Iranian J Biotech. 2018 April;16(2):e1933 **DOI:10.21859/ijb.1933**

Research Article

Prediction of Protein Sub-Mitochondria Locations Using Protein Interaction Networks

Adele Sadat Haghighat Hoseini [1], Mitra Mirzarezaee [1, 2*]

[1] Department of Computer Engineering, Science and Research branch, Islamic Azad University, Tehran, Iran
[2] School of Biological Science, Institute for Research in Fundamental Sciences (IPM), Tehran, Iran

[*] *Corresponding author*: Mitra Mirzarezaee, Department of Computer Engineering, Science and Research branch, Islamic Azad University, 1477893855, Tehran, Iran, Tel: +98 21 44865179, E-mail: mirzarezaee@srbiau.ac.ir

Received: 29 May. 2017; **Revised**: 11 Jan. 2018; **Accepted**: 13 Jan. 2018; **Published online:** 11 Aug. 2018

Background: Prediction of the protein localization is among the most important issues in the bioinformatics that is used for the prediction of the proteins in the cells and organelles such as mitochondria. In this study, several machine learning algorithms are applied for the prediction of the intracellular protein locations. These algorithms use the features extracted from protein sequences. In contrast, protein interactions have been less investigated.
Objectives: As protein interactions usually occur in the same or adjacent places, using this feature to find the location would be efficient and impressive. This study did not aim at increasing the total accuracy of the conducted research. The study has focused on the features of the proteins' interaction and their employment which lead to a higher accuracy.
Materials and Methods: In this study, we have examined the protein interaction network as one of the features for prediction of the protein localization and its effects on the prediction results. In this regards, we have gathered some of the most common features including Amino Acid Composition, Dipeptide Compositions, Pseudo Amino Acid Compositions (PseAAC), Position Specific Scoring Matrix (PSSM), Functional Domain, Gene Ontology information, and the Pair-wise sequence alignment. The results of the classification are compared to the ones using protein interactions. For achieving this goal different machine learning algorithms were tested.
Results: The best-obtained results of using single feature set obtained using SVM classifier for PseAAC feature. The accuracy of combining all features with PPI data, using the Decision Tree and Random Forest classifiers, was 82.49% and 83.35%, respectively. In another experiment, using just protein interaction data with the different cutting points resulted in obtaining an accuracy of 93.035% for the protein location prediction.
Conclusion: In total, it was shown that protein(s) interaction has a significant impact on the prediction of the mitochondrial proteins' location. This feature can separately distinguish the locations well. Using this feature the accuracy of the results is raised up to 5%.

Keywords: Machine learning; Mitochondria; Protein localization; Protein-Protein Interaction (PPI)

1. Background

As a subcellular organelle in the eukaryotic cells, mitochondrion plays an important role in the process of energy generation and energy metabolism (1).

It was proved that mitochondria contribute to many complex biological processes like programmed cell death and ionic homeostasis. Many diseases are related to the incorrect mitochondria function; thus the function of proteins in the mitochondria as a vital organelle of the cells is extremely important (1).

A mitochondrion, in general, can be divided into four distinct parts; i.e. the outer membrane, inter-membrane space, inner membrane, and the matrix. Proteins in each of these compartments have their own biological role. Understanding protein functions are experimentally very time consuming and costly (2). Therefore, some computational systems were developed for the protein subcellular location prediction (3-15).

Copyright © 2017 The Author(s); Published by National Institute of Genetic Engineering and Biotechnology. This is an open access article, distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/) which permits others to copy and redistribute material just in noncommercial usages, provided the original work is properly cited.

Protein-mitochondrial location prediction has attracted much of interest among scholars regarding mitochondrial bioinformatics. In this regard, Du and Li (1) have presented a method which was based on an extended version of pseudo-amino acid composition for predicting protein localization within the mitochondria called SubMito. They used leave-one-out classification method and the accuracy of the results was found 85.5% for the inner membrane, 94.5% for the matrix, and 51.2% for the outer membrane. The overall accuracy achieved with their method was 85.2 %. Afterwards, GP-Locby (16) presented a genetic programming approach. They used Du and Li dataset in their research. The overall prediction of the accuracy obtained by GP-Loc was 89%. Zeng and his colleges (17) prepared a dataset including 399 mitochondria proteins. Their method was a sequence based algorithm combined with the augmented Chou's pseudo amino acid composition (Chou'sPseAA) based on auto-covariance (AC). Their total obtained accuracy was 89.7%. Zakeri and his colleges (18) predicted the protein sub-mitochondrial locations based on data fusion of the various features extracted from protein sequences. They increased the overall accuracy to 94.7 %. Shi and his colleges (15) proposed a strategy of the discrete wavelet transform and obtained 93.38% for predicting sub-mitochondrial locations. Mei (19) used GO information and proposed a multi-kernel transfer learning model for the protein sub-mitochondrial localization (MK-TLM). Fan and Li (20) made another dataset which included 1,105 mitochondrial proteins with sequence identity less than 40%. They proposed a method by combining the amino acid composition, dipeptide composition, reduced physicochemical properties, gene ontology, evolutionary information, and pseudo-average chemical shift. The overall prediction accuracy was 93.57%. Lin and his colleges (2) constructed a dataset with sequence identity ≥25%. This dataset included 495 mitochondrial proteins. They used Support Vector Machine to predict the sub-mitochondrial locations by using over-represented tetra-peptides selected by the binomial distribution. Their overall obtained accuracy for this dataset was 91.1%. Ahmad Khurshid and his colleges (21) used several classification learners including K-Nearest Neighbor, Probabilistic Neural Network, and Support Vector Machine (SVM). Among the various classification algorithms, SVM achieved the highest accuracy which was 95.20 % accuracy on dataset SML3-317 and 95.11 % on dataset SML3-983. Also, another research has been done to predict protein localization. The study has used the protein interactions (PPI) (22)and has achieved acceptable results (23).

2. Objectives

In this study, we have focused on the protein interaction network as a new feature. At first, in this approach, the most common features of the protein sequences are extracted and the accuracy of the protein sub-mitochondrial prediction locations is calculated using several classification methods including Support Vector Machines, K Nearest Neighbors, Naive Bayes, Decision Tree, and Random Forest, respectively. Then, we have investigated the effect of adding protein interaction features and compared the results.

3. Materials and Methods

3.1. Dataset

Four datasets have been provided in order to predict protein sub-mitochondrial locations. The dataset M317 presented by Du and Li (1) contains 131 inner membrane proteins, 145 matrix proteins, and 41 outer membrane proteins. The second dataset, M399 created by Zeng and *et al* (17), contains 171 inner membrane proteins, 166 matrix proteins, and 62 outer membrane proteins. The third dataset M1105 by Fan and Li (20) has 589 inner membrane proteins, 280 matrix proteins, and 236 outer membrane proteins. The last dataset M495 derived by Lin. *et al*., (2) shows 254 inner membrane proteins, 132 matrix proteins, and 109 outer membrane proteins.

Since the effect of protein interactions on the protein locations is investigated here, the target dataset in this study is the proteins that belong to a particular organism. In this regards, the concept of interactions should be meaningful. Thus, a new dataset has been prepared.

In this research, the mitochondrial proteins' dataset has been extracted from Swiss- Prot. The total number of proteins are 547,085 including Human, Mouse, Bovine, *S. cerevisiae* and proteins from other organisms. Within this study, we merely have investigated the human proteins. To achieve a reliable and high-quality dataset, the following steps have been followed:

1- Proteins in more than one place in a mitochondria are removed.
2- Protein sequences are aligned in order to measure the similarity of their sequences using BLAST tool (1).
3- The proteins with E-value less than 0.0001 are eliminated. (1). This has been done in order to eliminate the false positive as much as possible. As the E value is lowered, the sequences are more similar and their confidence for the homology is increased.

Applying the aforementioned three steps, 435 proteins were obtained, from which 199 were inner membrane proteins (IM), 26 were intermembrane space

proteins (IMS), 132 were matrix proteins (M), and 78 were outer membrane proteins (OM). This obtained dataset is called M435. A list of all the proteins involved in this study is available in the **Supplementary File 1**.

3.2. Feature Vectors

3.2.1. Amino Acid Composition

The amino acid composition is a fraction of each amino acid in a protein and can be calculated for any of the 20 natural amino acids by the following equation:

$$\text{Fraction of AAC}_i = \frac{\text{total number of amino acids of type } i}{\text{total number of amino acids in protein}} \quad (1)$$

Where, *i* is any natural amino acid. The ACC extracted features for each protein is represented in a vector with 20 elements (18, 24, 25).

3.2.2. N-Peptide Composition

The N-peptide composition is the number of repeated occurrences of the amino acids in a consecutive protein sequence (18). When n is increased, the n-peptide compositions will maintain more general information from the sequences. If n =1, the n-peptide composition is the same as AAC, If n =2, the n-peptide composition called D-peptide composition (DP). For each protein, the DP feature has a 400 element feature vector. In most biological applications, n is 2 for efficient computing (26). D-peptide composition for each protein is calculated as:

$$\text{Fraction of } dipi = \frac{\text{total number of dip } i}{\text{total number of all dipeptides}} \quad (2)$$

For the prediction of protein in a cell, many researchers used this feature (18, 26, 27).

3.2.3. Pseudo Amino Acid Composition (PseAAC)

This feature is used in many previous researches (18, 28). The idea behind this feature was first introduced by Chou (2001) in order to avoid loss of sequence information. This feature vector is calculated as follows:

$$\mathrm{P} = \left[p_1, p_2, ..., p_{20}, p_{20+1}, ..., p_{20+\lambda}\right]^{\mathrm{T}} \quad (3)$$
$$(\lambda < L)$$

Where the first 20 vector elements are AACs and other elements are obtained from physical properties of the amino acids. In this study, hydrophobicity, hydrophilicity, and the side chain mass of the amino acids were used. To calculate PseAAC, the Web-serverat (http://www.csbio.sjtu.edu.cn/bioinf/PseAAC/) is used. Where λ is an integer parameter value. The different values of λ can create different features. In this research, the optimum value for λ achieved from 1 to 20 and 20 different feature vectors for representing a protein sample are created. The dimension of each feature vector is related to the value of λ and they can be acquired from Eq. 3 (18).

3.2.4. Functional Domain Composition

Proteins contain multiple domains and models. The function of a protein usually depends on the protein location. Therefore, functional domain composition (FD) can be used in sub-mitochondrial location prediction. For this purpose, InterPro dataset is used (29).

$$\mathrm{P} = \left[d_1 d_2 ... d_j ... d_{904}\right]^T$$
$$d_i = \begin{cases} 1 & \text{when a hit is found for P} \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

InterPro dataset contains a number of proteins with a known functional domain. The feature vector for our protein has 907 dimensions; for every protein in our dataset, if InterPro was very similar to a sequence section in the protein sequence and hit with them, we assigned 1 otherwise assigned 0 (18).

3.2.5. Position Specific Scoring Matrix

Position Specific Scoring Matrix (PSSM) is a Position-Specific Scoring Matrix which is made of the processes carried out in PSI-BLAST. In PSSM, each amino acid in the sequence is mapped to 20 integers. Each number indicates the amino acid substitution at that location of the sequence, with each of the 20 amino acids found in nature are in the process of evolution (30). PSI-BLAST is used to compare different sequences for finding the similar sequences and discovery of their evolutionary relationships (31, 32).

3.2.6. Smith-Waterman (Pair-SW) Pairwise Sequence

Pair-SW is the process of searching of the two sequences in order to find maximal levels of identical regions for the purpose of assessing the degree of similarity that may show functional, structural, and evolutionary relationships between the two biological sequences (33). Two different strategies are used for Pairwise sequence alignment namely: Global alignment and Local alignment. Global alignments are the most useful approach when the sequences in

the query are similar and are of equal size. A general global alignment technique is the Needleman-Wunsch algorithm (33).Local alignments are more beneficial for the dissimilar sequences with the different lengths. The Smith-Waterman algorithm is a general local alignment method. So according to our data set that contains different sequences in size, we used the Smith–Waterman algorithm.

3.2.7. Gene Ontology Information

The Gene Ontology (GO) is a major bioinformatics initiative for joining all the representations of the genes and their products and unifying them in one dataset. GO dataset contains: (a) cellular components referring to the place in cell where the gene is active; (b) molecular function; the biochemical activity of a gene product, and (c) biological process; the proceedings or collection of the molecular events with a specific beginning and the end (34). In this work, we just used the biological process and molecular function of a gene. Therefore, by mapping of the InterPro (http://www.ebi.ac.uk/interpro) entries to GO, we can get a list of data, and the mapping of each InterPro entrance to a GO number. To achieve this, a vector is formed. For every protein in our dataset, its GO related information is tagged.

3.2.8. Protein-Protein Interaction

Protein-protein interactions (PPIs) refer to special physical contacts between two or more proteins. Proteins which are in the interaction should be in the same or in adjacent locations. Therefore, it is expected that the prediction of the subcellular localization using protein interactions should be improved. In this study, the String dataset has been used to obtain protein interactions (35). The STRING database is a collection of various methods employed for identifying interactions. It enabled us to select interactions using various methods with different degrees of the accuracy and thus expands our choices for considering more interactions.

3.2.9. Combination of all Features

A combination of all input features including amino acid composition (AAC), D-peptide composition, Pseudo amino acid composition (PseAAC), Functional domain composition, Gene Ontology, Pairwise sequence alignment, and protein-protein interactions are also examined.

3.3. Base Learners

The following classification methods are used in this study. All these methods are implemented in the R programming environment.

3.3.1. Support Vector Machines

Support Vector Machine (SVM; Vapnik 1995) is a supervised learning algorithm used for classification which finds the discrimination function with the largest distance between two classes. In this study, the RBF kernel function is used. An important issue is how to optimize SVM in case of selected parameters. Different values for the parameters of the RBF kernel function; C and γ are examined to achieve the best possible accuracy. For a multiclass SVM classification, we use the one-against-one approach proposed by Yu (26).

3.3.2. K Nearest Neighbors

K Nearest Neighbors (KNN) is one of the simplest non-linear classifier. This classifier assigns the sample to the class with the highest vote among its k-nearest neighbors (36). In this study, Euclidean distance is used and the value of K is tested within the range of 1 to 10.

3.3.3. Naive Bayes

Naive Bayes (NB) classifier is a supervised learning algorithm. NB classifier is extremely scalable. Bayesian reasoning method is based on probability in order to draw inferences about the probability distribution of the optimal decision. (37, 38).

3.3.4. Decision Tree

A Decision (DT) Tree is a flowchart structure and a non-parametric supervised learning method used for classification and regression. Each internal node reflects a “test” of feature, each branch reflects the test results, and each leaf node represents a class label and path from the root to the leaves, which in turn indicates the classification rules (37). Decision Tree learning algorithms are generally recursive processes. In each step, one branch is selected. The important step in the Decision Tree algorithm is how to select the branches. Different algorithms employ different metrics to measure such as Gini Impurity used by the CART (classification and regression tree) algorithm, Information Gain used by the ID3, C4.5, and C5.0 Tree-Generation algorithms and Variance Reduction are used by the CART (39). Within this research, CART algorithm is used.

3.3.5. Random Forest

Random Forest (RF) is an ensemble method which is developed using bagging approach (40). Its main difference from Bagging is that its feature selection is random. This method involves multiple Decision Trees. K parameters contribute to the randomness. When K is equal to the total number of the features, the Decision Tree method is conducted typically. When

K=1, one feature is randomly selected. The proposed K is the logarithm of the number of properties in some studies; however, this number is the square root of the number of features. In an RF classification algorithm, two parameters should be specified by the user: M is a random subset of features and T is the number of trees. In this method, the value of T is considered 500.

3.4. Evaluation Criteria

In this study, numerable features of the protein sequences are considered to predict mitochondrial protein mitochondrial locations. The performances of each classifier are evaluated by 10-fold cross-validation and using prediction accuracy (ACC), sensitivity and specificity of each location, and the overall prediction accuracy according to the following formulas:

$$ACC(i) = \frac{TP(i)}{TP(i) + FN(i)} \quad (5)$$

$$ACC_{overall} = \frac{1}{m}\sum_{k=1}^{4} TP(k) \quad (6)$$

Where, *TP(i)* is the number of correctly predicted protein sequences that belong to location *i* (true positive), *TN(i)* is the number of correctly predicted protein sequences that do not belong to location i (true negative), *FP(i)* is the number of under -predicted protein sequences (false positive)and *FN(i)* is the number of over- predicted protein sequences (false negative). *m* is the total number of protein sequences, and *k* is the number of sub-mitochondrial locations representing OM, IMS, IM, and M, respectively (18).

4. Results

Results are provided in two sections: the first section represents investigation regarding the obtained classification accuracy for each feature set and the second section focuses on the protein interactions and their effects on the CCR of the obtained results. Also, each classifier will be evaluated by the cross-validation test.

4.1. Part 1: Analysis of the Results Using Different Feature Sets

In this section, the best-obtained results through application different feature sets and the combination of all features using the five different classifiers are reported and discussed.

PseAAC feature with λ=15 is the best-obtained accuracy from SVM classifier. As a result, the three best-obtained accuracy were from GO, PSSM, and FD features with 71.195%, 66.561%, and 66.169% as shown in **Table 1**.

For KNN classifier, PseAAC with a λ=15 is the best-obtained result and feature of accuracies of 69.62%, followed by the best accuracy for Pair-SW and PSSM from FD and features as shown in **Table 2**.

Results for Naïve Bayes (NB) classification indicate

Table 1. Best results using support vector machine for different feature sets.

Feature Set	**PseAAC (λ=15) with (C= 1, γ=0.01) (%)**	**GO (C = 6.5 and γ = 0.01) (%)**	**PSSM (C = 10 and γ = 0.01) (%)**	**FD (C = 100 and γ = 0.2) (%)**
Accuracy	76.49	71.20	66.56	66.17
IM	77.84	76.84	69.58	72.11
IMS	43	50	48.81	75
M	77.86	90.32	76.55	85.86
OM	93	73.53	59.71	60
Sensitivity				
IM	68.68	73.68	67.16	84.21
IMS	0	0	0	50
M	87.31	100	69.23	84.62
OM	93.15	50	40	20
Specificity				
IM	88.21	80	76	60
IMS	80	100	97.62	100
M	63.42	80.65	83.87	87.1
OM	80	97.06	79.41	100

Inner membrane proteins (IM), intermembrane space proteins (IMS), matrix proteins (M), outer membrane proteins (OM).

Table 2. Best results using KNN classifier for different feature sets.

Feature Set	PseAAC (λ=15) with (K=10) (%)	FD with (K=1) (%)	Pair-SW with (K=3) (%)	PSSM with (K=4) (%)
Accuracy	69.62	62.01	60.02	58.26
IM	70.21	68.74	70.31	59.05
IMS	48.63	75	40	50
M	70.33	81.39	80.67	63.03
OM	98	60	59.23	66.18
Sensitivity				
IM	68.42	89.47	70.33	52.11
IMS	0	50	0	0
M	69.23	69.23	30	70.54
OM	100	20	60	40
Specificity				
IM	80	48	70	80.83
IMS	100	100	99	100
M	77.42	93.55	80.33	64.52
OM	100	100	90.94	82.35

Inner membrane proteins (IM), intermembrane space proteins (IMS), matrix proteins (M), outer membrane proteins (OM).

Table 3. Best results using Naive Bayes for different feature sets.

Feature Set	PseAAC (λ=15) (%)	PseAAC (λ=2) (%)	PSSM(%)	PseAAC (λ=12) (%)
Accuracy	70.77	58.34	58.07	56.54
IM	71.84	75.58	58.42	74.21
IMS	49.51	48.81	46.43	48.81
M	73.92	81.64	72.33	76.18
OM	100	52.66	62.55	5.59
Sensitivity				
IM	55	63.16	36.84	68.42
IMS	0	0	0	0
M	71.43	92.31	76.92	84.62
OM	100	20	40	10
Specificity				
IM	82.61	88	80	80
IMS	95.24	97.62	92.86	97.62
M	72.41	70.79	67.74	67.74
OM	100	85.29	85.29	91.18

Inner membrane proteins (IM), intermembrane space proteins (IMS), matrix proteins (M), outer membrane proteins (OM).

the following order of the accuracy: PseAAC (λ=15), PseAAC (λ=2), PSSM, and PseAAC (λ=12) as presented in **Table 3**.

After testing all the selected classifiers with the common feature sets, the combination of all features was tested. The highest accuracy was obtained by combining all the features with PPI data using Decision Tree and Random Forest classifiers as found to be 82.49% and 83.35%, respectively. The best-obtained accuracy for this classification with PseAAC (λ=20), PseAAC (λ=15), and GO features are shown in **Table 4** and **5**.

Thus, the highest accuracy; without adding majority voting in protein interaction features at this point, was the combination of all the features with a PPI matrix in the Random Forest algorithm with 83.35% of the accuracy.

Table 4. Best results using Decision tree for different feature sets.

Feature Set	GO (%)	PseAAC (λ=15) (%)	PseAAC (λ=20) (%)	Mix of all feature with PPI (%)
Accuracy	66.32	73.45	57.86	82.49
IM	75.58	82.11	71.86	86.11
IMS	50	50	50	47.62
M	79.03	72.7	75.56	83.62
OM	58.53	100	60.59	100
Sensitivity				
IM	63.16	84.21	70.95	84.21
IMS	0	0	0	0
M	100	61.54	68.92	76.92
OM	20	100	20	100
Specificity				
IM	88	80	70	88
IMS	100	100	94.32	95.24
M	58.04	83.87	70.63	90.32
OM	97.06	100	89.13	100

Inner membrane proteins (IM), intermembrane space proteins (IMS), matrix proteins (M), outer membrane proteins (OM).

Table 5. Best results using Random forest for different feature sets.

Feature Set	GO (%)	PseAAC (λ=15) (%)	PseAAC (λ=20) (%)	Mix of all feature with PPI (%)
Accuracy	67.76	73.965	67.448	83.35
IM	69.47	79.78	73.28	90.74
IMS	50	50	50	50
M	76.55	75.12	68.62	91.32
OM	68.53	100	93.5	100
Sensitivity				
IM	78.95	90	83.5	89.47
IMS	0	0	0	0
M	69.23	57.14	50.64	90.31
OM	40	100	93.5	100
Specificity				
IM	60	69.57	64.32	92
IMS	100	100	92.15	100
M	83.87	93.1	86.62	92.33
OM	97.06	100	83.43	100

Inner membrane proteins (IM), intermembrane space proteins (IMS), matrix proteins (M), outer membrane proteins (OM).

4.2. Part 2: Protein-Protein Interaction Features for the Predicted Sub-Mitochondrial Protein Locations

4.2.1. The First Experiment: Majority Voting for PPI Data Set without Considering Scores

In the first experiment, the interaction weights are not taken into consideration; therefore, all interactions have the same values. As the location of each protein in training datasets is determined, the majority voting is used to label the test datasets. Since there might exist proteins with no interaction data, the proposed method was unable to predict the location of all proteins. From a set of 435 proteins, only 264 of these proteins have interactions. The result of this phase has the accuracy of 80.67%.

4.2.2. The Second Experiment: Majority Voting for PPI Data Set with Scores

Table 6. Number of available protein interactions based on different cutting points.

Range	Number of protein interactions
Interaction with a score greater than or equal to 200	251
Interaction with a score greater than or equal to 500	213
Interaction with a score greater than or equal to 700	201
Interaction with a score greater than or equal to 900	109

Table 7. Majority voting with different Cutting points.

Conditions	Results (%)
PPIs with score greater than or equal to 200	83.27
PPIs with score greater than or equal to 500	86.86
PPIs with score greater than or equal to 700	93.035
PPIs with score greater than or equal to 900	89.19

In the second experiment, scores of interactions are also considered. The interaction between the two proteins in the String dataset has a number between 0 and 1000. Four thresholds as cutting points of the interactions were taken into consideration: 200, 500, 700, and 900. It is obvious that a higher threshold will select just the interactions with a higher confidence, but the number of interactions decreases as well. **Table 6** lists proteins with four different cutting thresholds.

The results of applying majority voting to the protein interaction data with four different cutting points are shown in **Table 7**.

As shown in Table 7, increasing the cutting point for protein interactions can increase the accuracy of the results dramatically up to the 700 point; and the desired results will decrease afterward because the number of interactions will decrease when the cutting point is set at a very high rate. The best tested threshold for this purpose is 700. In comparison with the other tested approaches, it was shown that this method obtains favorable results. The summary of the results is shown in **Figure 1**. As it is shown in this Figure, the interaction of the protein with a score greater than or equal to 700 has given the highest degree of accuracy. It should be noted that the accuracy of the protein interactions is for a number of proteins and not for all the other proteins obtained in this project. So in the next experiment, we used a combination of all features that will be described in detail.

4.2.3. The Third Experiment: Combination of all Feature Sets with Protein Interactions

In the third experiment, the main question was whether adding protein interaction features will increase the classification accuracy or not. For increasing the classification accuracy, these experiment sets up were

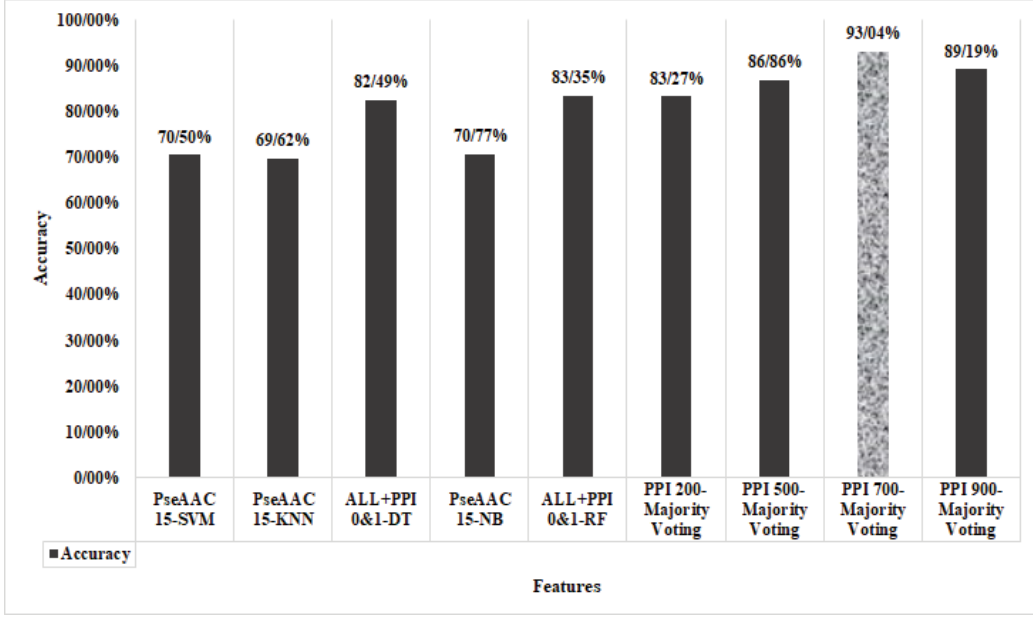

Figure 1. Comparison between best result of different classifiers with protein interaction in majority voting.

as follows: for the proteins with available protein interaction data, this feature was used and if such a data was not available the results of classification based on other features were used. In this experiment, different classifiers were also tested and the results for each classifier were reported separately in **Supplementary File 2**. As the results for SVM classifier show, the highest achieved accuracy was from PPI data with 700 as the cutting point and the PseAAC feature with λ = 15.

For the KNN classifier, the best results were from protein interactions with 700 as the cutting point and PseAAC feature with λ = 15. For the Decision Tree algorithm, the best-obtained results were for a combination of all features with the protein interactions of 700 as the cutting point. The best next result was for the same features with a cutting point of 500 on protein interactions.

In NB classifier, the highest accuracy was from protein interactions with 700 as the Cutting point and the PseAAC feature with λ = 15. Random Forest algorithm had the best result for this study by combining all features with PPI of 700 as the cutting point.

4.2.4. The Fourth Experiment

In the fourth experiment, our investigation was focused on the hypothesis that whether an increase in the number of proteins whose protein interaction is known can increase the classification accuracy or not.

In order to achieve this goal, it is supposed that the only existing proteins are those obtained from the PPI with the score of 700. This is because of the fact that the majority of voting with the accuracy of 93.035% was related to the protein interactions with the score of 700. As a result, it is supposed that we have only 201 proteins for all of which protein interactions exist. In the first step, all the datasets were combined so that datasets are placed in the same place and were provided with equal distribution. In the second step, the datasets were divided into 4 sections. Therefore, the combined datasets were divided into three groups with 50 proteins and one group with 51 proteins. In the third step, it was proposed to randomize and calculate with 50 proteins and their subsequent 50 interactions. This was done for every four groups and the accuracy was achieved; in the end, the average was calculated. In the fourth step, the second group was added to the previous group. Thus, more interactions were added, the majority voting was conducted, and the accuracy was achieved. In the final step, the third and the fourth groups were added; the fourth chosen group was included all the proteins and their interactions.

In this experiment, the protein interaction data with 700 as the cutting point was used, as this cutting point was the best results from the previous studies. For this purpose data was divided into four equal sections and the classification of the proteins based on their protein interaction data was done in four different steps. In each step, one more section of the data was added in the process of the decision making for the protein localization and the majority voting algorithm was used for the classification.

In this experiment, it was shown that if the protein interaction data increases, the result of the classification is also increased and the accuracy of the results will be improved. The results are shown in **Figure 2**.

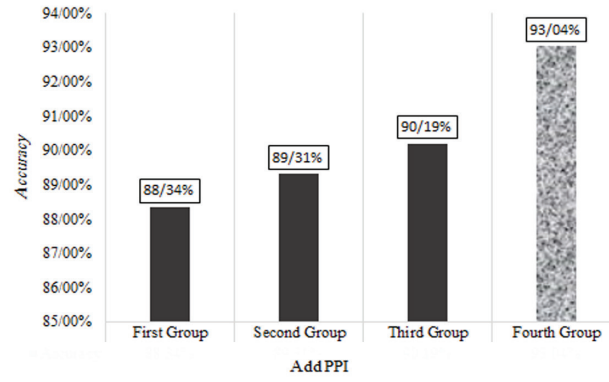

Figure 2. Increase in the number of PPI data and its effects on prediction of protein localization.

5. Discussion

The present study aimed at comparing the predictive method with and without using protein interactions in order to reveal that by taking the protein interaction into consideration, the predictive power will improve as well. As a result, the comparisons are merely made using the mentioned classifiers and the used databases. Four datasets M317(1), M399(17), M1105(20) and M495(2) have been provided to predict sub-mitochondrial locations. As mentioned before, in order to investigate the effect of protein interactions on the protein locations, a new dataset has been prepared that is called M435. Several methods are presented for the purpose of Machine Learning. These methods use protein sequences and amino acid assessment as their input and predictions performance of our classifiers were evaluated by calculating the prediction accuracy (ACC) (18).

In this study, different feature sets were tested to predict the location of the proteins in the mitochondria; PseAAC that used in SubMito (1) with λ=15 is the best-obtained accuracy from SVM, KNN and NB with

accuracies 76.49%, 69.62% and 70.77% . The highest accuracy was obtained by combining all the features with PPI data using Decision Tree and Random Forest classifiers with accuracies of 82.49% and 83.35% respectively. Random Forest classifier performs best in most cases (23). Then the effect of protein interaction data and its relation to the interactions were investigated. The interaction between the two proteins in the String dataset has a number between 0 and 1000. Four thresholds as cutting points of the interactions were taken into consideration that the best tested threshold for this purpose is 700. Random Forest algorithm had the best result for this study by combining all features with PPI of 700 as the cutting point resulted in obtaining an accuracy of 93.035%. The method aimed at showing how the previous accuracies, which have been achieved through a common method, will increase as we gain more information regarding protein interaction. As a result, as the number of the known interactions increases, the total accuracy will increase as well. Since protein interactions usually occur among the residing proteins in one location or adjacent locations, the use of this feature will be beneficial for solving the prediction issue of the protein locations.

The results show that interactions have a direct correlation with the location of the proteins. The proximity of the proteins which makes them interact is a factor in the identification of their locations. This feature is among the most important features in the prediction of the protein locations. Our experiments show that the prediction results improve when more protein interaction data is available. As well, it could be concluded that if the available data of the protein interactions increases, through the application of the new technologies, the results of this approach will be more promising.

6. Conclusion

Among the most important organelles of the eukaryote cells are mitochondria. Knowing the location of each protein in a mitochondrion is significant. One can predict the function of the proteins based on their locations. Therefore, the main purpose of this study was to predict the protein sub-mitochondrial locations. In the previous studies, different features from analyzing amino acids and protein sequences were studied. However, so far the interactions between proteins and their effects on the protein localization prediction were not studied. Therefore, in this study, we focused on this feature to make it known whether it could improve the results. Since protein interactions occur between the proteins in the same or adjacent locations, they may improve the accuracy of the results. For this purpose, we have used some of the well- known algorithms and used features like amino acid composition, D-peptide composition, pseudo amino acid composition, functional domain composition, position-specific scoring matrix, pairwise sequence alignment - Smith-Waterman, Gene ontology (GO) information, and then tested the effects of adding the protein interaction data as another feature. Different classification methods such as Support Vector Machines, K Nearest Neighbors, Naive Bayes, Decision Tree and Random Forest were tested. The best results were for Random Forest and the obtained accuracy was 88.96%. We have also tested the protein interaction data alone to investigate how it can predict the locations alone. The results were tested in different modes and it was shown that this feature can separately distinguish the locations well. Finally, the results of our experiments show that protein interaction data has a significant impact on the prediction of the protein localization in the mitochondria.

Acknowledgement

This research was supported in part by a grant from the Institute for Research in Fundamental Sciences (IPM), Tehran, Iran.

Supplementary Data

There were two supplementary files available in this study. The first supplementary file was attached to mention a list of all of the proteins involved in the present study. In the same way, the second supplementary file targets the results which was reported by different classifiers with various features.

References

1. Du P, Li Y. Prediction of protein submitochondria locations by hybridizing pseudo-amino acid composition with various physicochemical features of segmented sequence. *BMC Bioinformatics*. 2006;**7**(1):1-8. doi: 10.1007/s00726-011-1143-4
2. Lin H, Chen W, Yuan L-F, Li Z-Q, Ding H. Using Over-Represented Tetrapeptides to Predict Protein Submitochondria Locations. *Acta Bio Theor.* 2013;**61**(2):259-268. doi: 10.1007/s10441-013-9181-9
3. Du P, Cao S, Li Y. SubChlo: predicting protein subchloroplast locations with pseudo-amino acid composition and the evidence-theoretic K-nearest neighbor (ET-KNN) algorithm. *J Theor Biol.* 2009;**261**(2):330-335. doi: 10.1016/j.jtbi.2009.08.004
4. Du P, Li T, Wang X. Recent progress in predicting protein sub-subcellular locations. *Expert Rev Proteomics.* 2011;**8**(3):391-404. doi: 10.1586/epr.11.20
5. Huang W-L, Tung C-W, Ho S-W, Hwang S-F, Ho S-Y. ProLoc-GO: Utilizing informative Gene Ontology terms for sequence-based prediction of protein subcellular localization. *BMC Bioinformatics.* 2008;**9**(1):1-16. doi: 10.1186/1471-

2105-9-80
6. Huang W-L, Tung C-W, Huang H-L, Ho S-Y. Predicting protein subnuclear localization using GO-amino-acid composition features. *Biosystems.* 2009;**98**(2):73-79. doi: 10.1016/j.biosystems.2009.06.007
7. Huang W-L, Tung C-W, Huang H-L, Hwang S-F, Ho S-Y. ProLoc: Prediction of protein subnuclear localization using SVM with automatic selection from physicochemical composition features. *Biosystems.* 2007;**90**(2):573-581. doi: 10.1016/j.biosystems.2007.01.001
8. Jiang X, Wei R, Zhao Y, Zhang T. Using Chou's pseudo amino acid composition based on approximate entropy and an ensemble of AdaBoost classifiers to predict protein subnuclear location. *Amino Acids.* 2008;**34**(4):669-675. doi: 10.1007/s00726-008-0034-9
9. Lei Z, Dai Y. An SVM-based system for predicting protein subnuclear localizations. *BMC Bioinformatics.* 2005;**6**(1):1-8. doi: 10.1186/1471-2105-6-291
10. Lei Z, Dai Y. Assessing protein similarity with Gene Ontology and its use in subnuclear localization prediction. *BMC Bioinformatics.* 2006;**7**(1):1-10.doi: 10.1186/1471-2105-10-274
11. Li F-M, Li Q-Z. Using pseudo amino acid composition to predict protein subnuclear location with improved hybrid approach. *Amino Acids.* 2008;**34**(1):119-125. doi: 10.1007/s00726-007-0545-9
12. Mei S, Fei W. Amino acid classification based spectrum kernel fusion for protein subnuclear localization. *BMC Bioinformatics.* 2010;**11**(1):1-8. doi: 10.1186/1471-2105-11-S1-S17
13. Shen H-B, Chou K-C. Predicting protein subnuclear location with optimized evidence-theoretic K-nearest classifier and pseudo amino acid composition. *Biochem Biophys Res Commun*. 2005;**337**(3):752-726. doi: 10.1016/j.bbrc.2005.09.117
14. Shen H-B, Chou K-C. Nuc-PLoc: a new web-server for predicting protein subnuclear localization by fusing PseAA composition and PsePSSM. Protein Eng Des Sel. 2007 Nov;**20**(11):561-7. Epub 2007 Nov 10. doi: 10.1093/protein/gzm057
15. Shi S-P, Qiu J-D, Sun X-Y, Huang J-H, Huang S-Y, Suo S-B, et al. Identify submitochondria and subchloroplast locations with pseudo amino acid composition: Approach from the strategy of discrete wavelet transform feature extraction. (BBA) - *Molecular Cell Res.* 2011;**1813**(3):424-430. doi: 10.1016/j.bbamcr.2011.01.011
16. Nanni L, Lumini A. Genetic programming for creating Chou's pseudo amino acid based features for submitochondria localization. *Amino Acids.* 2008;**34**(4):653-660. doi: 10.1007/s00726-007-0018-1
17. Zeng Y-h, Guo Y-z, Xiao R-q, Yang L, Yu L-z, Li M-l. Using the augmented Chou's pseudo amino acid composition for predicting protein submitochondria locations based on auto covariance approach. *J Theor Biol.* 2009;**259**(2):366-372. doi: 10.1016/j.jtbi.2009.03.028
18. Zakeri P, Moshiri B, Sadeghi M. Prediction of protein submitochondria locations based on data fusion of various features of sequences. *J Theor Biol.* 2011;**269**(1):208-216. doi: doi.org/10.1016/j.jtbi.2010.10.026
19. Mei S. Multi-kernel transfer learning based on Chou's PseAAC formulation for protein submitochondria localization. *J Theor Biol.* 2012;**293**:121-30. doi: 10.1016/j.jtbi.2011.10.015
20. Fan G-L, Li Q-Z. Predicting protein submitochondria locations by combining different descriptors into the general form of Chou's pseudo amino acid composition. *Amino Acids*. 2012;**43**(2):545-555. doi: 10.1007/s00726-011-1143-4
21. Ahmad K, Waris M, Hayat M. Prediction of Protein Submitochondrial Locations by Incorporating Dipeptide Composition into Chou's General Pseudo Amino Acid Composition. *J Membr Biol.* 2016;**249**(3):293-304. doi: 10.1007/s00232-015-9868-8
22. Akbaripour-Elahabad M, Zahiri J, Rafeh R, Eslami M, Azari M. rpiCOOL: A tool for In Silico RNA–protein interaction detection using random forest. *J Theor Biol.* 2016;**402**(Supplement C):1-8.doi: 10.1016/j.jtbi.2016.04.025
23. Zahiri J, Mohammad-Noori M, Ebrahimpour R, Saadat S, Bozorgmehr JH, Goldberg T, et al. LocFuse: Human protein–protein interaction prediction via classifier fusion using protein localization information. *Genomics.* 2014;**104**(6, Part B):496-503. doi: 10.1016/j.ygeno.2014.10.006
24. Bhasin M, Raghava GPS. Classification of Nuclear Receptors Based on Amino Acid Composition and Dipeptide Composition. *J Bioll Chem*. 2004;**279**(22):23262-23266. doi:10.1074/jbc.M401932200
25. Bhasin M, Raghava GPS. ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. *Nucl Acids Res.* 2004;**32**(suppl 2):W414-W419.doi: 10.1093/nar/gkh350
26. Yu C-S, Lin C-J, Hwang J-K. Predicting subcellular localization of proteins for Gram-negative bacteria by Support Vector Machines based on n-peptide compositions. *Protein Sci.* 2004;**13**(5):1402-1406. doi: 10.1110/ps.03479604
27. Khan A, Majid A, Hayat M. CE-PLoc: An ensemble classifier for predicting protein subcellular locations by fusing different modes of pseudo amino acid composition. Comput Biol Chem. 2011;**35**(4):218-229. doi: 10.1016/j.compbiolchem.2011.05.003
28. Chou KC. Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins*. 2001;**43**(3):246-255. doi: 10.1002/prot.1035
29. Shen H-B, Chou K-C. Predicting protein fold pattern with functional domain and sequential evolution information. *J Theor Biol*. 2009;**256**(3):441-446. doi: 10.1002/prot.1035
30. Meshkin A, Ghafuri H. Prediction of relative solvent accessibility by support vector regression and best-first method. *Excli J*. 2010;9:29-38.
31. Xie D, Li A, Wang M, Fan Z, Feng H. LOCSVMPSI: a web server for subcellular localization of eukaryotic proteins using SVM and profile of PSI-BLAST. *Nucl Acids Res*. 2005;**33**(suppl 2):W105-W110.doi:10.1093/nar/gki359
32. Mundra P, Kumar M, Kumar KK, Jayaraman VK, Kulkarni BD. Using pseudo amino acid composition to predict protein subnuclear localization: Approached with PSSM. *Pattern Recognit Lett*. 2007;**28**(13):1610-1615. doi:10.1016/j.patrec.2007.04.001
33. Mullan L. Pairwise sequence alignment-it's all about us! *Brief Bioinform.* 2006;7(1):113-115. doi: 10.1093/bib/bbk008
34. Shibiao W, Mak MW, Kung SY, editors. Protein subcellular localization prediction based on profile alignment and Gene Ontology. *IEEE International Workshop on Machine Learning for Signal Processing*; 2011;18-21 Sept. doi: 10.1109/MLSP.2011.6064613

35. De Las Rivas J, Fontanillo C. Protein–Protein Interactions Essentials: Key Concepts to Building and Analyzing Interactome Networks. *PLoS Comput Biol.* 2010;**6**(6). doi: 10.1371/journal.pcbi.1000807
36. Huang Y, Li Y. Prediction of protein subcellular locations using fuzzy k-NN method. *Bioinformatics* (Oxford, Eng). 2004;**20**(1):21-28. doi: 10.1093/bioinformatics/btg366
37. Bulashevska A, Eils R. Predicting protein subcellular locations using hierarchical ensemble of Bayesian classifiers based on Markov chains. *BMC Bioinformatics.* 2006;**7**(1):1-13.
38. Zhang S-W, Pan Q, Zhang H-C, Shao Z-C, Shi J-Y. Prediction of protein homo-oligomer types by pseudo amino acid composition: Approached with an improved feature extraction and Naive Bayes Feature Fusion*. Amino Acids.* 2006;**30**(4):461-468. doi: 10.1007/s00726-006-0263-8
39. Zhou Z-H. *Ensemble Methods: Foundations and Algorithms*. Chapman \& Hall/CRC; 2012. 236 p.
40. Niu B, Jin Y-H, Feng K-Y, Lu W-C, Cai Y-D, Li G-Z. Using AdaBoost for the prediction of subcellular location of prokaryotic and eukaryotic proteins. *Mol Diversity.* 2008;**12**(1):41-45.doi: 10.1007/s11030-008-9073-0