Predict the Membrane Proteins at Various Subcellular Locations

M.Devipriyanga¹, E.Sivasankari M.E²
PG Scholar², Assistant Professor²
Department of CSE
Government College of Engineering, Tirunelveli, Tamilnadu, India

Abstract:
To knowing the protein functions which was held in cell, some biological functions be carried out throughout the human body. Protein subcellular localization prediction is important tool to understanding biological functions. After protein sequences are translated into correct subcellular locations, according biological functions will be done. The different number of methods can be proposed over recent years for subcellular localization prediction. Subcellular localization prediction processes explore two different approaches such as sequence based and database based approaches. Sequence based approach is to extracting features from amino acid sequences and based upon this features subcellular prediction was done. Database based approach is to conduct data mining from existing databases. This paper focuses on sequence-based approach for protein subcellular localization prediction using amino acid sequences. By using k-nearest neighbor (k-NN) method, we predict the subcellular localization by providing amino acid sequences as the input. The k-nearest neighbor method makes predictions based upon a weighted Euclidean distance method. This method achieves 90.9% accuracy for prediction into nine subcellular locations. Comparisons show that, k-nearest neighbor achieves better prediction accuracy than weighted Euclidean distance method.

Keywords: prediction, subcellular localization, k-nearest neighbor, function, weighted Euclidean distance.

I. INTRODUCTION
Membranes are crucial to the life of a cell. A cell is covered by the plasma membrane (cell envelope) which defines the cell boundaries and activities taken between cells. Proteins need to be at right spatio temporal content within a cell to properly exert their biological functions. The knowledge of a protein's subcellular can help to infer the functions of the protein. Protein localization may provide valuable information in target identification process for drug discovery. Even if the function of a protein is known, it is equally to find where and in what kind of environment the protein performs its function because one of the fundamental goals in cell biology and proteomics is to identify the function of proteins in the context of compartments that organize them in cellular environment. The experimental determination of protein subcellular localization is mostly accomplished by the following three approaches [1]. Cell fractionation, electron microscopy and fluorescence microscopy. These approaches are time-consuming, costly and might bear some sort of subjectivity [2]. Nowadays, the number of protein sequences in public data bank is increased rapidly. So it would be both time-consuming and costly for solving the problem of subcellular localization prediction. The raw sequences of proteins will be available in public databases such as SWISSPROT, PDB, NCBI, PIR, etc. The large amount of methods could be established for protein subcellular localization process. Some of those methods makes prediction based on the N-terminal signal sequences. The main disadvantage of those methods can have the big limitation such as length of the amino acid sequences can affect the prediction performance highly [3-5]. To overcome this disadvantage, newly proposed method for subcellular localization prediction is residue composition method which will take the input of amino acid sequences. Some others used artificial neural network approach to overcome this above specified disadvantages [6]. We focused the localization prediction into nine locations: chloroplast, endoplasmic reticulum, Golgi apparatus, Lysosome, Mitochondria, Nucleus, Peroxisome, Plasma membrane, Vacuole. Based upon the membrane protein contact will be contact with the lipid bilayer proteins are involved in different biological functions. In this paper we will provide K-nearest neighbor (K-NN) approach is used to predict protein subcellular localization prediction. This approach will take the input from weighted Euclidean distance that can be calculated from amino acid residue composition values. Weighted Euclidean distance method will mainly depend upon standard Euclidean distance method and it achieves higher accuracy in prediction process. K-nearest neighbor approach achieves better prediction accuracy compared to weighted Euclidean distance method.

II. MATERIALS AND METHODS
The protein contains three dimensional structural in nature. The primary structure of a protein contains linear amino acid sequences. Every amino acid sequence consist of a particular amount of amino acid residue combinations.

A. Protein Sequence Collection
Amino acid sequences of each protein were obtained from PSORT db. [7] (http://db.psort.org/php/gram_stain=negative &experimental_sc=LIST). Some protein sequences were located into multiple locations, that proteins sequences will be eliminated. Totally, 4588 proteins were collected from nine different subcellular locations: Chloroplast (562), Endoplasmic reticulum (638), Golgi apparatus (394), Lysosome (584), Mitochondria (397), Nucleus (205), Peroxisome (623), Plasma...
membrane (864), Vacuole (321)

**B. Composition value of residues in amino acid sequences**

The number of amino acid residues presented in the whole given protein sequence will be, 

$$x_i = n_i / \sum_j n_j$$

Where \(n_i\) and \(n_j\) are number of residues presented in subcellular location types \(i\) and \(j\). Based upon that concept, the average residue composition of particular subcellular location will be given by,

$$\bar{x}_i = n_i / \sum_j n_j$$

Where \(n_i\) and \(n_j\) are total number of residues presented in subcellular location types \(i\) and \(j\). The average residue composition value for nine subcellular locations will be calculated separately.

**C. Weighted Euclidean Distance**

All proteins presented in the training set, the average residue composition value will be calculated separately for all subcellular locations. Based this average value the training proteins are located into correct subcellular location. Then for a given test protein, distance between subcellular location and that test protein was calculated by,

$$D = \sqrt{\sum_i (x_{i,\text{test}} - \bar{x}_i)^2 / \bar{x}_i}$$

Where \(x_{i,\text{test}}\) will be composition of residues present in the given test protein and \(\bar{x}_i\) is residue composition of the taken subcellular location. Here the important notice will be that \(\sqrt{\sum_i (x_{i,\text{test}} - \bar{x}_i)^2}\) is referred to be standard Euclidean distance. Additionally in this calculation we use the factor \(1/\bar{x}_i\).So the distance will often referred to as weighted Euclidean distance. This distance measure will be calculated for all subcellular locations. From this calculation, the shortest value will be selected. The test protein was then assigned to a location to which its distance is the shortest. Comparing to least standard distance, weighted Euclidean distance will give more prediction accuracy.

**D. K-Nearest Neighbor Algorithm**

Here the distances between the test protein and train protein will be calculated by,

$$D = \frac{1}{k} (D_{\text{chp}-1} + D_{\text{chp}-2} + \cdots + D_{\text{chp}-k})$$

Where \(x_{i,\text{train}}\) will be residue composition of taken train protein and \(x_{i,\text{test}}\) will be residue be residue composition of taken test protein. For each test protein, its weighted Euclidean distances to every protein in the training set were calculated. Then for each subcellular location, K shortest distances were chosen. For example, let \(D_{\text{chp}-1}, D_{\text{chp}-2}, \ldots, D_{\text{chp}-k}\) be the \(K\) shortest distances between the test protein and proteins that locate at the chloroplast subcellular location. Then the distance between the test protein and the location of the outer membrane was given by,
E. Five-fold Cross-validation method
This method was used to determine previously established methods. The entire dataset will be divided into nine subsets. The identity between any proteins from two different subset was less than 25%. In this process eight subsets were used as a training set and one remaining set were testing set. This process will be continued over each subset will be used as testing set. The prediction accuracy achieved in this method was 79%.

F. Performance analysis
The performance analysis for this subcellular location process is fully based on the prediction accuracy. The accuracy metric for particular subcellular location will be determine by,

\[
Accuracy_i = \frac{TP_i}{N_i}
\]

Where,

- \(TP_i\) = numbers of true positives (correctly located amino acid residues)
- \(N_i\) = numbers of total proteins for location \(i\).

The total accuracy for all subcellular locations are calculated using the below specified formula.

\[
Accuracy = \sum_{i=1}^{10} \frac{TP_i}{\sum_{i=1}^{10} N_i}
\]

III. RESULTS AND DISCUSSIONS
A. Prediction using weighted Euclidean distance
First we calculated least weighted Euclidean distance for all nine subcellular locations. For each entered test protein, distance between the subcellular location and test protein will be calculated. Based upon the shortest distance the entered test protein will be assigned to a correct subcellular locations. Using five-fold cross-validation method least weight Euclidean distance will be achieved the accuracy of 71%. In comparison this process will repeated over least standard Euclidean distance method. This follows the same procedure but least weighted Euclidean distance will be replaced by least standard Euclidean distance. The comparison shows that the overall accuracy will be achieved by 66.1% This comparison shows that weighted Euclidean distance method will achieve the overall prediction accuracy than compared to the standard Euclidean distance method. This will be illustrated by the Table 1.

B. Prediction using K-NN Algorithm
We then predict subcellular locations of proteins using K-nearest neighbor method. For each entered test protein the distances between subcellular location and test protein will be calculated using weighted Euclidean distance method. This process will be repeated over all nine subcellular locations. Then among this distance measures the shortest distance will be taken. Based upon this shortest distance the given test protein will be assigned to a correct subcellular location. This method achieves the prediction accuracy will be 75.9%. So this algorithm increases overall accuracy than weighted Euclidean distance method.

C. Performance Analysis
The performance metric to analyze those subcellular location prediction will be accuracy. The prediction accuracy will be
calculated using above specified formula. This analysis will be illustrated using various methods.

**LEAST EUCLIDIAN DISTANCE METHOD**
For each protein, its Euclidian Distances to every protein location are calculated. The test protein is then assigned to the location to which its Euclidian distance is the shortest. Based on the accuracy calculation method this approach achieves an overall accuracy of 64.4%.

**LWED (LEAST WEIGHTED EUCLIDIAN DISTANCE) METHOD**
For each protein, its Weighted Euclidian Distances to every protein location are calculated. The test protein is then assigned to the location to which its weighted Euclidian distance is the shortest. Based on the accuracy calculation method this approach achieves an overall accuracy of 66.1%.

**K-NN ALGORITHM**
K-NN method to predict protein subcellular localization. For each test protein, its distance to a location is calculated based on the weighted Euclidian distance. Its distances to each of the ten locations are calculated. Then the protein is assigned to the location to which the distance is the least. Using this approach, the overall accuracy is improved to 73.3%.

The below table shows that the overall accuracies of the above specified three methods are obtained for the ten subcellular locations.

<table>
<thead>
<tr>
<th>METHODS</th>
<th>OVERALL ACCURACY</th>
</tr>
</thead>
<tbody>
<tr>
<td>Least Euclidian distance</td>
<td>64.4%</td>
</tr>
<tr>
<td>Least Weighted Euclidian distance</td>
<td>66.1%</td>
</tr>
<tr>
<td>K-NN method</td>
<td>73.3%</td>
</tr>
</tbody>
</table>

The following chart illustrates that compare the accuracy of three methods such as Least Euclidian Distance, Least Weighted Euclidian Distance and K-NN algorithm. From this diagram we can easily justify that K-NN method archives better accuracy than comparing other methods.

### IV. CONCLUSION AND FUTURE ENHANCEMENT

In this paper, we present a K-NN method for predicting subcellular location of proteins. We initialize this process by weighted Euclidean distance method and gradually improve the prediction performance using K-nearest neighbor algorithm and information collected from residue composition method. The K-NN approach achieves the prediction performance by 73.3%. K-NN method gain more prediction accuracy than weighted Euclidean distance method. The proposed K-NN approach makes subcellular location prediction based on weighted Euclidean distance. Compared with standard Euclidean distance method, weighted Euclidean distance method will achieve the more prediction accuracy. This can establish a relationship between a test protein and a group. The future work includes extending weighted Euclidean distance which include n-peptides and dividing amino acid groups and use reduce alphabet to encode amino acids. And also extend K-NN method which includes other information, e.g. signal peptides.

### V. REFERENCES

[1]. Efendin Nasibov, Cagin Kandemir-Cavas(2008),” Protein subcellular location prediction using optimally weighted fuzzy k-NN algorithm”, ELSEVIER.


[3]. Haifeng Fan and Haixing Wang(2011),” Predicting Protein Subcellular Location by AdaBoost.M1 Algorithm”, IEEE.

[4]. Jing Hu and Changhui Yan(2007),” Predicting Protein Subcellular Localizations Using Weighted Euclidian Distance”, IEEE.

[5]. Kuo-Chen Chou and David W. Elrod (1999), “Prediction of Membrane Protein Types and Subcellular Locations”, WILEY-LISS, INC.


[7]. Kuo-Chen Chou1 & Hong-Bin Shen(2008),” Cell-PLoc: a package of Web servers for predicting subcellular localization of proteins in various organisms”, ELSEVIER.

[8]. Minghui Wang and Ao Li(2005), “Improving prediction of protein subcellular localization using evolutionary information and sequence-order information”, IEEE.

[9]. Tien-ho Lin and Robert F(2011),” Discriminative Motif Finding for Predicting Protein Subcellular Localization”, IEEE.


[11].Wei-Zhong Lin and Xuan Xiao(2008),” Using Grey Model GM(2,1) and Pseudo Amino Acid Composition to Predict Protein Subcellular Location”, IEEE.
