein signal sequences and their cleavage sites, Proteins: Struct. Funct. Bioinform. 42 (1) (2001) 136–139.

[31] K.C. Chou, Prediction of signal peptides using scaled window, Peptides 22 (12) (2001) 1973–1979.

[32] H. Nielsen, J. Engelbrecht, S. Brunak, G. von Heijne, Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites, Protein Eng. 10 (1) (1997) 1–6.

[33] T.N. Petersen, S. Brunak, G. von Heijne, H. Nielsen, SignalP 4.0: discriminating signal peptides from transmembrane regions, Nat. Methods 8 (10) (2011) 785–786.

[34] W.Z. Lin, J.A. Fang, X. Xiao, K.C. Chou, iLoc-Animal: a multi-label learning classifier for predicting subcellular localization of animal proteins, Mol. BioSyst. 9 (4) (2013) 634–644.

[35] X. Xiao, P. Wang, W.Z. Lin, J.H. Jia, K.C. Chou, iAMP-2L: a two-level multi-label classifier for identifying antimicrobial peptides and their functional types, Anal. Biochem. 436 (2) (2013) 168–177.

[36] W. Chen, P.M. Feng, H. Lin, K.C. Chou, iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition, Nucleic Acids Res. 41 (e69) (2013), http://dx.doi.org/10.1093/nar/gks1450.

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

[38] K.C. Chou, H.B. Shen, Review: recent advances in developing web-servers for predicting protein attributes, Nat. Sci. 1 (2) (2009) 63–92.

[39] G. von Heijne, The distribution of positively charged residues in bacterial inner membrane proteins correlates with the trans-membrane topology, EMBO J. 5 (11) (1986) 3021–3027.

[40] K. Hiller, A. Grote, M. Scheer, R. Munch, D. Jahn, PrediSi: prediction of signal peptides and their cleavage positions, Nucleic Acids Res. 32 (2004) W375–W379.

[41] M. Wang, J. Yang, K.C. Chou, Using string kernel to predict signal peptide cleavage site based on subsite coupling model, Amino Acids 28 (4) (2005) 395–402.

[42] S.K. Bose, A. Browne, H.B. Kazemian, K. White, Use of artificial neural networks and effects of amino acid encodings in the membrane protein prediction problem, Prog. Pattern Recognition37–46.

[43] S.K. Bose, A. Browne, H.B. Kazemian, K. White, Classifying membrane proteins in the proteome by using artificial neural networks based on the preferential parameters of amino acids, In: J.A. Tenreiro Machado, B. Patkai, I.J. Rudas (Eds.), Intelligent Engineering Systems and Computational Cybernetics, Springer, 2009, pp. 63–71 http://dx.doi.org/10.1007/978-1-4020-8678-6_6.

[44] M. Seifert, Hidden Markov Models with Applications in Computational Biology: Model Extensions and Advanced Analysis of DNA Microarray Data, Südwestdeutscher Verlag für Hochschulschriften, 2013.

[45] S.R. Maetschke, M. Towsey, M.B. Boden, BLOMAP: an encoding of amino acids which improves signal peptide cleavage site prediction, in: Y.P. Phoebe Chen, L. Wong (Eds.), 3rd Asia Pacific Bioinformatics Conference, Singapore, 2005, pp. 141–150.

[46] S.K. Bose, The use of neural networks to identify and analyse membrane proteins in the proteome (Ph.D. thesis), London Metropolitan University, 2006.

[47] C. Pasquier, S.J. Hamodrakas, An hierarchical artificial neural network system for the classification of transmembrane proteins, Protein Eng. 12 (8) (1999) 631–634.

[48] D.J. Livingstone, Artificial Neural Networks: Methods and Applications, Methods in Molecular Biology, 2009 edition, Humana Press, 2011.

[49] Z. Yuan, J.S. Mattick, R.D. Teasdale, SVMtm: support vector machines to predict transmembrane segments, J. Comput. Chem. 25 (5) (2004) 632–636.

[50] H.B. Kazemian, K. White, D. Palmer-Brown, S.A. Yusuf, Applications of evolutionary SVM to prediction of membrane alpha-helices, Expert Systems with Applications 40 (9) (2013) 3412–3420.

[51] Z. Yuan, M.J. Davis, F. Zhang, R.D. Teasdale, Computational differentiation of N-terminal signal peptides and transmembrane helices, Biochem. Biophys. Res. Commun. 312 (4) (2003) 1278–1283.

[52] L. Wang, Support Vector Machines: Theory and Applications [electronic resource], SpringerLink Berlin, Springer, New York, 2005.

[53] A. Statnikov, C.F. Aliferis, D.P. Hardin, A Gentle Introduction to Support Vector Machines in Biomedicine, vol. 1: Theory and Methods, World Scientific Publishing, 2011.

[54] I. Steinwart, A. Christmann, Support Vector Machines [electronic resource], Springer, Dordrecht, 2008.

[55] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (3) (1995) 273–297.

[56] J. Heaton, Introduction to the Math of Neural Networks, Kindle edition, Heaton Research, Inc., 2012.

[57] P. Refaeilzadeh, L. Tang, H. Liu, Cross-Validation [electronic resource], Arizona State University, 2008.