SCIENTIFIC REPORTS

OPEN

A Novel Modeling in Mathematical Biology for Classification of Signal Peptides

Received: 21 July 2017
Accepted: 2 January 2018
Published online: 18 January 2018

Asma Ehsan[1], Khalid Mahmood[1], Yaser Daanial Khan[2], Sher Afzal Khan[3,5] & Kuo-Chen Chou[4]

The molecular structure of macromolecules in living cells is ambiguous unless we classify them in a scientific manner. Signal peptides are of vital importance in determining the behavior of newly formed proteins towards their destined path in cellular and extracellular location in both eukaryotes and prokaryotes. In the present research work, a novel method is offered to foreknow the behavior of signal peptides and determine their cleavage site. The proposed model employs neural networks using isolated sets of prokaryote and eukaryote primary sequences. Protein sequences are classified as secretory or non-secretory in order to investigate secretory proteins and their signal peptides. In comparison with the previous prediction tools, the proposed algorithm is more rigorous, well-organized, significantly appropriate and highly accurate for the examination of signal peptides even in extensive collection of protein sequences.

The classification of proteins bears vast interest to researchers across the world. For better understanding of the molecular structure of living cells, it is important to classify and categorize their macromolecules like proteins[1] in terms of their attributes. To examine the behavior of newly synthesized proteins towards cellular and extracellular positions in cells (for both eukaryotes and prokaryotes), signal peptide works like a "ZIP code"[2,3]. The study of signal can help to propose new medications for genetic remedial treatment which has become difficult for pharmaceutical chemists to develop more accurately[4,5]. A newly translated secretory protein starts to move from rough endoplasmic reticulum to Golgi transport vesicles and then Golgi cisternae leading to secretory transport vesicles. Consequently, these are secreted to the cell exterior surface. The secretory proteins enter the exterior surface of a cell via protein conducting channels (see Fig. 1). The signal peptides translocate the newly created proteins in the secretory pathway[6,7].

It is worth mentioning that the signal peptide of newly synthesized protein can diverge under abnormal circumstances from the exact path. This deviation results the protein ending up in an inappropriate cellular location and hence causes severe diseases. Hereafter, an accurate model for prediction of signal peptide is of much more crucial importance. Predictive results are momentous to examine cell functions and analyze its prospective and genetic purposes[8]. For the prediction of arbitrary signal sequences, a capable novel modeling with a consolidated approach is to be devised to provide more assiduous results. Everyday, a huge number of protein sequences are collected and entering into databanks. Indeed, it is extremely desirable to develop a robust, reliable and excellently accurate computational method for the prediction of signal sequences. A number of researchers presented several models in this respect[9–16]. A huge amount of literature is found pertaining to such prediction models[3,4,17–20]. PrediSi[21] associated the prediction techniques for secretory and non-secretory protein and proposed position weight matrix (PWM) approach for identification of cleavage site. SignalP[22] developed a method that uses a hidden Markov model (HMM) while Signal-CF[8] suggested a coupling fusion predictor that contrived through comprising the subsite coupling effects on protein sequences. Gunnar Von Heijne developed a prototype model depicting the nearby amino acids signal sequences[23]. Lal *et al.* devised a scheme for identifying proteins containing signal peptides and assigned a label SP to a protein after knowing that a protein contains a signal[24]. Extreme learning machine (ELM) and improved ELM were also devised to categorize the avalanche of protein

[1]University of the Punjab, Department of Mathematics, Lahore, 54500, Pakistan. [2]University of Management and Technology, School of Systems and Technology, Department of Computer Science, Lahore, 54770, Pakistan. [3]King Abdul Aziz University, Faculty of Computing and Information Technology in Rabigh, Jeddah, 21577, Saudi Arabia. [4]Gordon Life Science Institute, San Diego, CA, 92130, USA. [5]Abdul Wali Khan University, Department of Computer Sciences, Mardan, Pakistan. Correspondence and requests for materials should be addressed to Y.D.K. (email: yaser.khan@umt.edu.pk)

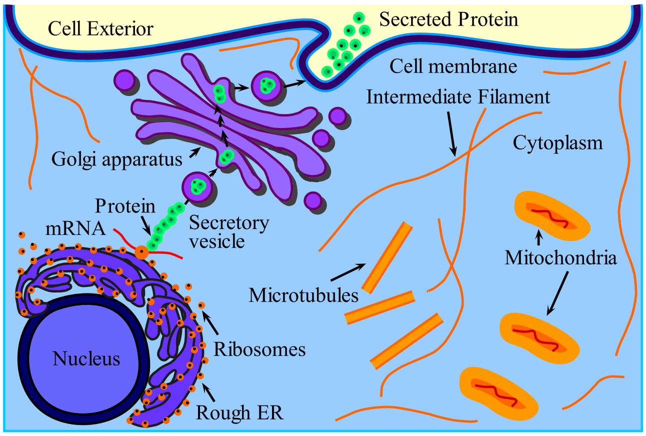


Figure 1. Protein secretion: Ribosomes deposits the protein in endoplasmic reticulum (ER), protein exits ER and enters Golgi apparatus for processing and later it exits Golgi and enters into the cell exterior.

Datasets	Self-consistency Test	Cross validation	Jackknife Test
Eukaryotic(%)	95.00	93.03	92.98
Gram-positive(%)	93.50	87.90	87.85
Gram-negative(%)	97.70	97.09	97.04

Table 1. Prediction accuracies for secretory protein for prokaryotes (Gram-positive & Gram-negative) and eukaryotes.

sequences[25,26]. Some researcher have worked on identifying protein sub-cellular localization using "Support Vector Machine" (SVM) by targeting amino acid composition[27,28]. The SVMs and other machine learning classifiers (ANN, random forest) have been widely used in the field of bioinformatics, and some predictors have been established based on these classifiers, such as PSFM-DBT[29], 2L-piRNA[30], Pse-Analysis[31], ProtDec-LTR[32], ProtDec-LTR2.0[33], etc. Some powerful protein analysis methods have been proposed for the formulation of biological sequences, such as Pse-in-One[31], repDNA[34], based on different functions to produce feature vector for biological sequences. HMMTOP was developed for the prediction of localization of helical transmembrane[35]. Kuo-Chen Chou devised quasi-sequence-order effect for the prediction of sub-cellular localization[36] and afterwards classified protein sequences employing Pseudo Amino Acid Composition[37]. Multiple classifications of protein are offered in[38]. The perception of difference between signal sequence and its peptide chain is manipulated by predictors SS- and SP-indexes[39]. SPOCTOPUS: a combined predictor of signal peptides and membrane protein topology based on neural networks, hidden Markov model (HMM) and dynamic programming algorithm is proposed in[40]. Various researcher have used different mathematical and statistical models for solving various classification problems[41–43]. Yaser *et al.*[44,45] proposed another methodological rigorous examination for the prediction of membrane protein.

Results

In order to validate the current model, a comparative analysis with existing models[8,14,21,22] is established. It is found that the techniques suggested in PrediSi[21], Neural Network (NN)[14] and hidden Markov model (HMM)[22] were less efficient than the Signal-CF[8]. Although the PredSi denied schemes such as NN and HMM[14,22]. Moreover, Signal-CF is a predictor established using pseudo amino acid composition, scaled window and subsite coupling effect, and is an improvement over the PredSi and SignalP methods. Also, few disadvantages of the previous predictors were observed. For example, the feature vector in Chou's scheme was dependent on a the value of a variable λ.

As for the current framework, the occurrence of each residue with a special weight factor has been incorporated for all protein sequences. This yielded a comprehensive description for any arbitrary sequence. A better accuracy rate has been achieved as compared to previous models. Self-consistency test, 10-fold cross validation and jackknife testing were conducted to ensure the accuracy of the proposed model. These tests are elaborated in Tables 1 and 2. Table 1 provides accuracy outcomes for the prediction of secretory protein of eukaryotic, Gram-positive and Gram-negative datasets whereas Table 2 represents the same statistical analysis for non-secretory proteins.

In previous works, most of the methodologies were evaluated by self-consistency, cross validation or/and jackknife testing. The proposed model is validated by all the three testing procedures simultaneously. Firstly, self-consistency test is performed. This is applied to the trained neural network predictor. Datasets for eukaryotic, Gram-positive and Gram-negative are separated into subsets of positives and negatives with comparable random

Datasets	Self-consistency Test	Cross validation	Jackknife Test
Eukaryotic(%)	94.60	92.35	92.30
Gram-positive(%)	91.50	86.55	86.50
Gram-negative(%)	95.80	94.37	94.32

Table 2. Prediction accuracies of non-secretory protein for prokaryotes (Gram-positive & Gram-negative) and eukaryotes.

Protein's Sort	dataset	PridiSi	NN(SignalP)	HMM(SignalP)	Signal-CF	Proposed Model
Secretory	Eukaryotic	72.66	82.11	78.73	76.50	**95.00**
	Gram-positive	78.39	77.97	75.42	80.50	**93.50**
	Gram-negative	86.54	86.54	87.07	87.80	**97.70**
Non-secretory	Eukaryotic	98.31	99.21	97.74	99.78	**94.60**
	Gram-positive	97.89	93.25	99.16	99.79	**91.50**
	Gram-negative	95.70	96.24	99.10	99.82	**95.80**

Table 3. A comparison of proposed model for **Self-consistency** test in order to predict secretory and non-secretory protein sequences.

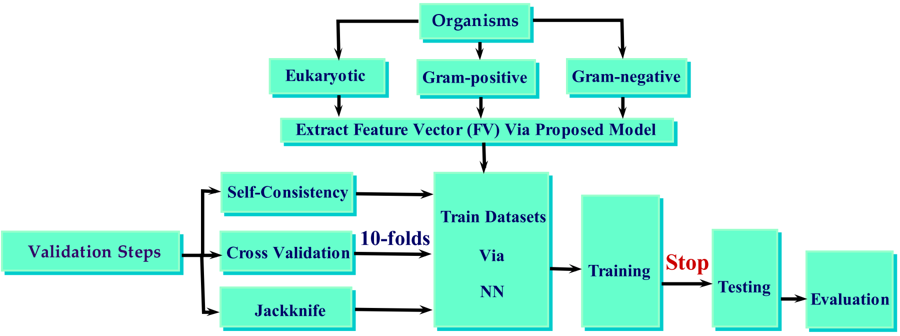


Figure 2. Flowchart illustrating the procedure of training, testing and evaluation.

sizes. After the completion of the training process, the test is executed. In further steps, the other two validation techniques, namely cross validation and jackknife are also evaluated.

The validation process is demonstrated in the following flow chart (see Fig. 2). Secretory and non-secretory datasets are divided into ten equal units for 10-fold cross validation. Evaluation is performed based on performance in ten dissimilar and disjoint subsets of test data and is translated into ten distinctive confusion matrices together with the ROC (Receiver operating characteristic) graphs. The overall prediction accuracy is achieved by averaging all the outcomes. The prediction accuracies are estimated for each subset separately. This procedure is necessary for cross validation and jackknife testing as well. The ROC is a graphical interpretation for the performance of a classifier which illustrates the false positive rate (FPR) against the true positive rate (TPR). The area under the curve signifies the accuracy of the system. ROC graph yielded as a result of self consistency test for the three organisms dataset in given below (see Fig. 3).

A comparative analysis of the proposed model with existing techniques is elaborated in Table 3. It is found that the previous accuracy rates varied from 72% to 87% whereas the accuracy of the proposed model for Self-consistency test lies between 95% to 97%.

In order to assess the accuracy of the prediction model four distinct metrics were utilized. The expected success rate and performance of the proposed model is predicted by these statistical measures. To understand these statistical metrics in easiest way, the prediction scales are expanded as discussed in[46,47]. The true prediction rates for secretory Υ^{+} and non-secretory Υ^{-} for three describes organisms categories are given by

$$\Upsilon^{+} = \frac{\mathbb{T}^{+} - \mathbb{T}^{+}_{-}}{\mathbb{T}^{+}} \tag{1}$$

$$\Upsilon^{-} = \frac{\mathbb{T}^{-} - \mathbb{T}^{-}_{+}}{\mathbb{T}^{-}} \tag{2}$$

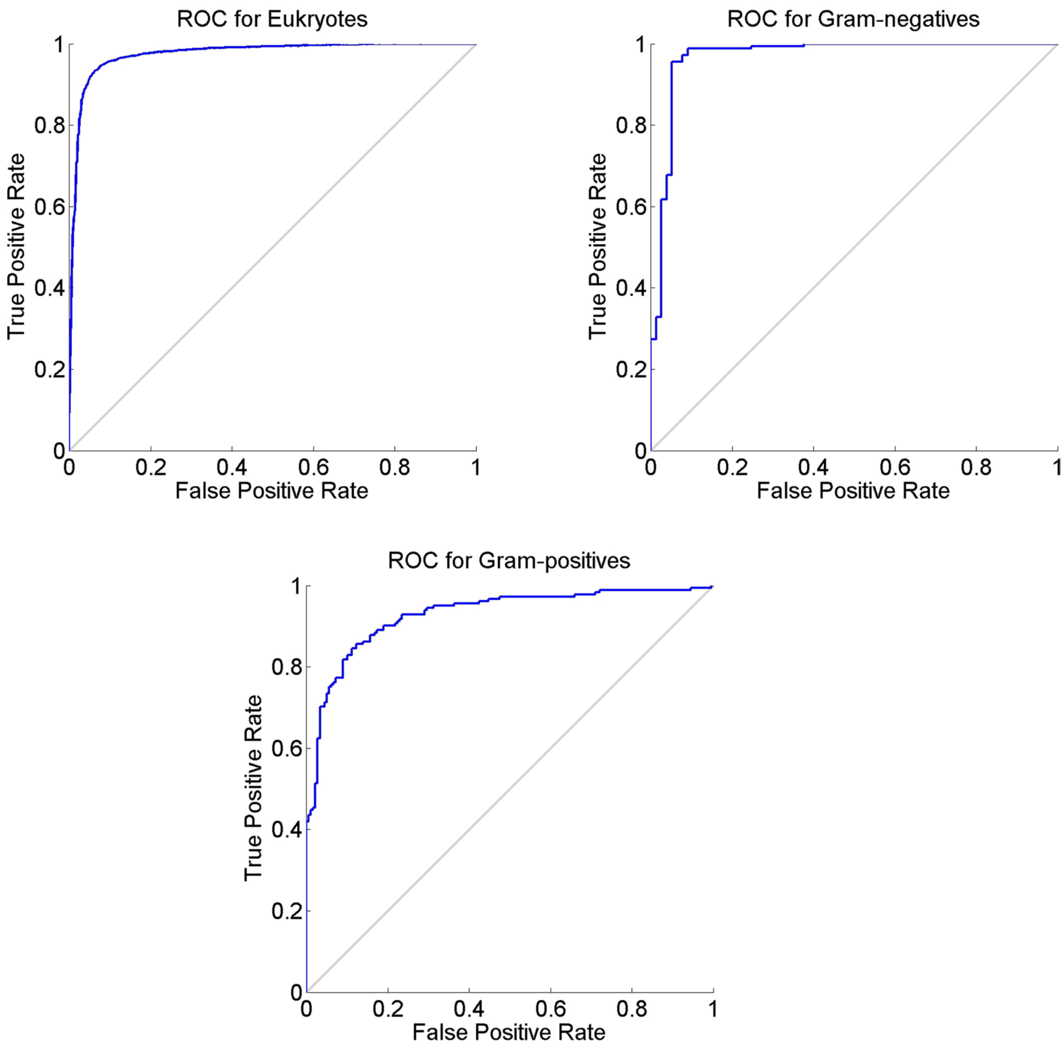


Figure 3. ROC of Self-consistency test for Eukaryotes, Gram Negatives and Gram Positives.

where $\mathbb{T}^+$ and $\mathbb{T}^+_-$ represents the total number of the predicted secretory proteins and incorrectly identified non-secretory protein sequences. Similarly $\mathbb{T}^-$ and $\mathbb{T}^-_+$ denotes the total number of anticipated non-secretory peptides and falsely recognized secretory protein chains. In general the prediction rate is defined by

$$\Upsilon = \frac{\Upsilon^+\mathbb{T}^+ + \Upsilon^-\mathbb{T}^-}{\mathbb{T}^+ + \mathbb{T}^-} = 1 - \frac{\mathbb{T}^+_- + \mathbb{T}^-_+}{\mathbb{T}^+ + \mathbb{T}^-} \tag{3}$$

From the equations (1) to (3), it is clear that in case there is no incorrectly predicted secretory and non-secretory protein sequences i.e., $\mathbb{T}^+_- = \mathbb{T}^-_+ = 0$ then, $\Upsilon^+ = \Upsilon^- = 1$ and the complete prediction accuracy rate is $\Upsilon = 1$. Conversely, when $\mathbb{T}^+_- = \mathbb{T}^-_+ \neq 0$ then the prediction rate would be lesser than 1.

Furthermore, it is helpful to emphasize the importance of equation (4) which is frequently found in literatures for observing the performance superiority of a predictor. Particularly, its advantages have been analyzed and endorsed by a series of significant studies published very recently[48–53].

$$\begin{cases} Sn = \dfrac{TP}{TP + FN} \\ Sp = \dfrac{TN}{TN + FP} \\ Acc = \dfrac{TP + TN}{TP + TN + FP + FN} \\ MCC = \dfrac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)\,(FN + TN)\,(FP + TN)\,(TP + FN)}} \end{cases} \tag{4}$$

Datasets	TP	TN	FP	FN	Acc	Mcc	Sn	Sp
Eukaryotic	5648	1297	392	92	0.94	0.81	0.98	0.77
Gram-positive	159	148	22	31	0.85	0.71	0.84	0.87
Gram-negative	177	71	6	6	0.95	0.90	0.97	0.92

Table 4. Depicts the validation metrics for the proposed model.

where TP, TN, FP and FN represents the true positive, true negative, false positive and false negative values respectively while Sn, Sp, Acc and MCC denotes the values for sensitivity, specificity, accuracy and Mathew's correlation coefficient.

The equation (4) can be rewritten in the form of the symbols given in (3)

$$\begin{cases} TP &= \mathbb{T}^{+} - \mathbb{T}_{-}^{+} \\ TN &= \mathbb{T}^{-} - \mathbb{T}_{+}^{-} \\ FP &= \mathbb{T}_{+}^{-} \\ FN &= \mathbb{T}_{-}^{+} \end{cases} \tag{5}$$

By substituting (5) into (4) along and utilizing (3), we have

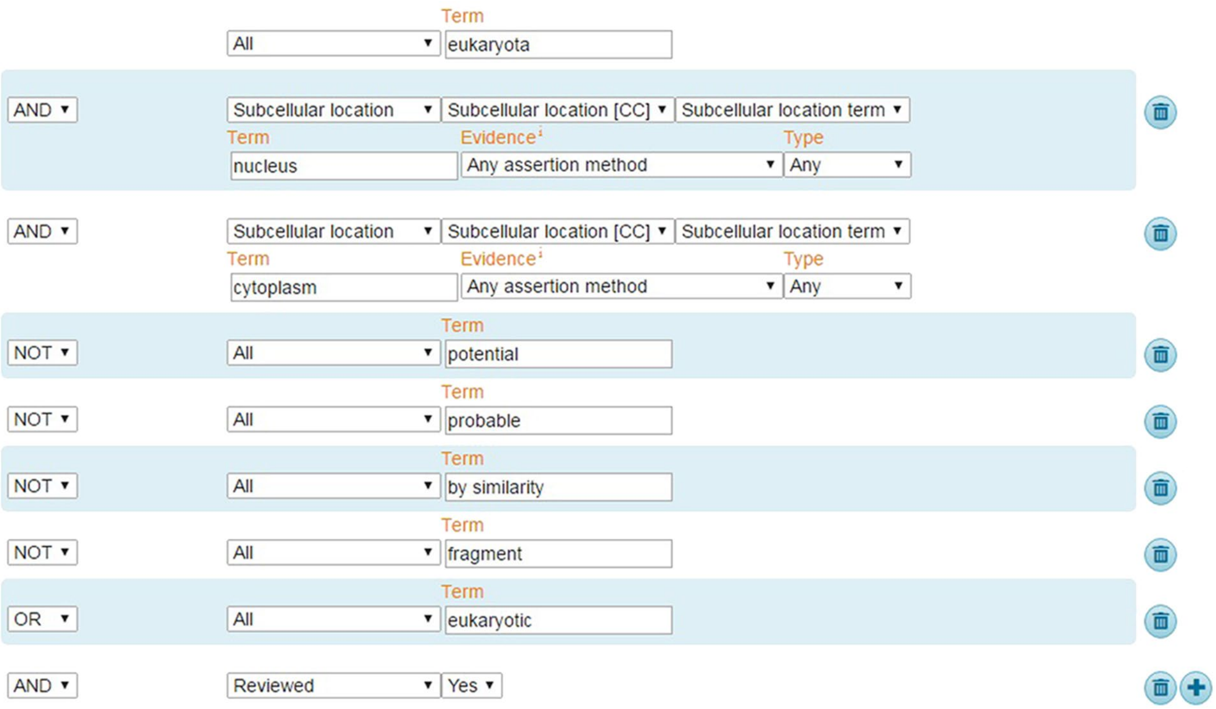
$$\begin{cases} Sn &= 1 - \dfrac{\mathbb{T}_{-}^{+}}{\mathbb{T}^{+}} \\ Sp &= 1 - \dfrac{\mathbb{T}_{+}^{-}}{\mathbb{T}^{-}} \\ Acc &= \Upsilon = 1 - \dfrac{\mathbb{T}_{-}^{+} + \mathbb{T}_{+}^{-}}{\mathbb{T}^{+} + \mathbb{T}^{-}} \\ MCC &= \dfrac{1 - \left(\frac{\mathbb{T}_{-}^{+}}{\mathbb{T}^{+}} + \frac{\mathbb{T}_{+}^{-}}{\mathbb{T}^{-}}\right)}{\sqrt{\left(1 + \frac{\mathbb{T}_{+}^{-} - \mathbb{T}_{-}^{+}}{\mathbb{T}^{+}}\right)\left(1 + \frac{\mathbb{T}_{-}^{+} - \mathbb{T}_{+}^{-}}{\mathbb{T}^{-}}\right)}} \end{cases} \tag{6}$$

It is worth mentioning that from (6) that if $\mathbb{T}_{-}^{+} = 0$ then $Sn = 1$ signifying that no secretory protein sequences is incorrectly predicted as non-secretory protein sequence. In other case when $\mathbb{T}_{-}^{+} = \mathbb{T}^{+}$ implies that $Sn = 0$ represents that all the secretory protein sequences were incorrectly predicted as non-secretory protein chains. Similarly if $\mathbb{T}_{+}^{-} = 0$ yields $Sp = 1$ then it signifies that no non-secretory sequence was incorrectly predicted similarly if $\mathbb{T}_{+}^{-} = \mathbb{T}^{-}$ gives $Sp = 0$ then it shows that all the non-secretory sequences were falsely predicted as secretory sequences. On the other hand the prediction accuracy metric $Acc = \Upsilon = 1$ when there is no incorrectly predicted sequences for secretory Υ^{+} as well as for non-secretory proteins Υ^{-} i.e., $\mathbb{T}_{-}^{+} = \mathbb{T}_{+}^{-} = 0$. A value of $\mathbb{T}_{-}^{+} = \mathbb{T}^{+}$ and $\mathbb{T}_{+}^{-} = \mathbb{T}^{-}$ indicates that all the secretory Υ^{+} and non-secretory Υ^{-} sequences were falsely predicted hence yielding an overall accuracy $Acc = \Upsilon = 0$. Additionally, Matthew correlation coefficient (MCC) is frequently used to assess the performance of binary classifications. MCC is designed in such a way that the disparity in the size of positive or negative samples in the comprehensive dataset does not bias the overall outcome.

The statistical analysis using independent set testing for computing these metrics mentioned above is given in Table 4 with a noteworthy MCC value for the three segregated Eukrayotic, Gram-positive and Gram-negative datasets is 0.81, 0.71 and 0.90 respectively.

Discussion

Undoubtedly, the examination and analysis of patterns and sequence is quite convoluted when there are avalanche of protein sequences with diverse lengths. The statistical formulation of these sequences along with construction of robust data set to produce assiduous results was a challenging task. These cumbersome tasks have been rectified by the proposed technique. A performance comparison over the existing and previous predictors[8,14,21,22] has been analyzed and a worth seeing prediction accuracy using the proposed technique was observed. The idea was to recognize the query proteins as secretory or non-secretory and hence identify the presence of signal peptides, further in the next step the cleavage site for the signal peptide is identified based on "(−3, −1) -Rule", as described by Von Heijni[10]. It is obseved that signal peptides are more important than protein synthesization. The insufficiency of protein in distribution or contribution is cause of health hazards. So, the protein synthesization is not enough to perform cell functions properly but the compartmentalization of proteins to their relevent loci is of vital importance. Involvement of deviated protein with beneficial cells function is responsible for severe health diseases including Neurodegenerative disorder and cells death[54,55]. Hence, the computational capability to classify a protein as secretory or a non-secretory protein with a high level of accuracy bears great significance. In this process of anticipation, a large-scaled benchmark data set had been selected and incorporated into the prediction model. Even for huge data set, signal sequences along with their cleavage site was predicted at the cost of minor computational overhead with a high accuracy. The proposed algorithm can also be applied in future works to solve several other problems such as identification of Post Translational Modification (PTM) sites[56–58],



Term
All ▾ eukaryota

AND ▾ Subcellular location ▾ Subcellular location [CC] ▾ Subcellular location term ▾
Term: nucleus Evidence: Any assertion method ▾ Type: Any ▾

AND ▾ Subcellular location ▾ Subcellular location [CC] ▾ Subcellular location term ▾
Term: cytoplasm Evidence: Any assertion method ▾ Type: Any ▾

NOT ▾ All ▾ Term: potential

NOT ▾ All ▾ Term: probable

NOT ▾ All ▾ Term: by similarity

NOT ▾ All ▾ Term: fragment

OR ▾ All ▾ Term: eukaryotic

AND ▾ Reviewed ▾ Yes ▾

Figure 4. Shows a sample query to extract dataset.

DNA-binding protein prediction[59], protein-protein interaction prediction[60], etc. The prediction accuracy of the proposed algorithm using self-consistency, cross validation and jackknife tests were reported as **95.00%**, **93.50%** and **97.70%** (Self-consistency), **93.03%**, **87.90%** and **97.90%** (Cross validation) and **92.98%**, **87.85%** and **97.04%** (Jackknife) for the three organisms namely eukaryotic, Gram-positive and Gram-negative respectively.

Methods

Material. Since user-friendly and publicly accessible web-servers represent the future direction for developing practically more useful models[61–63], we shall make efforts in our future work to provide a web-server for the method presented in this paper. A current updated edition of the well-known database uniprot wasn employed to obtain an inclusive benchmark dataset. The proposed model was trained over the assimilated dataset and subsequently validated in the following steps:

- The eukaryotic or eukaryota, Gram-positive and Gram-negative are the organisms which have been incorporated for the desired prediction via proposed model. A benchmark data set related to these organisms was composed. A query was operated for the extraction of non-secretory protein dataset in the entire database. Through this inquiry, those protein sequences were selected which were marked in OC (Organism Classification) field as eukaryotic or eukaryota, Gram-positive or Gram-negative. Amongst these extracted entities, the eukaryotic entries annotated with subcellular locations as the nucleus or cytoplasm were selected while Gram-positive and Gram-negative entries annotated with subcellular location as cytoplasm were selected as a non-secretory protein sequences shown below (see Fig. 4). The vague extraction glossed with keywords like "by similarity" and "fragment" were discarded.
- In taxonomy field, the keyword eukaryota was combed to obtain a training dataset for secretory protein sequences. The proteins annotated as signal peptides were only selected excluding those containing ambiguous keywords like potential, probable, fragment and by similarity in FT field. Out of those selected protein sequences, in which the preamble signal peptide was surely contained, the first 100 residues were extracted. The Gram-positive and Gram-negative organisms data was also extracted in a similar fashion.
- The defined steps were adopted strictly to collect a first-rate benchmark data set of signal peptides for both secretory and non-secretory proteins in the three organisms namely eukaryotic, Gram-positive and Gram-negative.
- Table 5, shows that 28,220 eukaryotic, 934 Gram-positive and 1046 Gram-negative samples were found in all 30,200 secretory proteins. After deleting similar entities, these are reduced to 25327 unique values. Similarly for all non-secretory proteins, 5595, 908 and 441 were found in eukaryotic, Gram-positive and Gram-negative respectively and 6,944 conjointly. After deleting repeated items, these are reduced to 6426 unique values. The benchmark dataset extracted for the proposed model is more diverse and comprehensive than the benchmark data set employed by existing predictors. Within the training dataset the secretory samples are considered as positive and non-secretory samples as negative as shown in flowchart given below (see Fig. 5).

Organisms	Eukaryotic (K)	Gram-positive (G_+)	Gram-negative (G_-)
Secretory	28,220	934	1,046
Non-secretory	5,595	908	441
Total	33,815	1,842	1,487

Table 5. The data set used to test the proposed model includes secretory and non-secretory protein sequences in all three organisms.

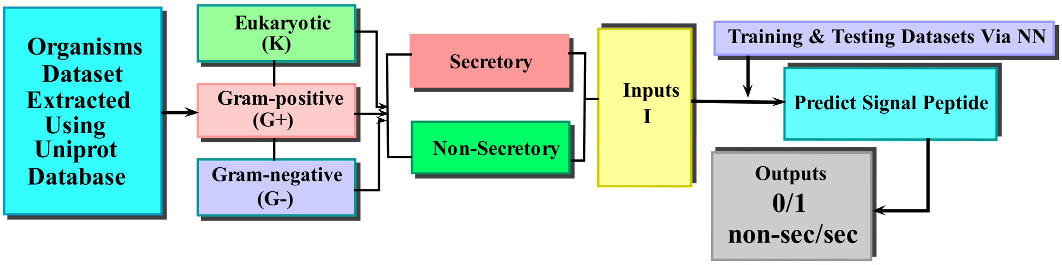


Figure 5. A flowchart depicting the process starting from data extraction and ultimately leading to validation of results.

In the above flow chart for the sake of convenience inputs expressed as positive (negative) in the direction of secretory (non-secretory) for eukaryotic (**K**), Gram-positive(G_+) and Gram-negative(G_-) are documented as follows.

$$
\begin{aligned}
I_K &= K^+ \oplus K^- \\
I_{G_+} &= G_+^+ \oplus G_+^- \\
I_{G_-} &= G_-^+ \oplus G_-^-
\end{aligned}
$$

where I_K, I_{G_+} and I_{G_-} are the inputs from eukaryotic, Gram-positive and Gram-negative organisms for the collection of secretory (K^+, G_+^+, G_-^+) and non-secretory (K^-, G_+^-, G_-^-) protein sequences. And the symbol $\oplus$ denotes the combination of both secretory and non-secretory protein sequences. The primary structure of secretory and non-secretory proteins can be found in Supplementary Tables S1 and S2 respectively.

Method. Consider a protein sequence **S** comprising of N amino acid residues.

$$\mathbf{S} = \xi_1\xi_2\xi_3\xi_4\xi_5\xi_6\xi_7\cdots\xi_N \tag{7}$$

where ξ_1, is the first amino acid residue, ξ_2 is the second amino acid residue and so on up till the last residue ξ_N within the polypeptide chain **S**. Also, N represents the length of protein sequence (7). In order to predict the signal peptide in a convenient manner a computational algorithm has been proposed. This algorithm preserves the sequence order effect and is carried out by assimilating the whole sequence data together with the occurrence of each amino acid residue $\Lambda_{\hat{a}}$ of type $\hat{a}: 1 \le \hat{a} \le 20$ (any one of residue among twenty amino acid residues). The whole structure is based on expressions (8) to (11). Expression (8) is associated with the number of occurrence $\Lambda_{\hat{a}}$ of residue $\hat{a}$ and with the possible number of correlated factors ϑ of $\hat{a}$ with itself such that $(\Lambda_{\hat{a}} - 1)!\vartheta(\xi_{\hat{a}}, \xi_{\hat{a}})$. Subsequently, expression (9) represents the mean factors M_0, M_i and M_j which are linked with the difference factors of $\hat{a}$ at their corresponding positions and is appended with constraint (10). Moreover, M_i varies according to the number of difference factors and each difference factor associates an exclusive mean term. The difference is labeled as $(q-p)_i$, where p and q are respective positions of $\hat{a}$ in the polypeptide chain given as $\xi_p = \xi_q = \xi_{\hat{a}}$. Also subscript i represents the number of the possible difference factors between similar residues excluding first and last time occurrence of residue, $\hat{a}$ depends upon the n number of $\hat{a}$ residues in sequence. Likewise M_j is fitted to factor $(N-r)$, where r is the position of the n^{th} occurrence of the residue $\hat{a}$ in sequence (7) such that $\xi_r = \xi_{\hat{a}}$ and $1 \le p < q < r \le N$ represents the $\hat{a}$ residue at their corresponding positions.

$$\Lambda_{\hat{a}} + (\Lambda_{\hat{a}} - 1)!\,\vartheta(\xi_{\hat{a}}, \xi_{\hat{a}}) \tag{8}$$

$$
\begin{gathered}
\left[(p-0)_0 M_0 + \sum_{q>p}(q-p)_i M_i + (N-r)_n M_{j=n}\right]; \\
1 \le p < q < r \le N,\ i = 1, 2, 3, \ldots n-1
\end{gathered} \tag{9}
$$

$$(9) \Rightarrow \begin{cases} \sum\limits_{q>p}(q-p)_i M_i + (N-r)_n M_{j=n}, & \mathit{if} \quad 1 \le p < q,\ r < N \\ (p-0)_0 M_0 + \sum\limits_{q>p}(q-p)_i M_i + (N-r)_n M_{j=n}, & \mathit{if} \quad 1 < p < q,\ r < N \\ (p-0)_0 M_0 + \sum\limits_{q>p}(q-p)_i M_i, & \mathit{if} \quad 1 < p < q,\ r = N \end{cases} \tag{10}$$

Combining expression (8) and (9) and using constraint (10) yields the template for manipulating feature component related to $\hat{a}$, given in (11) and (12).

$$\begin{aligned} &\Lambda_{\hat{a}} + (\Lambda_{\hat{a}} - 1)!\,\vartheta(\xi_{\hat{a}}, \xi_{\hat{a}}) + \left[(p-0)_0 M_0 + \sum_{q>p}(q-p)_i M_i + (N-r)_n M_{j=n}\right], \\ &i = 1, 2, 3, \ldots, n-1 \end{aligned} \tag{11}$$

Or

$$\begin{aligned} \Lambda_{\hat{a}} + (\Lambda_{\hat{a}} - 1)!\,\vartheta(\xi_{\hat{a}}, \xi_{\hat{a}}) + [&(p-0)_0 M_0 + (q-p)_1 M_1 + (q-p)_2 M_2 \\ &+ (q-p)_3 M_3 + \cdots + (q-p)_{n-1} M_{n-1} + (N-r)_n M_{j=n}] \end{aligned} \tag{12}$$

whereas M_i and M_j are count factors of residue $\hat{a}$ with other nineteen residues occurring before and after $\hat{a}$ respectively and can be defined in terms of X and Y, as elaborated in equations (13) and (14).

$$\begin{aligned} M_i &= \{X_i + Y_i\}, \quad i = 1, 2, 3, \ldots, n-1 \\ M_{j=n} &= \{X + Y\} \end{aligned} \tag{13}$$

where

$$\begin{aligned} X &= X_i = \frac{1}{38}\left[\sum_{\substack{k=1 \\ k \neq \hat{a}}}^{20} f_k\, \vartheta(\xi_k, \xi_{\hat{a}}) + \sum_{\substack{k=1 \\ k \neq \hat{a}}}^{20} f_0\, \vartheta(\xi_{\hat{a}}, \xi_k)\right] \\ Y &= Y_i = \frac{1}{38}\left[\sum_{\substack{k=1 \\ k \neq \hat{a}}}^{20} f_0\, \vartheta(\xi_k, \xi_{\hat{a}}) + \sum_{\substack{k=1 \\ k \neq \hat{a}}}^{20} f_k\, \vartheta(\xi_{\hat{a}}, \xi_k)\right] \end{aligned} \tag{14}$$

where f_k, $1 < k < 20$ represents the frequency of pair function ϑ related to residue $\hat{a}$ with other nineteen amino acid residues and in case of non-occurrence, it is denoted by f_0. Consider $\rho_{l,m}$, it describes pair function ϑ for all amino acid residues with each other and while another pair function $\vartheta(\xi_l,\xi_m)$ in term of $\rho_{l,m}$ is defined as $\vartheta(\xi_l,\xi_m) = \rho_{l,m}$; $l = m = 1, 2, 3 \ldots 20$, shown in matrix (15). Subsequently, expression (16) is the complete representation for all possible pair factors regarding X and Y followed by (17), When pair $\vartheta(\xi_l,\xi_m)$ exists then $\rho_{l,m}$ is established as 1 otherwise its assigned a zero entry. Furthermore, (16) admits to (15) with entries $\rho_{l,m}$, $\rho_{l,l}$ and $\rho_{m,l}$ indicating lower triangular matrix for X. Consequently, diagonal entries represents the combination among similar residues and upper triangular matrix for Y.

$$\begin{pmatrix} \rho_{1,1} & \rho_{1,2} & \rho_{1,3} & \cdots & \rho_{1,20} \\ \rho_{2,1} & \rho_{2,2} & \rho_{2,3} & \cdots & \rho_{2,20} \\ \rho_{3,1} & \rho_{3,2} & \rho_{3,3} & \cdots & \rho_{3,20} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{20,1} & \rho_{20,2} & \rho_{20,3} & \cdots & \rho_{20,20} \end{pmatrix} \tag{15}$$

$$\begin{aligned} \begin{pmatrix} 0 & 1 & 1 & \ldots & 1 \\ 1 & 0 & 1 & \ldots & 1 \\ 1 & 1 & 0 & \ldots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & 1 & \ldots & 0 \end{pmatrix} &= M_i = M_j = \{X_i + Y_i\} = \{X + Y\} \\ &= \begin{pmatrix} 0 & 0 & 0 & \ldots & 0 \\ 1 & 0 & 0 & \ldots & 0 \\ 1 & 1 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & 1 & \ldots & 0 \end{pmatrix} + \begin{pmatrix} 0 & 1 & 1 & \ldots & 1 \\ 0 & 0 & 1 & \ldots & 1 \\ 0 & 0 & 0 & \ldots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & 0 \end{pmatrix} \end{aligned} \tag{16}$$

where

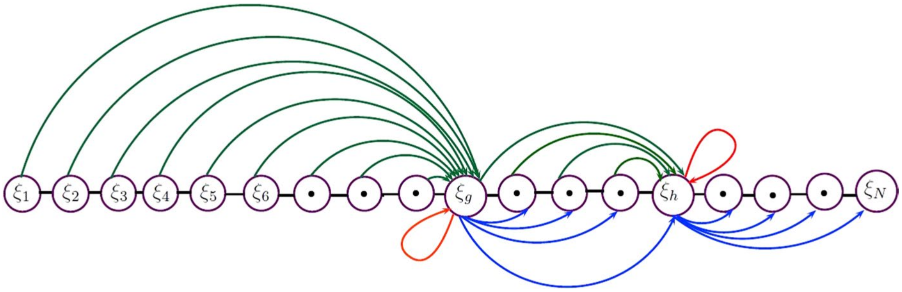


Figure 6. An illustration of the structural scheme of proposed technique.

$$\rho_{l,m} = \begin{cases} 1, & when \quad \vartheta(\xi_l, \xi_m) \quad exists \quad for \quad both \quad l = m \quad or \quad l \neq m \\ 0, & otherwise \end{cases} \tag{17}$$

Extension of (11) gives the idea for manipulating feature components for all twenty amino acid residues in matrix form, as given in (18).

$$\Lambda_{\hat{a}} + (\Lambda_{\hat{a}} - 1)! \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix} + \frac{1}{38}\Bigg[(p - 0)_0 M_0 + \sum_{q>p}(q - p)_i \begin{pmatrix} 0 & 1 & 1 & \dots & 1 \\ 1 & 0 & 1 & \dots & 1 \\ 1 & 1 & 0 & \dots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & 1 & \dots & 0 \end{pmatrix}$$
$$+ (N - r)_n \begin{pmatrix} 0 & 1 & 1 & \dots & 1 \\ 1 & 0 & 1 & \dots & 1 \\ 1 & 1 & 0 & \dots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & 1 & \dots & 0 \end{pmatrix}\Bigg] \tag{18}$$

Using equation (14) in (12) contributes as a component of feature vector corresponding to residue $\hat{a}$ as given in equations (19) and (20).

$$\begin{aligned} \Pi_{\hat{a}} \;=\; & \Lambda_{\hat{a}} + (\Lambda_{\hat{a}} - 1)!\,\vartheta(\xi_{\hat{a}}, \xi_{\hat{a}}) \\ & + \frac{1}{38}\left[(p - 0)_0\left\{\sum_{\substack{k=1\\k\neq\hat{a}}}^{20} f_k\,\vartheta(\xi_k, \xi_{\hat{a}}) + \sum_{\substack{k=1\\k\neq\hat{a}}}^{20} f_k\,\vartheta(\xi_{\hat{a}}, \xi_k)\right\}_0\right. \\ & + \left\{(q - p)_1\left\{\sum_{\substack{k=1\\k\neq\hat{a}}}^{20} f_k\,\vartheta(\xi_k, \xi_{\hat{a}}) + \sum_{\substack{k=1\\k\neq\hat{a}}}^{20} f_k\,\vartheta(\xi_{\hat{a}}, \xi_k)\right\}_1\right. \\ & + (q - p)_2\left\{\sum_{\substack{k=1\\k\neq\hat{a}}}^{20} f_k\,\vartheta(\xi_k, \xi_{\hat{a}}) + \sum_{\substack{k=1\\k\neq\hat{a}}}^{20} f_k\,\vartheta(\xi_{\hat{a}}, \xi_k)\right\}_2 \\ & + (q - p)_3\left\{\sum_{\substack{k=1\\k\neq\hat{a}}}^{20} f_k\,\vartheta(\xi_k, \xi_{\hat{a}}) + \sum_{\substack{k=1\\k\neq\hat{a}}}^{20} f_k\,\vartheta(\xi_{\hat{a}}, \xi_k)\right\}_3 \\ & \left. + \cdots + (q - p)_{n-1}\left\{\sum_{\substack{k=1\\k\neq\hat{a}}}^{20} f_k\,\vartheta(\xi_k, \xi_{\hat{a}}) + \sum_{\substack{k=1\\k\neq\hat{a}}}^{20} f_k\,\vartheta(\xi_{\hat{a}}, \xi_k)\right\}_{n-1}\right\} \\ & \left. + (N - r)_n\left\{\sum_{\substack{k=1\\k\neq\hat{a}}}^{20} f_k\,\vartheta(\xi_k, \xi_{\hat{a}}) + \sum_{\substack{k=1\\k\neq\hat{a}}}^{20} f_k\,\vartheta(\xi_{\hat{a}}, \xi_k)\right\}_n\right] \end{aligned} \tag{19}$$

Or

$$\begin{aligned}\Pi_{\hat{a}} \;=\;& \Lambda_{\hat{a}} + (\Lambda_{\hat{a}} - 1)!\,\vartheta(\xi_{\hat{a}}, \xi_{\hat{a}})\\ &+ \frac{1}{38}\left[(p - 0)_0\left[\left\{\sum_{\substack{k=1\\k\neq\hat{a}}}^{20} f_k\,\vartheta(\xi_k, \xi_{\hat{a}}) + \sum_{\substack{k=1\\k\neq\hat{a}}}^{20} f_k\,\vartheta(\xi_{\hat{a}}, \xi_k)\right\}\right]_0\right.\\ &+ \sum_{q>p}(q - p)_i\left[\left\{\sum_{\substack{k=1\\k\neq\hat{a}}}^{20} f_k\,\vartheta(\xi_k, \xi_{\hat{a}}) + \sum_{\substack{k=1\\k\neq\hat{a}}}^{20} f_k\,\vartheta(\xi_{\hat{a}}, \xi_k)\right\}\right]_i\\ &\left.+ (N - r)_n\left[\left\{\sum_{\substack{k=1\\k\neq\hat{a}}}^{20} f_k\,\vartheta(\xi_k, \xi_{\hat{a}}) + \sum_{\substack{k=1\\k\neq\hat{a}}}^{20} f_k\,\vartheta(\xi_{\hat{a}}, \xi_k)\right\}\right]_n\right],\ i = 1, 2, 3, \ldots, n - 1.\end{aligned} \tag{20}$$

To understand the structural scheme of proposed model consider g_{th} term of sequence given in equation (7), say, ξ_g, which reflects the first alphabetical letter of amino acid residues say 'A'. Notice its occurrences as well as corresponding positions in the sequence. ξ_g makes pair with its contiguous residues before and after the g_{th} residue in the terms $\vartheta(\xi_k,\xi_g)$ and $\vartheta(\xi_g,\xi_k)$ represented by green and blue curved lines and pairs ξ_g by itself represented by red loops (see Fig. 6). This process will be continued until next ξ_h occurs at h_{th} position such that $\xi_g = \xi_h = A$. Similarly same steps will be applied for ξ_j. The feature component corresponding to residue "A" is interpreted in equation (21).

$$\begin{aligned}\Pi_A \;=\;& \Lambda_A + (\Lambda_A - 1)!\,\vartheta(A, A)\\ &+ \frac{1}{38}\left[(p_g - 0)_0\left[\left\{\sum_{\substack{k=1\\k\neq A}}^{20} f_k\,\vartheta(\xi_k, A) + \sum_{\substack{k=1\\k\neq A}}^{20} f_k\,\vartheta(A, \xi_k)\right\}\right]_0\right.\\ &+ (q_h - p_g)\left[\left\{\sum_{\substack{k=1\\k\neq A}}^{20} f_k\,\vartheta(\xi_k, A) + \sum_{\substack{k=1\\k\neq A}}^{20} f_k\,\vartheta(A, \xi_k)\right\}\right]\\ &\left.+ (N - r_h)\left[\left\{\sum_{\substack{k=1\\k\neq A}}^{20} f_k\,\vartheta(\xi_k, A) + \sum_{\substack{k=1\\k\neq A}}^{20} f_k\,\vartheta(A, \xi_k)\right\}\right]\right]\end{aligned} \tag{21}$$

where $k = 1, 2, 3 \ldots, 20$ represents the ordinal values of twenty amino acid residues in alphabetical order and for more simplification assume that $\xi_1, \xi_2, \xi_3, \ldots, \xi_{20}$ represents 20 amino acids in alphabetical order labeled as: A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W and Y also ξ_{21} onwards the 20 residues cyclically repeats themselves then let $\Pi_1, \Pi_2, \Pi_3, \ldots, \Pi_{20}$ be their corresponding feature components. The set of twenty feature components is given in equation (22).

$$\begin{aligned}\Pi_1 \;=\;& \Lambda_1 + (\Lambda_1 - 1)!\,\vartheta(\xi_1, \xi_1)\\ &+ \frac{1}{38}\left[(p - 0)_0\left[\left\{\sum_{\substack{k=1\\k\neq 1}}^{20} f_k\,\vartheta(\xi_k, \xi_1) + \sum_{\substack{k=1\\k\neq 1}}^{20} f_k\,\vartheta(\xi_1, \xi_k)\right\}\right]_0\right.\\ &+ \sum_{q>p}(q - p)_i\left[\left\{\sum_{\substack{k=1\\k\neq 1}}^{20} f_k\,\vartheta(\xi_k, \xi_1) + \sum_{\substack{k=1\\k\neq 1}}^{20} f_k\,\vartheta(\xi_1, \xi_k)\right\}\right]_i\\ &\left.+ (N - r)_n\left[\left\{\sum_{\substack{k=1\\k\neq 1}}^{20} f_k\,\vartheta(\xi_k, \xi_1) + \sum_{\substack{k=1\\k\neq 1}}^{20} f_k\,\vartheta(\xi_1, \xi_k)\right\}\right]_n\right],\ i = 1, 2, 3, \ldots, n - 1.\end{aligned} \tag{22}$$

$$\begin{aligned}\Pi_2 = {} & \Lambda_2 + (\Lambda_2 - 1)!\,\vartheta(\xi_2, \xi_2)\\ & + \frac{1}{38}\Bigg[(p - 0)_0\Bigg[\Bigg\{\sum_{\substack{k=1\\ k\neq 2}}^{20} f_k\,\vartheta(\xi_k, \xi_2) + \sum_{\substack{k=1\\ k\neq 2}}^{20} f_k\,\vartheta(\xi_2, \xi_k)\Bigg\}\Bigg]_0\\ & + \sum_{q>p}(q - p)_i\Bigg[\Bigg\{\sum_{\substack{k=1\\ k\neq 2}}^{20} f_k\,\vartheta(\xi_k, \xi_2) + \sum_{\substack{k=1\\ k\neq 2}}^{20} f_k\,\vartheta(\xi_2, \xi_k)\Bigg\}\Bigg]_i\\ & + (N - r)_n\Bigg[\Bigg\{\sum_{\substack{k=1\\ k\neq 2}}^{20} f_k\,\vartheta(\xi_k, \xi_2) + \sum_{\substack{k=1\\ k\neq 2}}^{20} f_k\,\vartheta(\xi_2, \xi_k)\Bigg\}\Bigg]\Bigg],\quad i = 1, 2, 3, \ldots, n - 1.\end{aligned}$$

$$\begin{aligned}\Pi_3 = {} & \Lambda_3 + (\Lambda_3 - 1)!\,\vartheta(\xi_3, \xi_3)\\ & + \frac{1}{38}\Bigg[(p - 0)_0\Bigg[\Bigg\{\sum_{\substack{k=1\\ k\neq 3}}^{20} f_k\,\vartheta(\xi_k, \xi_3) + \sum_{\substack{k=1\\ k\neq 3}}^{20} f_k\,\vartheta(\xi_3, \xi_k)\Bigg\}\Bigg]_0\\ & + \sum_{q>p}(q - p)_i\Bigg[\Bigg\{\sum_{\substack{k=1\\ k\neq 3}}^{20} f_k\,\vartheta(\xi_k, \xi_3) + \sum_{\substack{k=1\\ k\neq 3}}^{20} f_k\,\vartheta(\xi_3, \xi_k)\Bigg\}\Bigg]_i\\ & + (N - r)_n\Bigg[\Bigg\{\sum_{\substack{k=1\\ k\neq 3}}^{20} f_k\,\vartheta(\xi_k, \xi_3) + \sum_{\substack{k=1\\ k\neq 3}}^{20} f_k\,\vartheta(\xi_3, \xi_k)\Bigg\}\Bigg]_n\Bigg],\quad i = 1, 2, 3, \ldots, n - 1.\\ & \vdots\end{aligned}$$

$$\begin{aligned}\Pi_{20} = {} & \Lambda_{20} + (\Lambda_{20} - 1)!\,\vartheta(\xi_{20}, \xi_{20})\\ & + \frac{1}{38}\Bigg[(p - 0)_0\Bigg[\Bigg\{\sum_{\substack{k=1\\ k\neq 20}}^{20} f_k\,\vartheta(\xi_k, \xi_{20}) + \sum_{\substack{k=1\\ k\neq 20}}^{20} f_k\,\vartheta(\xi_{20}, \xi_k)\Bigg\}\Bigg]_0\\ & + \sum_{q>p}(q - p)_i\Bigg[\Bigg\{\sum_{\substack{k=1\\ k\neq 20}}^{20} f_k\,\vartheta(\xi_k, \xi_{20}) + \sum_{\substack{k=1\\ k\neq 20}}^{20} f_k\,\vartheta(\xi_{20}, \xi_k)\Bigg\}\Bigg]_i\\ & + (N - r)_n\Bigg[\Bigg\{\sum_{\substack{k=1\\ k\neq 20}}^{20} f_k\,\vartheta(\xi_k, \xi_{20}) + \sum_{\substack{k=1\\ k\neq 20}}^{20} f_k\,\vartheta(\xi_{20}, \xi_k)\Bigg\}\Bigg]_n\Bigg],\quad i = 1, 2, 3, \ldots, n - 1.\end{aligned}$$

The set of above twenty feature components are based on three properties of amino acids, namely hydrophobicity, hydrophilicity and side chain mass of amino acids. Each property associates a set of sixty components so collectively this admits 180 components manipulated by using equations (23) to (25), where *s* represents the number of attributes for amino acid residues in succinct representation, for $s = 1, 2, 3$ it corresponds to hydrophobicity, hydrophilicity and side chain mass of amino acids respectively.

$$\begin{aligned}\vartheta(\xi_l, \xi_m) = {} & \sqrt{\Delta_s{}^*(\xi_l)^2\,|\Delta_s{}^*(\xi_l) - \Delta_s{}^*(\xi_m)|^2}\\ & + \frac{|(\Delta_1^*(\xi_l) - \overline{\Delta}_1^*(\hat{a}))\,(\Delta_2^*(\xi_m) - \overline{\Delta}_2^*(\hat{a})|}{\sqrt{\sum_{l=1}^{20}(\Delta_1^*(\xi_l) - \overline{\Delta}_1^*(\hat{a}))^2\sum_{m=1}^{20}(\Delta_2^*(\xi_m) - \overline{\Delta}_2^*(\hat{a}))^2}}\end{aligned} \tag{23}$$

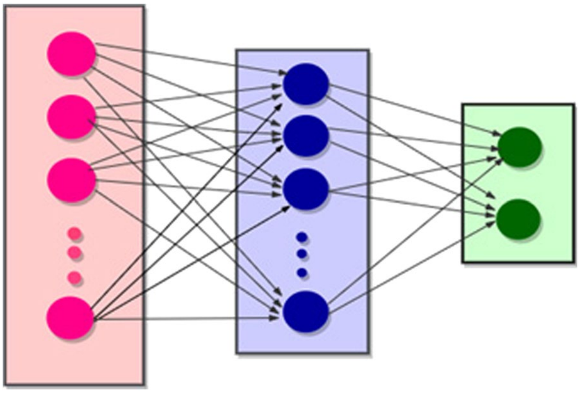


Figure 7. An illustration of the input, hidden and output layers of the neural network.

$$\vartheta(\xi_l, \xi_m) = \sqrt{\Delta_s^*(\xi_l)^2 |\Delta_s^*(\xi_l) - \Delta_s^*(\xi_m)|^2} + \frac{|(\Delta_1^*(\xi_l) - \overline{\Delta}_1^*(\hat{a}))\,(\Delta_3^*(\xi_m) - \overline{\Delta}_3^*(\hat{a})|}{\sqrt{\sum_{l=1}^{20}(\Delta_1^*(\xi_l) - \overline{\Delta}_1^*(\hat{a}))^2 \sum_{m=1}^{20}(\Delta_3^*(\xi_m) - \overline{\Delta}_3^*(\hat{a}))^2}} \tag{24}$$

$$\vartheta(\xi_l, \xi_m) = \sqrt{\Delta_s^*(\xi_l)^2 |\Delta_s^*(\xi_l) - \Delta_s^*(\xi_m)|^2} + \frac{|(\Delta_2^*(\xi_l) - \overline{\Delta}_2^*(\hat{a}))\,(\Delta_3^*(\xi_m) - \overline{\Delta}_3^*(\hat{a})|}{\sqrt{\sum_{l=1}^{20}(\Delta_2^*(\xi_l) - \overline{\Delta}_2^*(\hat{a}))^2 \sum_{m=1}^{20}(\Delta_3^*(\xi_m) - \overline{\Delta}_3^*(\hat{a}))^2}} \tag{25}$$

where Δ_1^*, Δ_2^*, Δ_3^* represents the normalized hydrophobicity, hydrophilicity and side-chain mass respectively. The values used in (23) to (25) are normalized with (26), and standardized with preferred range such that (−R, R), where R is the number in which $\hat{a}$ amino acids are being standardized. The hydrophobicity values are taken from Tanford C.[64], hydrophilicity assesses are assumed from Hopp T. P., Woods K. R.[65] and side-chain mass values are generally available in any biochemistry text book.

$$\begin{aligned} \Delta_1^*(\hat{a}) &= \left[\frac{2R}{\left(\Delta_{1_{(max)}} - \Delta_{1_{(min)}}\right)}\left(\Delta_1(\hat{a}) - \Delta_{1_{(max)}}\right)\right] + R \\ \Delta_2^*(\hat{a}) &= \left[\frac{2R}{\left(\Delta_{2_{(max)}} - \Delta_{2_{(min)}}\right)}\left(\Delta_2(\hat{a}) - \Delta_{2_{(max)}}\right)\right] + R \\ \Delta_3^*(\hat{a}) &= \left[\frac{2R}{\left(\Delta_{3_{(max)}} - \Delta_{3_{(min)}}\right)}\left(\Delta_3(\hat{a}) - \Delta_{3_{(max)}}\right)\right] + R \end{aligned} \tag{26}$$

Feature set is characterized with a vector having two hundred and twenty (220) elements, where the first sixty components based on the hydrophobic nature of amino acids, second sixty represents their hydrophilic nature, next sixty components contain information regarding side chain mass of 20 amino acids and the last forty elements reflect the positions as well as composition of individual amino acids residues. This novel predictor establishes outstanding outcomes towards recognition of roll down protein sequences. These extracted vectors obtained for training data are further used to train Neural Networks (NN) based classifier.

Neural networks are one of the most powerful techniques used to solve decision problems. They work by receiving labelled inputs and hence gain experience which help them to develop an opinion regarding arbitrary input for test purposes. After training process is completed the network seemingly behaves in a way that makes it capable to classify each given input within an acceptable degree of accuracy. During the learning process the network adjusts its weights such that the error is minimized which essentially translates into improved learning and increased accuracy[17]. A multilayer neural network was used to tackle this problem (see Fig. 7). The feature vector constructed for the prediction of signal peptides consists of 220 coefficients. Its connectionist architecture consists of 220 input layer neurons, 50 hidden layer neurons and two output neurons that discern among secretory and non-secretory poly-peptide chains. The training of the multilayered neural network is performed using back propagation method. In order to reduce the error and increase the prediction accuracy gradient descent technique was used along with an adaptive learning rate.

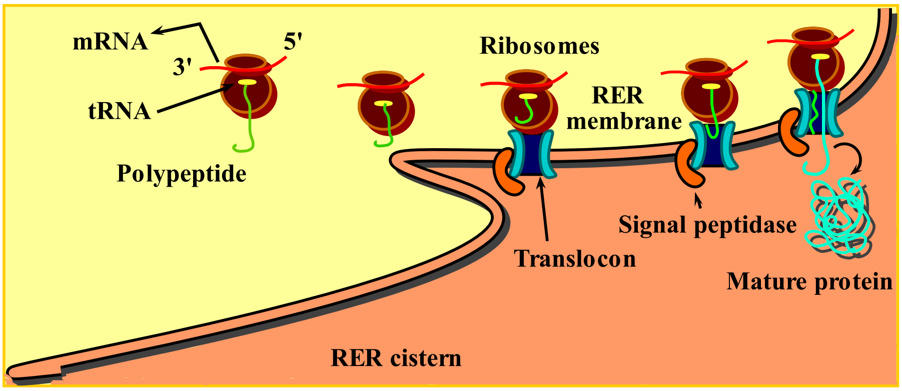


Figure 8. Signal peptidase is an enzyme that removes the signal of translocated primary proteins from the membrane to exhibit its mature form when they are substituted from a cytoplasmic position of synthesis to extracytoplasmic regions. Ultimately, these cleaved signal peptides are directed towards secretory tract.

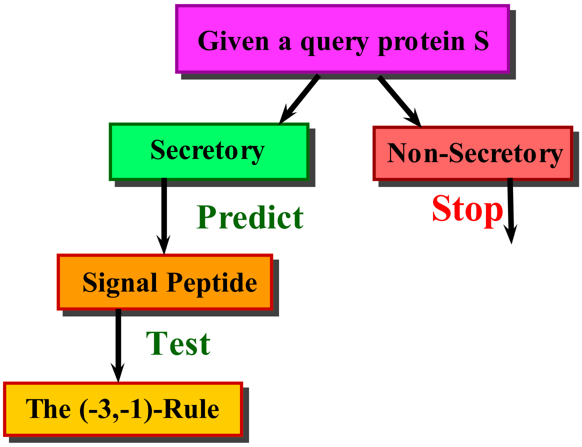


Figure 9. Pictorial representation shows how to predict the signal peptide and its cleavage site by means of proposed tool and the "(−3, −1)"-rule.

Respective output units are ultimately brought together through individual units of input, output representation[19].

Signal peptidase perform a momentous role in order to cleave signal sequence and the mature peptide from the nascent protein. Signal peptide is found in the vicinity of N-terminus site of protein sequence. Customarily, signal consists of 3–60 amino acid residues[8]. Translocon duct allows the passing of signal sequence transversely (see Fig. 8). Peptidases firstly confronts a nascent protein within endoplasmic reticulum (ER)[66]. Signal peptidase is also encountered in prokaryotes[67] When protein builds appurtenance for mitochondria and chloroplasts.

A signal cleavage site for signal peptide is revealed by dividing concatenation into 3 parts, N-terminal part is basic positively charged and is labeled as the n-region, the central hydrophobic region is labeled as h-region and the C-terminal part describes the more polar site of the sequence and is represented as a c-region. Signal is encountered through the n-region to h-region by occupying positions −3 and −1. The application of "(−3, −1) -rule" proves very productive in identifying signal cleavage site directly. As discussed previously the dataset is divided in two categories, eukaryotes and prokaryotes (Gram-positive and Gram-negative) to predict the signal sequence and its cleavage site in both categories. Signal is mostly encountered embedded within the concatenation segregating the signal sequence and mature protein chain. The cleavage site of a secretory protein is determined by following these steps: First count the amino acid residues at each position in the sequence, formally, P($\hat{a}$, i), where P is the count factor for the occurrence of residue of type $\hat{a}$ at position i. Subsequently build Weight-matrices Q($\hat{a}$, i) by dividing all counts by their diverse expected abundance in proteins, primarily, $\langle P(\hat{a})\rangle$. Taking the natural logarithms of these entities for all sequences arrayed from their accepted sites of removed peptide chain between positions −1 and +1, t follows in equation (27).

$$Q(\hat{a}, i) = ln\left(\frac{P(\hat{a}, i)}{\langle P(\hat{a})\rangle}\right) \tag{27}$$

Eukaryotic and prokaryotic Signal concatenation splitter follows the "(−3, −1) -rule"[10,23]. Statistical analysis shows that any residue out of Ala, Ser, Gly, Cys, Thr is placed at −1 location respect to the cleavage site while −3 site is occupied by Asp, Glu, Lys, Arg, Asn, Gln, (but not Phe, His, Tyr, Try), furthermore, there is no Pro residue between −3 to +1. Similarly, for Prokaryotic proteins same rule applies with a different set of residues. The −1 location is occupied by any of Ala, Gly, Ser, Thr whereas −3 is occupied by any of Ala, Gly, Leu, Ser, Thr, Val, also, −7 and −8 is mostly occupied by Leu or any other hydrophobic residue other than Val, Phe. In addition, it was also submitted that Proline (Pro) necessarily will be missed in the region −3 over +1.

A scanning algorithm was developed to search the cleavage site pattern. It transcribed a weight matrix onto the polypeptide sequence. The residues bearing significance in identifying the cleavage site were assigned higher non-zero values while others were substituted by a zero value, hence the primary sequence was transformed into a vector. The algorithm worked by identifying a spike among the neighboring elements of the vector[10,23]. Systematic drawing reflects the overall procedure involved in predicting a signal peptide and determining its cleavage site (see Fig. 9).

References

1. De Souza, G. A., Leversen, N. A., Målen, H. & Wiker, H. G. Bacterial proteins with cleaved or uncleaved signal peptides of the general secretory pathway. *J. proteomics* **75**, 502–510 (2011).
2. Von Heijne, G. The signal peptide. *J. Membr. Biol.* **115**, 195–201 (1990).
3. Zheng, N. & Gierasch, L. M. Signal sequences: the same yet different. *Cell* **86**, 849–852 (1996).
4. Shen, H.-B. & Chou, K.-C. Signal-3L: A 3-layer approach for predicting signal peptides. *Biochem. biophysical research communications* **363**, 297–303 (2007).
5. Hagmann, M. Protein zip codes make nobel journey. *Sci.* **286**, 666–666 (1999).
6. Blobel, G. & Dobberstein, B. Transfer of proteins across membranes. i. presence of proteolytically processed and unprocessed nascent immunoglobulin light chains on membrane-bound ribosomes of murine myeloma. *The J. cell biology* **67**, 835–851 (1975).
7. Rapoport, T. A. Transport of proteins across the endoplasmic reticulum membrane. *Science-New York Then Washington-* **258**, 931–931 (1992).
8. Chou, K.-C. & Shen, H.-B. Signal-CF: a subsite-coupled and window-fusing approach for predicting signal peptides. *Biochem. biophysical research communications* **357**, 633–640 (2007).
9. McGeoch, D. J. On the predictive recognition of signal peptide sequences. *Virus research* **3**, 271–286 (1985).
10. Von Heijne, G. A new method for predicting signal sequence cleavage sites. *Nucleic acids research* **14**, 4683–4690 (1986).
11. Folz, R. J. & Gordon, J. I. Computer-assisted predictions of signal peptidase processing sites. *Biochem. biophysical research communications* **146**, 870–877 (1987).
12. Ladunga, I., Czako, F., Csabai, I. & Geszti, T. Improving signal peptide prediction accuracy by simulated neural network. *Bioinforma.* **7**, 485–487 (1991).
13. Arrigo, P., Giuliano, F., Scalia, F., Rapallo, A. & Damiani, G. Identification of a new motif on nucleic acid sequence data using kohonen's self-organizing map. *Bioinforma.* **7**, 353–357 (1991).
14. Nielsen, H., Engelbrecht, J., Brunak, S. & Von Heijne, G. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein engineering* **10**, 1–6 (1997).
15. Emanuelsson, O., Nielsen, H. & Von Heijne, G. ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. *Protein Sci.* **8**, 978–984 (1999).
16. Chou, K.-C. Using subsite coupling to predict signal peptides. *Protein Eng.* **14**, 75–79 (2001).
17. Jagla, B. & Schuchhardt, J. Adaptive encoding neural networks for the recognition of human signal peptide cleavage sites. *Bioinforma.* **16**, 245–250 (2000).
18. Menne, K. M., Hermjakob, H. & Apweiler, R. A comparison of signal sequence prediction methods using a test set of signal peptides. *Bioinforma.* **16**, 741–742 (2000).
19. Reinhardt, A. & Hubbard, T. Using neural networks for prediction of the subcellular location of proteins. *Nucleic acids research* **26**, 2230–2236 (1998).
20. Frank, K. & Sippl, M. J. High-performance signal peptide prediction based on sequence alignment techniques. *Bioinforma.* **24**, 2172–2176 (2008).
21. Hiller, K., Grote, A., Scheer, M., Münch, R. & Jahn, D. PrediSi: prediction of signal peptides and their cleavage positions. *Nucleic acids research* **32**, W375–W379 (2004).
22. Nielsen, H. & Krogh, A. Prediction of signal peptides and signal anchors by a hidden markov model. *In Ismb*, vol. 6, 122–130 (1998).
23. Heijne, G. Patterns of amino acids near signal-sequence cleavage sites. *The FEBS J.* **133**, 17–21 (1983).
24. Lal, P., Au-Young, J., Reddy, R., Murry, L. E. & Mathur, P. Signal peptide-containing proteins. US Patent 5,932,445 (1999).
25. Wang, D. & Huang, G.-B. Protein sequence classification using extreme learning machine. In *Neural Networks*, *2005. IJCNN'05. Proceedings. 2005 IEEE International Joint Conference on*, vol. 3, 1406–1411 (IEEE, 2005).
26. Cao, J. & Xiong, L. Protein sequence classification with improved extreme learning machine algorithms. *BioMed research international* **2014** (2014).
27. Höglund, A., Dönnes, P., Blum, T., Adolph, H.-W. & Kohlbacher, O. MultiLoc: prediction of protein subcellular localization using n-terminal targeting sequences, sequence motifs and amino acid composition. *Bioinforma.* **22**, 1158–1165 (2006).
28. Pearson, W. R. & Lipman, D. J. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci.* **85**, 2444–2448 (1988).
29. Zhang, J. & Liu, B. Psfm-dbt: identifying dna-binding proteins by combing position specific frequency matrix and distance-bigram transformation. *Int. journal molecular sciences* **18**, 1856 (2017).
30. Liu, B., Yang, F. & Chou, K.-C. 2l-pirna: A two-layer ensemble classifier for identifying piwi-interacting rnas and their function. *Mol. Ther. Acids* **7**, 267–277 (2017).
31. Liu, B., Wu, H., Zhang, D., Wang, X. & Chou, K.-C. Pse-analysis: a python package for dna/rna and protein/peptide sequence analysis based on pseudo components and kernel methods. *Oncotarget* **8**, 13338 (2017).
32. Liu, B., Chen, J. & Wang, X. Application of learning to rank to protein remote homology detection. *Bioinforma.* **31**, 3492–3498 (2015).
33. Chen, J., Guo, M., Li, S. & Liu, B. Protdec-ltr2. 0: an improved method for protein remote homology detection by combining pseudo protein and supervised learning to rank. *Bioinforma.* **33**, 3473–3476 (2017).
34. Liu, B., Liu, F., Fang, L., Wang, X. & Chou, K.-C. Repdna: a python package to generate various modes of feature vectors for dna sequences by incorporating user-defined physicochemical properties and sequence-order effects. *Bioinforma.* **31**, 1307–1309 (2014).
35. Tusnady, G. E. & Simon, I. The hmmtop transmembrane topology prediction server. *Bioinforma.* **17**, 849–850 (2001).
36. Chou, K.-C. Prediction of protein subcellular locations by incorporating quasi-sequence-order effect. *Biochem. biophysical research communications* **278**, 477–483 (2000).

37. Chou, K.-C. Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins: Struct. Funct. Bioinforma.* **43**, 246–255 (2001).
38. Diplaris, S., Tsoumakas, G., Mitkas, P. A. & Vlahavas, I. Protein classification with multiple algorithms. *In Panhellenic Conference on Informatics*, 448–456 (Springer, 2005).
39. Gomi, M., Sonoyama, M. & Mitaku, S. High performance system for signal peptide prediction: Sosuisignal. *Chem-bio informatics journal* **4**, 142–147 (2004).
40. Viklund, H., Bernsel, A., Skwark, M. & Elofsson, A. SPOCTOPUS: a combined predictor of signal peptides and membrane protein topology. *Bioinforma.* **24**, 2928–2929 (2008).
41. Khan, Y. D., Ahmad, F. & Anwar, M. W. A neuro-cognitive approach for iris recognition using back propagation. *World Appl. Sci. J.* **16**, 678–685 (2012).
42. Khan, Y. D., Khan, S. A., Ahmad, F. & Islam, S. Iris recognition using image moments and k-means algorithm. *The Sci. World J.* **2014** (2014).
43. Khan, Y. D., Ahmed, F. & Khan, S. A. Situation recognition using image moments and recurrent neural networks. *Neural Comput. Appl.* **24**, 1519–1529 (2014).
44. Butt, A. H., Khan, S. A., Jamil, H., Rasool, N. & Khan, Y. D. A prediction model for membrane proteins using moments based features. *BioMed research international* **2016** (2016).
45. Butt, A. H., Rasool, N. & Khan, Y. D. A treatise to computational approaches towards prediction of membrane protein and its subtypes. *The J. membrane biology* **250**, 55–76 (2017).
46. Chen, W., Feng, P.-M., Lin, H. & Chou, K.-C. iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. *Nucleic acids research* **41**, e68–e68 (2013).
47. Xu, Y., Ding, J., Wu, L.-Y. & Chou, K.-C. iSNO-PseAAC: predict cysteine s-nitrosylation sites in proteins by incorporating position specific amino acid propensity into pseudo amino acid composition. *PLoS One* **8**, e55844 (2013).
48. Lin, H., Deng, E.-Z., Ding, H., Chen, W. & Chou, K.-C. iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition. *Nucleic acids research* **42**, 12961–12972 (2014).
49. Jia, J., Liu, Z., Xiao, X., Liu, B. & Chou, K.-C. pSuc-Lys: predict lysine succinylation sites in proteins with pseaac and ensemble random forest approach. *J. theoretical biology* **394**, 223–230 (2016).
50. Qiu, W.-R., Xiao, X., Xu, Z.-C. & Chou, K.-C. iPhos-PseEn: identifying phosphorylation sites in proteins by fusing different pseudo components into an ensemble classifier. *Oncotarget* **7**, 51270 (2016).
51. Zhang, C.-J. *et al*. iOri-Human: identify human origin of replication by incorporating dinucleotide physicochemical properties into pseudo nucleotide composition. *Oncotarget* **7**, 69783–69793 (2016).
52. Chen, W. *et al*. iRNA-AI: identifying the adenosine to inosine editing sites in rna sequences. *Oncotarget* **8**, 4208 (2017).
53. Liu, B., Wang, S., Long, R. & Chou, K.-C. iRSpot-EL: identify recombination spots with an ensemble learning approach. *Bioinforma.* **33**, 35–41 (2016).
54. Rane, N. S., Chakrabarti, O., Feigenbaum, L. & Hegde, R. S. Signal sequence insufficiency contributes to neurodegeneration caused by transmembrane prion protein. *The J. cell biology* **188**, 515–526 (2010).
55. Castro-Fernandez, C., Maya-Nunez, G. & Conn, P. M. Beyond the signal sequence: protein routing in health and disease. *Endocr. Rev.* **26**, 479–503 (2004).
56. Xu, Y., Wen, X., Shao, X.-J., Deng, N.-Y. & Chou, K.-C. iHyd-PseAAC: Predicting hydroxyproline and hydroxylysine in proteins by incorporating dipeptide position-specific propensity into pseudo amino acid composition. *Int. journal molecular sciences* **15**, 7594–7610 (2014).
57. Wei, L., Xing, P., Shi, G., Ji, Z.-L. & Zou, Q. Fast prediction of protein methylation sites using a sequence-based feature selection technique. *IEEE/ACM Transactions on Comput. Biol. Bioinforma.* (2017).
58. Wei, L., Xing, P., Tang, J. & Zou, Q. PhosPred-RF: a novel sequence-based predictor for phosphorylation sites using sequential information only. *IEEE Transactions on NanoBioscience* (2017).
59. Wei, L., Tang, J. & Zou, Q. Local-DPP: An improved dna-binding protein prediction method by exploring local evolutionary information. *Inf. Sci.* **384**, 135–144 (2017).
60. Wei, L. *et al*. Improved prediction of protein–protein interactions using novel negative samples, features, and an ensemble classifier. *Artif. Intell. Medicine* (2017).
61. Xing, P., Su, R., Guo, F. & Wei, L. Identifying n6-methyladenosine sites using multi-interval nucleotide pair position specificity and support vector machine. *Sci. Reports* **7** (2017).
62. Wei, L. *et al*. CPPred-RF: a sequence-based predictor for identifying cell-penetrating peptides and their uptake efficiency. *J. Proteome Res.* **16**, 2044–2053 (2017).
63. Su, R. *et al*. Detection of tubule boundaries based on circular shortest path and polar-transformation of arbitrary shapes. *J. microscopy* **264**, 127–142 (2016).
64. Tanford, C. Contribution of hydrophobic interactions to the stability of the globular conformation of proteins. *J. Am. Chem. Soc.* **84**, 4240–4247 (1962).
65. Hopp, T. P. & Woods, K. R. Prediction of protein antigenic determinants from amino acid sequences. *Proc. Natl. Acad. Sci.* **78**, 3824–3828 (1981).
66. Milstein, C., Brownlee, G., Harrison, T. M. & Mathews, M. A possible precursor of immunoglobulin light chains. *Nat.* **239**, 117–120 (1972).
67. Paetzel, M., Karla, A., Strynadka, N. C. & Dalbey, R. E. Signal peptidases. *Chem. reviews* **102**, 4549–4580 (2002).

Author Contributions

A.E. proposed an algorithm and conducted experiments, K.M. supervised the results. Y.D. supervised the validation process. S.K. worked out the appropriateness of the suggested model and validation techniques. K.C. originally formulated the problem contributed equally to analyze and improve the results. All authors reviewed the manuscript.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-018-19491-y.

Competing Interests: The authors declare that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.