Unfolding the Locale of Membrane Proteins within Cellular Alcove by SVM

Mehwish Faiz[1,2], Saad Jawaid Khan[1,*], Fahad Azim[2], Sumaya Abid[1], Areej Ahmed[1],

[1] Department of Biomedical Engineering, Ziauddin University (FESTM), Karachi, Pakistan.
[2]Department of Electrical Engineering, Ziauddin University (FESTM), Karachi, Pakistan.
[*]Corresponding author: sj.khan@zu.edu.pk

Abstract

The Membrane Proteins (MPs) on the cell membrane, is an ideal targets for drug delivery owing to their distinct location on the cell membrane and intracellular structure. To locate the multitude of MPs with better accuracy, we implemented SVM on the Benchmark MP Database, and MemLoci Dataset. A total of 3000 proteins are selected with 1000 plasma membrane proteins, 1000 internal membrane proteins, and 1000 organelle membrane proteins. The amino acid sequence of the proteins is converted into Pseudo Amino Acid Code. After feature extraction, the data is trained and tested, yielding an overall accuracy of 78%. The output is displayed through a Graphical User Interface (GUI) which discloses the category of MP based on the cellular site where they are residing.

Keywords—Membrane Proteins, location, Pseudo Amino Acid Code, Graphical User Interface (GUI), Support Vector Machine (SVM)

1 Introduction

Living cells and the cell organelles are encapsulated by a protective layer of lipid, known as a membrane. The proteins associated with these membranes are the membrane protein (MP). These proteins have a 20% to 30% contribution to the entire genome which is cipher only for membrane proteins [1]. MP act as boundaries or barriers as they separate the cell its organelles from the outer environment. They also allow specific molecules or nutrients to enter or exit. Their transport mechanism relies on acting as protein channels and ions pump to maintain homeostasis. They act as carriers, receptors for signal transduction, intracellular and extracellular signaling molecules, and also as enzymes, thus playing an extensive role to carry out cellular activities [2]-[3]. Every membrane protein executes a distinct function, which is directly associated with its locale within the cellular alcove. Moreover, the physicochemical characteristics of the MP depend on Hydrophobicity, Hydrophilicity, and mass of the side chain along with the basic framework of the sequence of amino acids. These MPs have a legitimate structural orientation in 3D space based on the presence of Alpha Helical regions and Beta Barrel proteins [4].

Based on the evolutionary and functional basis, the MP in eukaryotes is categorized as the plasma membrane, internal membranes, and organelle membrane protein. The proteins existing on the plasma membrane are the plasma membrane proteins. Internal Membrane Proteins are the MP that originated from the preliminary prokaryotic plasma membrane and are resident of the components of the endomembrane system of the cell. The habitat of the Organelle Membrane Proteins is the mitochondria and plastid. Thus, this segregation of MP refers to a diversified location in the cell which exemplifies their particular tasks [5].

As the number of protein sequences is increasing rapidly in the post-genomic era [6]-[7], experimental biological techniques are antiquated, specifically, when dealing with a large number of proteins. Thus, computational techniques have turned out with the key advantages of rescuing time and expense. Multiple computational methods including SVM, KNN, gene ontology, Fuzzy KNN, and tolerance predictors have been implemented to elucidate the structures and functions of membrane proteins and their types [8]-[10].

Machine Learning algorithms extract the locale

ISSN: 2523-0379 (Online), ISSN: 1605-8607 (Print)
DOI: https://doi.org/10.52584/QRJ.2101.05

information of a protein from its monomer, the amino acids. However, this amino acid composition does not provide adequate information as it only contains the native twenty components, thus, re-vealing its composition only. The sequence order of amino acids is omitted which affects the results. This issue is resolved by introducing the concept of Pseudo Amino acid Composition (PseAAC) which incorporate the sequence order into an amino acid composition. This incorporation of the arrangement of amino acids in the protein chain has drastically improved the out-come of the machine learning algorithms in terms of accuracy [11]-[14].

2 Literature Survey

The exploration of the literature in this arena has the given findings.

In 2008, an article reported their work on the localization of outer membrane proteins only through some combined features and implemented SVM to discriminate outer membrane proteins from the other types [15].

An article was published in 2011, in which researchers identified the locations of internal, organelle, and plasma membrane proteins on eukaryotes through their designed predictor named 'MemLoci' and performed a ten-fold cross-validation test on the extracted dataset from SwissProt and achieved 70% accuracy [6].

Then again in 2011, a research question based on the identification of the location of membrane proteins was answered by using HMMs and SVM techniques by taking data from SwissProt. The location of cell membrane protein, internal, and organelle membrane protein was identified with an accuracy of 70% [16].

In 2012, research was conducted on membrane proteins by using the Projected Gene Ontology Score algorithm to sort out the location problem. However, this method yields an accuracy of 87% but the limitation is that it can discriminate between 2 locations only [17].To improve the localization of membrane proteins, multiple attributes were blended in a parallel fashion, then a principal component analytic technique was implemented to reduce the dimensionality, this work was reported in 2012 [18]. SVM ensemble-based approach was adopted in 2014 to get an accuracy of 63% [19]. In 2018, an innovative approach BUSCA was adopted by combining different methods to segregate the location of membrane proteins from globular proteins [5]. In 2019, Tommaso Orioli et al. tried to resolve the localization problem for single-pass and multi-pass membrane proteins. The proposed approach yields a better outcome for single-pass proteins, however, for multi-pass proteins promising results are achieved [15].

In 2020, 'SCLpred-MEM' was proposed to locate whether the unknown protein belongs to membrane protein or non-membrane protein. This method makes use of the dataset from UniProtKB and achieved 81.25% accuracy in five-fold cross-validation. However, this predictor yielded a better outcome but the lacking is that it only discriminates whether the protein is MP or not [20]. The literature search reveals that machine learning approaches disclosing the location of membrane proteins are very few, as the majority of the predictors available are for identifying the type of Membrane proteins. Another flaw is that only SVM has been implemented on the membrane proteins Benchmark dataset, MemLoci, with 70% accuracy. Any predictor with a better feature extraction technique including Pseudo Amino Acid Composition can improve the prediction results for membrane proteins localization research problems [21]-[22]. Moreover, the review of the various research articles demonstrates that the Pseudo amino acid composition has not been used for the MemLoci dataset. Thus, we implemented SVM on the MemLoci dataset with Pseudo Amino Acid composition.

3 Materials and Methods

3.1 Dataset

Mem Loci database is a freely accessible archive, a collection of the membrane proteins with plasma membrane, internal membrane, and organelle membrane proteins as three categories. This database is initially extracted from the Swiss Prot database and is managed by Biocomputing Group – University of Bologna. The redundancy of the dataset is already reduced by incorporating a cut-off threshold value of 20% and is available at https://mu2py.biocomp.unibo.it/memloci/. From this database, we extracted 3000 sequences with 1000 sequences of each class of membrane proteins i.e., internal, organelle, and plasma membrane proteins. The sequences were downloaded in FASTA format.

3.2 Feature Extraction

The FASTA Format of the amino acid sequences is then converted into Pseudo Amino Acid by employing Chou's Pseudo Amino Acid Converter (PseAAC) http://www.csbio.sjtu.edu.cn/bioinf/PseAAC/. The

6 distinguishing features that exemplify the properties of membrane proteins are Hydrophobicity, Hydrophilicity, pK1(alpha-COOH), pK2 (NH3), pI (Isoelectric point) and Mass of the side chain. These features are used to evaluate the impact of various amino acid positions along the polypeptide chain of proteins.

We take account of these characteristics while converting into PseAAC for the feature vector. A protein PseAAC is indicated by more than 20 different parameters. The first 20 factors relate to elements of their amino acid makeup with the information of specific amino acids that are present along the protein chain, whilst the further factors unfold the sequence order information with the nitty-gritty of the location of 1st amino acid, then 2nd amino acid and so on. We selected the following optimum parameters for the conversion of Amino Acid sequences into Pseudo Amino Acid Com-position as:

- Type 2 PseAA Mode, which is the series correlation type and generates discrete values in the form of $20 + i * \lambda$, where the first 20 values generated on output are Amino Acid values and reaming values are related to the characteristic values of the amino acid sequence.
- 0.05 Weight Factor, which is the order effect [23],[24],[25].
  The protein sequence entered into the PseAAC converter (available at: http://www.csbio.sjtu.edu.cn/bioinf/PseAAC/) yields numeric values as the output in which :
  - The first 20 values are of amino acid codes.
  - Rest are the values of correlation factors.

3.3 Flow Chart of the Proposed Approach

The proposed approach is based on the extraction of data from a benchmark membrane protein database for which the MemLoci Database is selected. Then feature extraction is done through Pseudo Amino Acid Converter. After that, MATLAB Software is used for cross-validation and classification. The outcome of the proposed predictor is displayed through GUI. Fig 1. depicts the flow diagram of the proposed framework.

3.4 Support Vector Machine

SVM is a widely used labeled training data classifier based on discriminating the data into different classes by generating a hyperplane. This hyperplane split the data into two distinct categories for binary classification. However, for multi-class classification, the one-vs-all approach is used. To unfold the

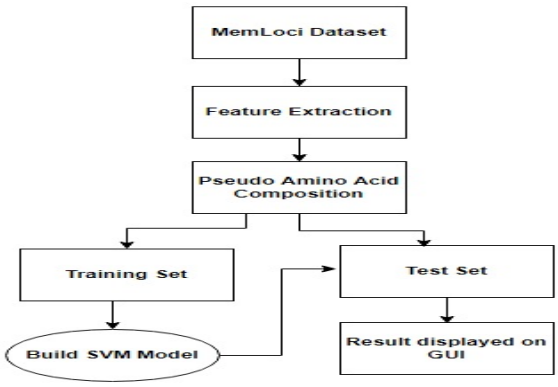

Fig. 1: Flow chart of the proposed Model

location of three distinct classes of MP, this approach is employed which makes use of three distinct binary SVM to discriminate the proteins of a single Class from the other two classes by creating three hyperplanes. Thus, a hyperplane line segregates the Plasma membrane proteins from the Internal Membrane and Organelle Membrane Proteins. Then Internal Membrane Proteins are isolated from the Plasma membrane and Organelle Membrane Proteins by another hyperplane line. Mathematically, it can be written as:

$$\{PMP\}vs\{notPMP\}, \{IMP\}vs \{notIMP\}, \{OMP\}vs\{notOMP\} \quad (1)$$

where
$PMP = PlasmaMembraneProteins$
$IMP = InternalMembraneProteins$
$OMP = OrganelleMembraneProteins$

The pictorial representation of the working of the multi-class classification of the SVM approach is depicted in Figure 2.

3.5 Implementation of the Classifier

SVM is applied using MATLAB software and a graphical user interface is created for prediction. The one-vs-all technique is used to train SVM for the classification of multiple classes. Where a dataset of 3000 is disintegrated into training and testing sets with 2500 sequences for training purposes and 500 sequences for testing purpose. Fig 3. depicts the proportion of the training and testing dataset employed in the study.

For the evaluation of the performance of the model, the hold-out cross-validation technique was used. Then on the basis of probability, the protein class with the

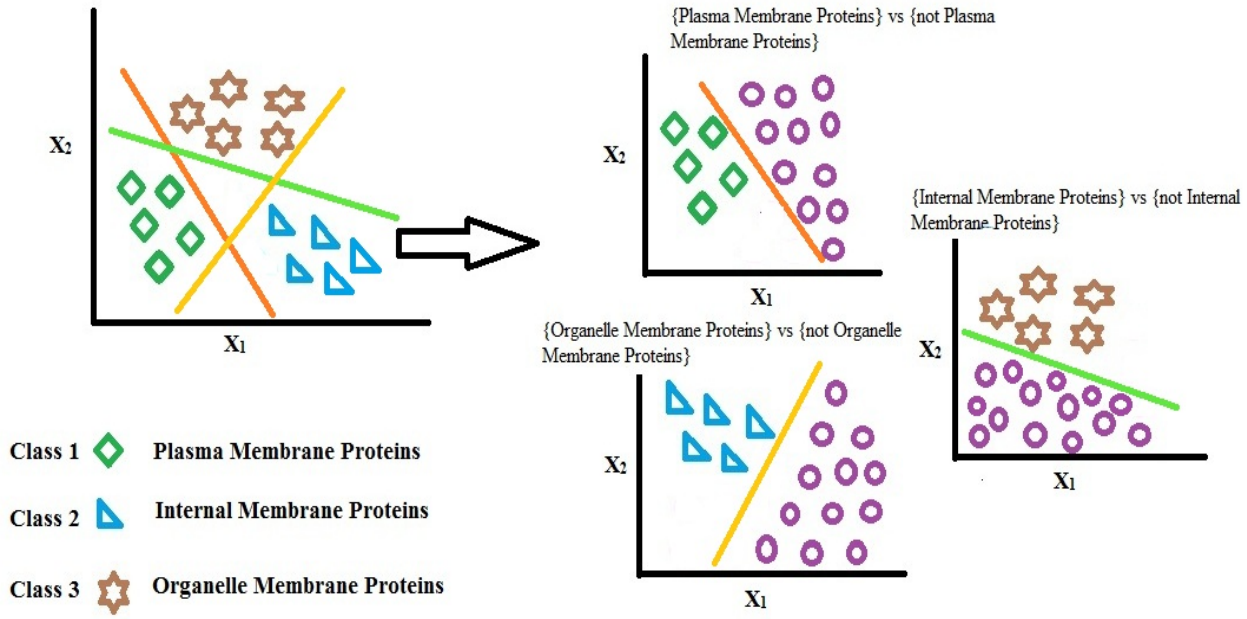

Fig. 2: One vs all SVM technique to classify 3 distinct classes of membrane proteins

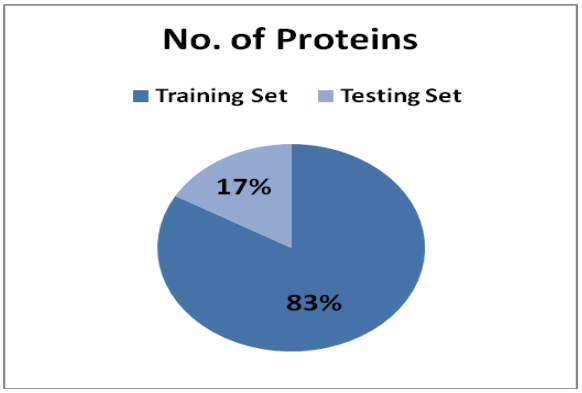

Fig. 3: Training and Testing set for the proposed approach

TABLE 1: Overall accuracy yielded by implementing SVM

ML Algorithm	Training Data	Testing Data	Validation	Overall Accuracy
SVM	2500 proteins	500 proteins	Hold out Cross Validation	78%

highest probability is detected as the revealed location. This probabilistic model gives an overall accuracy of 78% and a confusion matrix is also generated. Table 1. manifests the implementation of the SVM classifier on the MemLoci dataset.

4 Results

The predicted output is displayed on the GUI in the form of probability of all the classes with the forecast of the location of the protein by revealing the actual class of the protein in a separate bar as the locale of membrane protein either as an internal membrane protein or Organelle Membrane Proteins or Plasma Membrane Proteins in terms of probability. The occurrence of sequence similarity is also depicted in terms of probability between all three protein types, with the highest probability representing the actual membrane protein.

Fig 5. depicts the confusion matrix for the three types of Membrane Proteins with distinct True positive values (TP), True Negative values (TN), False Positive values (FP), and False Negative values (FN).

To evaluate the performance of the predictor, six indicators by published recommendations have been applied [26]. These indicators are based on a confusion matrix, where data items are categorized as true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), and actual conditions are contrasted with predicted results. Accuracy is a substantial parameter to arbitrate the efficiency of the predictor and is dependent on the number of truly classified sequences and the total number of test sequences. Mathematically,

$$Accuracy = \frac{TP + TN}{FP + FN + TP + TN} \qquad (2)$$

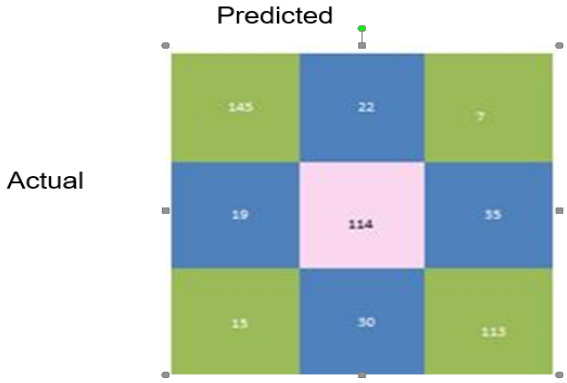

Fig. 4: Confusion Matrix for the proposed approach

The ratio of actual positive results to all positive predictions is known as the positive predictive value (PPV). Another terminology used for PPV is Precision and can find out by,

$$PPV = \frac{TP}{FP + TP} \tag{3}$$

Similarly, the ratio of true negative results to all negative predictions is known as the negative predictive value (NPV).

$$NPV = \frac{TN}{FN + TN} \tag{4}$$

The False positive rate(FPR) is obtained by False positive and True negative values.

$$FPR = \frac{FP}{FP + TN} \tag{5}$$

The TP rate divided by the total number of positive conditions is the sensitivity. It is also referred to as true positive rate (TPR) and given as,

$$Sensitivity = \frac{TP}{FN + TP} \tag{6}$$

The rate of TP over all of the negative outcomes is known as the specificity or true negative rate (TNR).

$$Specificity = \frac{TN}{FP + TN} \tag{7}$$

Matthew's correlation coefficient is regarded as a balanced indicator as it incorporates all the values of the contingency table.

$$MCC = \frac{TPxTN - FPxFN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{8}$$

Specificity, Sensitivity, PPV, and NPV do not accurately reflect all facets of performance as these parameters are estimated by using only half of the data in the contingency table. Therefore, accuracy and MCC are more balanced, representative, and all-encompassing than the line- or column-wise metrics.

Table 2 depicts the performance of the predictor for all the proteins with the values of PPV, NPV, Sensitivity, Specificity, and Accuracy.

The above scoring indexes for the performance of the proposed predictor indicate that the internal membrane proteins are better predicted as compared to other classes as the Accuracy and the Mathews Co-relation Coefficient (MCC) have the highest values.

Table 3 depicts the variation in the outcomes of the proposed approach as compared to other predictors. The parameters compared are FPR, Recall, and MCC with MCC as a more authentic statistical measure that only yields a high score if the pre-diction performed well in each of the four categories of the confusion matrix (true positives, false negatives, true negatives, and false positives). Similarly, a higher value of Recall is associated with a good prediction quality. It is evident from Table III that the Recall, and MCC for the Internal Membrane Proteins, Organelle Membrane Proteins, and Plasma Membrane Proteins are improved with the adopted method in comparison with other approaches.

Table 3 portrays the accuracies for the three classes of membrane proteins with the proposed approach and Mem-Loci predictor. With the adopted methodology the accuracy obtained for the Internal Membrane Proteins is 87%, for Organelle Membrane Proteins is 78% and for Plasma Membrane Proteins is 82%. However, the outcome in the form of accuracy is abating for all the 3 categories of MP with MemLoci Predictor.

5 Conclusion

The environment in which proteins function is exclusively determined by their subcellular localization. As a result, subcellular localization affects protein function by regulating the availability and access to all different kinds of interacting molecules including biomolecules or drugs. Thus, understanding protein localization frequently plays a key role in defining the physiological function of speculative and recently found proteins. Machine Learning Classification aids in classifying different membrane proteins thus if a new membrane protein is discovered we can easily categorize it by inputting the amino acid sequence in the predictor. Moreover, the correct categorization of a protein can be helpful in Computer Aided Drug

TABLE 2: Performance Evaluation of the Proposed PredictorPerformance Evaluation of the Proposed Predictor

	PPV	NPV	FPR	Sensitivity	Specificity	Accuracy	MCC
Internal Membrane Proteins	0.81	0.90	0.10	0.83	0.89	0.87	0.72
Organelle Membrane Proteins	0.68	0.83	0.15	0.67	0.84	0.78	0.52
Plasma Membrane Proteins	0.72	0.86	0.12	0.71	0.87	0.82	0.59

TABLE 3: Comparison of the outcome of the proposed approach with MemLoci Predictor

	MemLoci Predictor			Proposed Approach		
	FPR (%)	Accuracy (%)	MCC	FPR (%)	Accuracy (%)	MCC
Internal Membrane Proteins	30	72	0.42	10	87	0.72
Organelle Membrane Proteins	9	70	0.60	15	78	0.52
Plasma Membrane Proteins	15	56	0.43	12	82	0.59
Overall		66			78	

Designing in which different drugs (molecules) are checked by their interaction with the binding sites of proteins. Therefore, we implemented SVM on the benchmark dataset MemLoci with Pseudo Amino Acid Composition as the chosen feature. Multiclass segregation is performed to categorize Membrane proteins into three distinct classes with optimum results as compared to the MemLoci predictor. Moreover, our approach delves one level further than the existing techniques, it can act as a supplement to those algorithms.

References

[1] J. D. Qiu, X. Y. Sun, J. H. Huang, and R. P. Liang, "Prediction of the types of membrane proteins based on discrete wavelet transform and support vector machines," (in eng), Protein J, vol. 29, no. 2, pp. 114-9, Feb 2010, doi: 10.1007/s10930-010-9230-z.

[2] T. W. Allen and F. Separovic, "Membrane protein structure and function," (in eng), Biochim Biophys Acta, vol. 1818, no. 2, p. 125, Feb 2012, doi: 10.1016/j.bbamem.2011.12.015.

[3] S. Galdiero, M. Galdiero, and C. Pedone, "beta-Barrel membrane bacterial proteins: structure, function, assembly and interaction with lipids," (in eng), Curr Protein Pept Sci, vol. 8, no. 1, pp. 63-82, Feb 2007, doi: 10.2174/138920307779941541

[4] A. Elofsson and G. von Heijne, "Membrane protein structure: prediction versus reality," (in eng), Annu Rev Biochem, vol. 76, pp. 125-40, 2007, doi: 10.1146/annurev.biochem.76.052705.163539.

[5] C. Savojardo, P. L. Martelli, P. Fariselli, G. Profiti, and R. Casadio, "BUSCA: an integrative webserver to predict subcellular localization of proteins," (in eng), Nucleic Acids Res, vol. 46, no. W1, pp. W459-w466, Jul 2 2018, doi: 10.1093/nar/gky320.

[6] A. Pierleoni, P. L. Martelli, and R. Casadio, "MemLoci: predicting subcellular localization of membrane proteins in eukaryotes," (in eng), Bioinformatics, vol. 27, no. 9, pp. 1224- 30, May 1 2011, doi: 10.1093/bioinformatics/btr108.

[7] P. Du, Y. Tian, and Y. Yan, "Subcellular localization prediction for human internal and organelle membrane proteins with projected gene ontology scores," (in eng), J Theor Biol, vol. 313, pp. 61-7, Nov 21 2012, doi: 10.1016/j.jtbi.2012.08.016.

[8] K. C. Chou and Y. D. Cai, "Using GO-PseAA predictor to identify membrane proteins and their types," (in eng), Biochem Biophys Res Commun, vol. 327, no. 3, pp. 845-7, Feb 18 2005, doi: 10.1016/j.bbrc.2004.12.069.

[9] S. Wan, M. W. Mak, and S. Y. Kung, "Mem-ADSVM: A two-layer multi-label predictor for identifying multi-functional types of mem-brane proteins," (in eng), J Theor Biol, vol. 398, pp. 32-42, Jun 7 2016, doi: 10.1016/j.jtbi.2016.03.013.

[10] M. Arif, M. Hayat, and Z. Jan, "iMem-2LSAAC: A two-level model for discrimination of membrane proteins and their types by extend-ing the notion of SAAC into chou's pseudo amino acid composition," (in eng), J Theor Biol, vol. 442, pp. 11-21, Apr 7 2018, doi: 10.1016/j.jtbi.2018.01.008.

[11] K. C. Chou, "Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes," (in eng), Bioinformatics, vol. 21, no. 1, pp. 10-9, Jan 1 2005, doi: 10.1093/bioinformatics/bth466.

[12] A. H. Butt, N. Rasool, and Y. D. Khan, "Predicting membrane proteins and their types by extracting various

sequence features into Chou's general PseAAC," (in eng), Mol Biol Rep, vol. 45, no. 6, pp. 2295-2306, Dec 2018, doi: 10.1007/s11033-018-4391-5.

[13] X. B. Zhou, C. Chen, Z. C. Li, and X. Y. Zou, "Using Chou's amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes," (in eng), J Theor Biol, vol. 248, no. 3, pp. 546-51, Oct 7 2007, doi: 10.1016/j.jtbi.2007.06.001.

[14] K. Ahmad, M. Waris, and M. Hayat, "Prediction of Protein Submitochondrial Locations by Incorporating Dipeptide Composition into Chou's General Pseudo Amino Acid Composition," (in eng), J Membr Biol, vol. 249, no. 3, pp. 293-304, Jun 2016, doi: 10.1007/s00232-015-9868-8.

[15] L. Zou, Z. Wang, and Y. Wang, "Prediction of outer membrane proteins using support vector machine with combined features," (in chi), Sheng Wu Gong Cheng Xue Bao, vol. 24, no. 4, pp. 651-8, Apr 2008, doi: 10.1016/s1872-2075(08)60034-5.

[16] A. Pierleoni, V. Indio, C. Savojardo, P. Fariselli, P. L. Martelli, and R. Casadio, "MemPype: a pipeline for the annotation of eukaryotic membrane proteins," (in eng), Nucleic Acids Res, vol. 39, no. Web Server issue, pp. W375-80, Jul 2011, doi: 10.1093/nar/gkr282

[17] P. Du, Y. Tian, and Y. Yan, "Subcellular localization prediction for human internal and organelle membrane proteins with projected gene ontology scores," (in eng), J Theor Biol, vol. 313, pp. 61-7, Nov 21 2012, doi: 10.1016/j.jtbi.2012.08.016.

[18] D. Yu, X. Wu, H. Shen, J. Yang, Z. Tang, and Y. Qi, "Enhancing membrane protein subcellular localization prediction by parallel fu-sion of multi-view features," (in eng), IEEE Trans Nanobioscience, vol. 11, no. 4, pp. 375-85, Dec 2012, doi: 10.1109/tnb.2012.2208473.

[19] S. Mei, "SVM ensemble based transfer learning for large- scale membrane proteins discrimination," (in eng), J Theor Biol, vol. 340, pp. 105-10, Jan 7 2014, doi: 10.1016/j.jtbi.2013.09.007.

[20] M. Kaleel, L. Ellinger, C. Lalor, G. Pollastri, and C. Mooney, "SCLpred-MEM: Subcellular localization prediction of membrane proteins by deep N-to-1 convolutional neural networks," (in eng), Proteins, vol. 89, no. 10, pp. 1233-1239, Oct 2021, doi: 10.1002/prot.26144.

[21] K. Shigene et al., "Translation of Cellular Protein Localization Using Convolutional Networks," (in eng), Front Cell Dev Biol, vol. 9, p. 635231, 2021, doi: 10.3389/fcell.2021.635231.

[22] G. Pan, C. Sun, Z. Liao, and J. Tang, "Machine and Deep Learning for Prediction of Subcellular Localization," (in eng), Methods Mol Biol, vol. 2361, pp. 249-261, 2021, doi: 10.1007/978-1- 0716-1641-3_15.

[23] K.-C. Chou, "Prediction of protein cellular attributes using pseudo-amino acid composition," Proteins: Structure, Function, and Genetics, vol. 44, no. 1, pp. 60–60, 2001.

[24] K.-C. Chou, "Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes," Bioinformatics, vol. 21, no. 1, pp. 10–19, 2004.

[25] K.-C. Chou and Y.-D. Cai, "Prediction of membrane protein types by incorporating amphipathic effects," Journal of Chemical Information and Modeling, vol. 45, no. 2, pp. 407–413, 2005.

[26] M. Vihinen, "How to evaluate performance of prediction methods? Measures and their interpretation in variation effect analysis", BMC Genomics, vol. 13, no. 4, p. S2, 2012. Available: 10.1186/1471-2164-13-s4-s