International Journal of Biomedical Data Mining

Govindan and Nair, Biomedical Data Mining 2015, 3:2
DOI: 10.4172/2090-4924.1000109

Research Article Open Access

Sequence Features and Subset Selection Technique for the Prediction of Protein Trafficking Phenomenon in Eukaryotic Non Membrane Proteins

Geetha Govindan* and Achuthsankar S Nair

Department of Computational Biology and Bioinformatics, University of Kerala, Thiruvananthapuram, India

Abstract

Protein trafficking or protein sorting is the mechanism by which a cell transports proteins to the appropriate position in the cell or outside of it. This targeting is based on the information contained in the protein. Many methods predict the subcellular location of proteins in eukaryotes from the sequence information. However, most of these methods use a flat structure to perform prediction. In this work, we introduce ensemble methods to predict locations in the eukaryotic protein-sorting non membrane pathway hierarchically. We used features that were extracted exclusively from full length protein sequences with feature subset selection for classification. Sequence driven features, sequence mapped features and sequence autocorrelation features were tested with ensemble learners and classifier performances were compared with and without feature subset selection technique.

This study shows the new features extracted from full length eukaryotic protein sequences are effective at capturing biological features among compartments in eukaryotic non membrane pathways at two levels. Feature subset selection techniques helped to reduce the time taken for building the classification model.

Keywords: Sequence driven features; Sequence mapped features; Autocorrelation; Ensemble classifier; Pathways; Protein sorting

Introduction

Eukaryotic cells are organized into several membrane bound compartments. In order to perform the function; newly formed proteins get sorted and are delivered to various compartments in the non-membrane and trans-membrane pathways [1]. This protein sorting process in the pathway is very complex and still not clearly understood. But the most important principle of protein trafficking is that each protein has the information on its final localization site as a part of its amino acid sequence [2]. In 1983, Nishikawa, Kubota and Ooi had conducted investigations into predicting subcellular locations based on amino acid compositions. They had reported that the amino acid compositions have the discriminating ability to classify subcellular locations.

Prediction of protein localization sites in the pathways from the amino acid sequence has implications both for the function of the protein and its possibility of interacting with other proteins in the same compartment [3-5]. Protein sorting pathway in eukaryotes can be represented hierarchically like a tree structure [1,6,7]. Pathway at root level differentiates non membrane and trans-membrane proteins. Non membrane protein pathway can be further divided into secretory and non-secretory types. In a secretory pathway, proteins are delivered to the endoplasmic reticulum (ER), and then transported to other related locations.ER signal sequences, located in the N-terminal sequence, control this protein transport. In the non-secretory pathway, proteins with organelle-specific signal sequences are imported into the nucleus or mitochondria, according to their signal sequence type. The remaining proteins are located in the cytosol which lacks sorting signals [8,9] and some are localized by binding with another protein.

A wide variety of methods have been tried throughout the years in order to predict the subcellular localisation of proteins from their amino acid sequences (Olof). These methods differ in terms of sequence features as input data, techniques employed, time and cost to make the prediction about location. The success of computational prediction relies on the extraction of relevant biological features from the sequence and the computational techniques used [10-14]. Studies by Nakashima and Nishikawa [15], have shown that secretory and intracellular proteins differ significantly in their amino acid compositions and in residue pair frequencies. Hence in our study, priority was given to the features that can be extracted from the full length protein sequence based on various coding schemes without referencing external databases or external server generated outputs.

For computation, we used ensemble learning [16-20] hierarchically, (Figure 1) by mimicking the protein trafficking phenomenon; which is incorporated from the location descriptions provided by the Gene Ontology consortium (GO) [21] with the sequence features as input. First, this approach was used to classify the subcellular location of proteins. Second, this study was extended to determine whether the use of feature subset selection improves the prediction performance at various levels of hierarchy.

Materials and Methods

Data set

We used the recently published eukaryotic data set of LocTree2 [17] having 1682 proteins for testing and comparing. This is a manually curated database with experimental annotations for the subcellular localizations of proteins. In this dataset sequence bias was reduced through UniqueProt [22]. This bias reduction ascertained that no pair of proteins in the set had BLAST2 [23] HSSP-value (HVAL)> 0 [24,25]. We formed our data set ASN_G_1677 from this by verifying

***Corresponding author:** Geetha Govindan, Department of Computational Biology and Bioinformatics, University of Kerala, Thiruvananthapuram, India, Tel: +91-471-2308759; E-mail: geetha@sctimst.ac.in

Received February 14, 2014; **Accepted** November 22, 2014; **Published** February 15, 2015

Citation: Govindan G, Nair AS (2015) Sequence Features and Subset Selection Technique for the Prediction of Protein Trafficking Phenomenon in Eukaryotic Non Membrane Proteins. Biomedical Data Mining 3: 109. doi: 10.4172/2090-4924.1000109

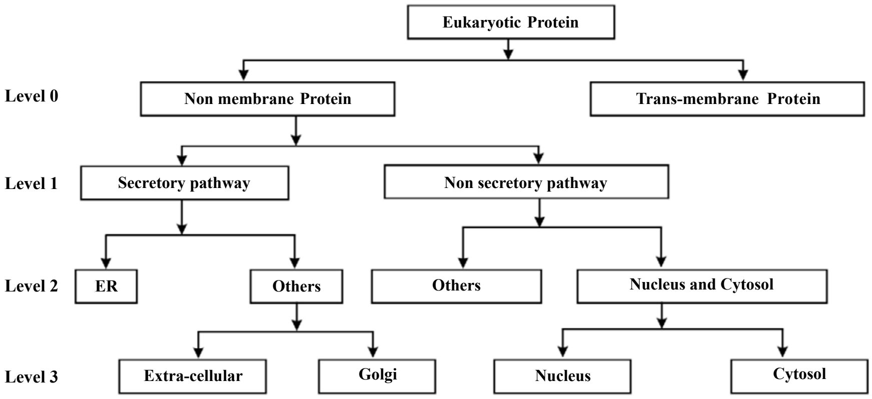

Figure 1: Hierarchical structures of compartments in protein trafficking based on cellular sorting used for this study. (Adopted from references [16-21]).

Pathway and subcellular location	No. of proteins
Trans-Membrane	**245**
Non-Membrane	**1432**
Chloroplast	133
Cytosol	212
Endoplasmic reticulum	10
Extra-cellular space	595
Golgi apparatus	2
Mitochondria	136
Nucleus	321
Peroxisome	6
Plastid	14
Vacuole	3
Total	**1677**

Table 1: Number of proteins in the data set ASN_G_1677.

with UniProt release 2013_05 [26] for the protein sequence and for the explicit annotation of subcellular localization. Annotations based on non-experimental findings ('potential', 'probable', or by 'similarity') and with multiple localization were excluded. The final data set had 1677 eukaryotic sequences with no over representation of a particular sub cellular protein. The list with pathway and subcellular location within the non-membrane pathway is mentioned in Table 1.

Sequence feature formation

The sequence feature extraction performed in this study can be classified into three groups. The first group is a method of converting the protein sequence into a numeric sequence by replacing each amino acid with its equivalent numeric values, counts etc. The second group is based on mapping amino acids into sub groups and the third group is based on features obtained from calculations based on autocorrelation.

1. Features directly from sequence. (Sequence driven feature)
2. Features by mapping the sequence. (Sequence mapped feature)
3. Features from sequence autocorrelation. (Sequence autocorrelation feature)

Sequence driven feature-amino acid dipeptide composition: (Dipeptide): A dipeptide is a molecule consisting of two amino acids joined by a single peptide bond and gives a feature vector with a dimension of 400 from the 20 amino acid combinations. The advantage of dipeptide sequence composition over amino acid composition is that it encapsulates global information about the fraction of amino acids as well as sequence order [27].

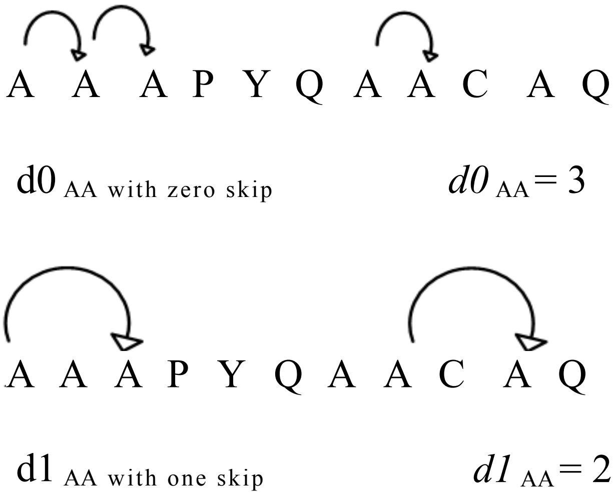

Figure 2: Amino acid di-peptide count with skips.

Consider a protein sequence AAAPYQAACAQ.

The dipeptide count with 0 skips, d0, is calculated by counting all pairs of amino acid conditions with no skips. In Figure 2, the count of d0AA is shown as 3, and one skip d1AA is counted as 2. The dipeptide count, 'dNxx', counts pairs with N skips between them.

The feature vector using the occurrence frequency count of a dipeptide to represent a protein sequence is formulated as follows:-

Given a protein sequence P with m amino acid residues, P= [R1 R_2 R_3 R_4 R_5 R_6 R_7 ...... R_m], where R_1, R_2 ….. $R_m$is the residues, we can map the sequence to a fixed length feature vector for each skip as P={f_1 $f_2 f_{3\ldots\ldots} f_{400}$}, where f_1, f_2 are the 400 native dipeptide occurrences (AA, AC, AD…… CA, CC, CD …. YV, YW, YY) counts in P.

The feature vector of the sequence' AAAPYQAACAQ'for dipeptided_0, d_1, $d_2$is as follows:-

Feature vector for occurrence frequency d_0

= [3 1 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 1 0 0 0 0 ……0] (1)

Feature vector for occurrence frequency d_1

= [2 1 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 1 0 0 0 0 0 ……0] (2)

Feature vector for occurrence frequency d_2

=[1 0 0 0 0 0 0 0 0 0 0 0 1 2 0 0 0 0 0 1 0 0 0 0 0 ……0] (3)

Each protein sequence is represented as three separate numeric counts of its dipeptide d_0, d_1 and d_2, each having 400 components. The feature vector having 1200 attributes is obtained by concatenating the corresponding vectors of d0, d_1 and d_2.

Sequence mapped feature (composition, transition, distribution (CTD)): The different properties of the amino acids result from the structural variations of the R groups. There are four different classes of amino acids determined by the side chains: (1) non-polar and neutral, (2) polar and neutral, (3) acidic and polar, (4) basic and polar. The twenty amino acids forming the protein sequence can also be divided into several groups based on their properties. Important properties are (5) charge, (6) hydrophilicity or hydrophobicity, (7) size, and (8) functional groups.

Twenty amino acids can be mapped into 1–3 groups by replacing each amino acid code with its group code. From the mapped sequence, features called Composition, Transition and Distribution (CTD) can be calculated.

Composition is the number of amino acids of a particular property divided by the total number of amino acids. Transition characterizes the percentage frequency with which amino acids of a particular property are followed by amino acids of a different property. Distribution measures the chain length within which the first, 25%, 50%, 75% and 100% of the amino acids of a particular property are located.

Through this method, amino acids are grouped into three classes according to their property types, as shown in Table 2, and are encoded by the numeric indices 1, 2, 3. The attributes of charge, hydrophobicity, normalized van der Waals volume; polarity, Polaris ability, secondary structure and solvent accessibility are used as properties [28-33].

Consider a sample sequence ‘RKEDQNGASTPHYCLVIMFW’. According to hydrophobicity grouping, this sequence is encoded as “11111122222223333333”.

Composition is the global percentage for each encoded class in the sequence. In this example the total count of 1, 2, 3 is 6, 7, 7 and hence composition is calculated as 6/20, 7/20 and 7/20.

$$Composition = \frac{N_e}{N} \tag{4}$$

where e=1, 2, 3.N_e is the number of e in the encoded sequence and N is the total length of the sequence.

The transition from class 1 to 2 is the percentage frequency with which 1 is followed by 2 or 2 is followed by 1 in the encoded sequence.

The transition descriptor is calculated as

$$Tmn = \frac{\mathrm{Nmn} + \mathrm{Nnm}}{N-1} \tag{5}$$

Where mn=“12”, “13”,”23” and Nnm, and Nnm are the numbers of dipeptide encoded as “mn” and “nm” respectively in the sequence.N is the length of the sequence. For the given sample sequence, Transition=2/19.

The distribution descriptor describes the distribution of each property in the sequence. There are five distribution descriptors for each property and they are the position percentages in the sequence for the first residue, 25% of the residues, 50% of the residues, 75% of the residues and 100% of the residues.

The CTD calculation is performed for 7 properties for each protein sequence after dividing each sequence into three equal segments. In total, 21 x 3 attributes for a sequence and 441 attributes for 7 properties comprise the final feature vector.

Sequence autocorrelation features (Autocorrelation Descriptors (ACD)): Sequence autocorrelation-based features are based on the Tobler’s First law of geography: “everything is related to everything else but nearby things are more related than distant things” [34] Sequence autocorrelation-based features also assume that “the disturbances in each area are systematically related to those in adjacent areas” [35]. Spatial autocorrelation is positive when nearby things are similar and negative when they are dissimilar**.** It measures the degree to which near and distant things are related. This concept helps to analyze the dependency among the features of sequences in each location.

Autocorrelation features are calculated based on the distribution of amino acid properties along the sequence. Amino acid indices related to hydrophobicity are used for calculation after replacing each amino acid with its equivalent normalized index as P_i. Three autocorrelation descriptors are used as features. They are normalized Moreau-Broto autocorrelation descriptors [36,37] Moran auto-correlation descriptors

Sl	Property	Group 1	Group 2	Group 3
1	**Charge**	**Neutral**	**Negatively charged**	**Positively charged**
	Amino acids	A,C,F,G,H,I,L,M,N,P,Q,S,T,V,W,Y	D, E	K, R
2	**Hydrophobicity**	**Hydrophobicity**	**Neutral**	**Polar**
	Amino acids	C,F,I,L,M,V,W	A,G,H,P,S,T,Y	D, E, K, N, Q, R
3	**Normalised Vander Waals volume**	**0-2.78**	**2.95-4.0**	**4.03-8.08**
	Amino acids	A,C,D,G,P,S,T	E, I, L, N, Q, V	F,H,K,M,R,W,Y
4	**Polarity**	**4.9-6.2**	**8.0-9.2**	**10.4-13.0**
	Amino acids	C,F,I,L,M,V,W,Y	A, G, P, S, T	D,E,H,K,N,Q,R
5	**Polarisability**	**0 - .108**	**0.128-0.186**	**0.219-0.409**
	Amino acids	A, D, G, S, T	C,E,I,L,N,P,Q,V	F,H,K,M,R,W,Y
6	**Secondary Structure**	**Coil**	**Helix**	**Strand**
	Amino acids	D,G,N,P,S	A, E, H, K, L, M, Q, R	C,F,I,T,V,W,Y
7	**Solvent Accessibility**	**Buried**	**Intermediate**	**Exposed**
	Amino acids	A, C, F, G, I, L, V, W	H,M,P,S,T,Y	D, E, K, N, R, Q

Table 2: Amino acid attributes and division of the amino acids into groups.

[38] and Geary autocorrelation descriptors [39].

The Moreau-Broto autocorrelation descriptor is defined as

$$MB(\mathrm{d}) = \sum_{i=1}^{N-d} PiPi+d \text{ Where d =1, 2, 3 upto Max.lag} \qquad (6)$$

d is the lag of the autocorrelation, N is the length of the sequence, and P_i and P_{i+d} are the amino acid index value of the selected property at position I and i+d, respectively. Max. Lag is the maximum value of the lag. The normalized Moreau-Broto autocorrelation descriptors are defined as

$$\text{Normalised Moreau – Broto autocorrelation descriptor} = \frac{\mathrm{MB}(\mathrm{d})}{\mathrm{N-d}} \qquad (7)$$

The Moran autocorrelation descriptor is defined as

$$Moran(d) = \frac{\frac{1}{N-d}\sum_{i=1}^{N-d}(\mathrm{P}_i-\overline{P})(\mathrm{Pi+d}-\overline{\mathrm{P}})}{\frac{1}{N}\sum_{i=1}^{N}(\mathrm{Pi}-\overline{\mathrm{P}})^2}\mathrm{d}=1,2,3\ldots,30 \qquad (8)$$

$$\overline{P} = \frac{\sum_{i=1}^{N} P_i}{N} \qquad (9)$$

Where P_i, P_{i+d} have the same meaning as above.

The Geary autocorrelation descriptor is defined as

$$Geary(d) = \frac{\frac{1}{2(N-d)}\sum_{i=1}^{N-d}(P_i-P_{i+d})^2}{\frac{1}{N-1}\sum_{i=1}^{N}(P_i-\overline{P})^2} D=1,2,3....30\,. \qquad (10)$$

Where P^{-}, Pi, Pi+d have the same meaning as above. 3510 features from 39 amino acid properties with 30 lag form the sequence feature vector for autocorrelation.

The combined feature vector from three groups had 5151 elements.

Feature subset selection

Feature subset selection is used as a pre-processing step in machine learning methods. The performance of a classifier depends on the number of features, sample size and algorithm complexity. Feature selection is effective in removing irrelevant and redundant features, increasing efficiency in learning tasks, improving learning performance like predictive accuracy, and enhancing comprehensibility of learned results [40-42]. In this study, a fast filter method called FCBF (Fast Correlation Based Filter) which can identify relevant features as well as redundancy among relevant features without pairwise correlation analysis was adopted [43,44]. The FCBF filter algorithm is designed for high-dimensional data and has been shown effective in removing both irrelevant features and redundant features. This algorithm has two stages: the first stage is based on relevance analysis, aimed at ordering the input variables depending on a relevance score and the second stage is a redundancy analysis, aimed at selecting predominant features from the relevant set obtained in the first stage. With FCBF, we were able to reduce the numbers of features to the range of 100 from 5151.

Computational techniques used (Ensemble Learning)

Ensemble learning is an effective method that has been adopted to combine multiple machine learning algorithms to improve overall prediction accuracy by aggregating the predictions of all algorithms [45]. Multiple learners (base learners) are trained to solve the same problem. It is a set of classifiers whose individual decisions are combined in some way (typically by weighted or unweighted voting) to classify new examples. This method averages over multiple classification models; and each model have different input feature vectors. Weak individual models are transformed into strong ensemble models.

The aim of using the ensemble method is to achieve more accurate classification (on training data) as well as better generalization (on unseen data). These ensemble techniques reduce the small sample size problem which is critical in biological applications and multiple prediction models can be tested with different feature sets. The two most popular classifiers based on the ensemble method, are Bagging [46] and Ada boostM1 [47]. In this study, these two methods were used to predict protein trafficking in the pathway.

Bagging generates new training sets by sampling with the substitution of the training data while boosting adopts an adaptive sampling by using all instances of iteration. In both methods, multiple classifiers are combined using a simple voting system to create a Meta classifier. In Bagging, each classifier has the vote of the same strength, whereas boosting assigns different voting strengths to classifiers based on their accuracy.

Performance evaluation

Basic ensemble based classifiers; Adaboost M1 and Bagging were trained to classify the location compartment of proteins in the pathway using WEKA [48]. Two tests were carried out with ASN_G_1677 dataset for performance evaluation at all levels in the hierarchy as shown in Figure 1. 5 fold cross-validation test (randomly partitioning the dataset into equally sized training and test sets; training on 4 sets and testing with5thset and averaging the results) and (2) independent data test (training on one set and testing with another test set by dividing the dataset into two equal sized random groups). The classifier performance evaluation parameters Specificity, Sensitivity, Accuracy, Mathew correlation coefficient [49], Positive predictive value [50], Negative predictive value [50] and Receiver operating characteristic [51] were calculated at all levels as per the below equations.

$$\text{Specificity } (\mathrm{S_p}) = \frac{TN}{TN+FP}$$

$$\text{Sensitivity } (\mathrm{S_n}) = \frac{TP}{TP+FN}$$

Mathew's Correlation coefficient ($\mathrm{M_{cc}}$) =

$$\frac{TP.TN-FP.FN}{Sqrt\left((TP+FN).(TP+FP).(TN+FN)(TN+FP)\right)}$$

Accuracy($\mathrm{A_{cc}}$)=

$$\frac{True\,Positive(TP)+True\,Negative(TN)}{True\,Positive(TP)+True\,Negative(TN)+False\,Positive(FP)+False\,Negative(FN)}$$

$$\text{Positive predictive value } (\mathrm{P_{pv}}) = \frac{True\,Positive(TP)}{True\,Positive(TP)+False\,Positive(FP)}$$

$$\text{Negative predictive value } (\mathrm{N_{pv}}) = \frac{True\,Negative(TN)}{True\,Negative(TN)+False\,Negative(FN)}$$

Results and Discussion

Here the final 1677 protein sequences were represented in two groups; by combining the three different sequence features with and without feature subset selection. As is well known, 5 fold cross-validation test and independent data test were performed on these two feature groups to evaluate the quality of the classifier. Tables 3 and 4 shows the performance evaluation parameter summary of classifiers against these two feature groups. Parameters S_p, S_n, M_{cc}, P_{pv}, N_{pv},A_{cc}, ROC and time taken to build the model were obtained from the two tests at various levels of the pathway for the two feature groups using two classifiers. Mcc which is regarded as a balanced measure even for data groups of different sizes; reported 0.5 at level 0 (between non membrane and trans-membrane pathway) and level 1 (between secretory and non-secretory pathway) for both tests. Both tests with feature subset selection; enhanced the average value of Mcc to 0.6. In level 2, between the pathway ER, others and in level 3 between extracellular and Golgi; though the positive predictive value is higher, Mcc value is less than zero. Hence there is disagreement between prediction and observation due to small and unbalanced data size at these levels.

ROC analysis provides a systematic tool for quantifying the impact of variability among individuals' decision thresholds. The term receiver operating characteristic (ROC) originates from the use of radar during World War II. Just as American soldiers deciphered a blip on the radar screen as a German bomber, a friendly plane, or just noise, radiologists face the task of identifying abnormal tissue against a complicated background. As radar technology advanced during the war, the need for a standard system to evaluate detection accuracy became apparent. ROC analysis was developed as a standard methodology to quantify a signal receiver's ability to correctly distinguish objects of interest from the background noise in the system.

Comparison

In this study, a hierarchical system for the prediction of protein subcellular localization was tested. In order to roundly assess our method, we carried a comparison with the published report of LOCtree [16] and LocTree2 [17]. LOCtree used the amino acid composition, composition of the 50 N terminal residues, pseudo amino acid composition from three secondary structure states and Signal server [52] outputs as a feature vector on support vector machine. LocTree2 used the profiles created by BLAST-ing [23].

Results reported by LocTree2 [17] is directly comparable to ours in terms of selection of a dataset with no feature subset selection. The overall accuracy mentioned in LocTree2 [17] is the positive predictive value (P_{pv}) based on the fivefoldcross-validation experiments and comparison with our P_{pv} values at all levels is shown in Table 5.

Tables 3 and 4 show that at level 0 between the non-membrane and trans-membrane pathway; 5 fold cross-validation and independent data test based on Adaboost M, Bagging reported accuracies above 89% with Mcc above 0.5, with and without feature subset selection. In 5 fold cross validation test; Bagging with feature subset selection reported accuracy similar to LocTree2 in level 1 between the secretory and non-secretory path way and in level 2 between Nucleus, Cytosol – others pathway with Mcc value greater than 0.44. At Level 3 between nucleus and cytosol pathway; Bagging reported positive predictive value of 62% with Mcc of 0.34 while LocTree2 reported 67%.

Conclusion

Protein transport to compartments is a topic which is now also poorly understood. Major protein localisation prediction methods

Classifier	Adaboost								Bagging							
Feature	Sp	Sn	Mcc	Ppv	Npv	Acc	ROC	Time in Sec. – to build the model	Sp	Sn	Mcc	Ppv	Npv	Acc	ROC	Time in Sec. – to build the model
Di-peptide+ CTD +ACD	50%	96%	0.521	92	67	89%	0.90	28.97	50%	99%	0.625	92	87	92%	0.93	65.98
Di-peptide+ CTD +ACD with FCBF feature subset selection (75 features)	59%	96%	0.601	93	72	91%	0.92	0.38	55%	99%	0.653	93	88	92%	0.93	0.84

Table 3: Performance evaluation summary of classifiers against features for the 5 fold cross-validation test at all levels of hierarchy for dataset ASN_G_1677. Performance evaluation of classification of proteins between non-membrane and trans membrane pathway at Level 0.

Classifier	Adaboost								Bagging							
Feature	Sp	Sn	Mcc	Ppv	Npv	Acc	ROC	Time in Sec. – to build the model	Sp	Sn	Mcc	Ppv	Npv	Acc	ROC	Time in Sec. – to build the model
Di-peptide+ CTD +ACD (5151 features)	86%	69%	0.559	78	79	79%	0.85	26.11	87%	77%	0.645	81	84	83%	0.91	78.5
Di-peptide+ CTD +ACD with FCBF feature subset selection (77 features)	86%	73%	0.591	79	81	80%	0.87	0.36	87%	73%	0.613	81	81	81%	0.88	1.23

Table 3a: Performance evaluation summary of classifiers against features for the 5 fold cross-validation test at all levels of hierarchy for dataset ASN_G_1677. Performance evaluation of classification of proteins between secretory and non-secretory pathway at Level 1.

Classifier	Adaboost								Bagging							
Feature	Sp	Sn	Mcc	Ppv	Npv	Acc	ROC	Time in Sec. – to build the model	Sp	Sn	Mcc	Ppv	Npv	Acc	ROC	Time in Sec. – to build the model
Di-peptide+ CTD +ACD (5151 features)	0%	99%	0.01	98	0	98%	0.58	10.3	0%	100%	<0	98	<0	98%	0.76	18.86
Di-peptide+ CTD +ACD with FCBF feature subset selection (32 features)	10%	99%	0.15	99	25	98%	0.93	0.02	0%	100%	<0	98	<0	98%	0.79	0.05

Table 3b: Performance evaluation summary of classifiers against features for the 5 fold cross-validation test at all levels of hierarchy for dataset ASN_G_1677. Performance evaluation of classification of proteins between ER and others at Level 2.

Classifier	**Adaboost**								**Bagging**							
Feature	Sp	Sn	Mcc	Ppv	Npv	Acc	ROC	Time in Sec. – to build the model	Sp	Sn	Mcc	Ppv	Npv	Acc	ROC	Time in Sec. – to build the model
Di-peptide+ CTD +ACD (5151 features)	82%	63%	0.454	66	80	75%	0.8	14.58	87%	56%	0.459	71	78	76%	0.82	40.34
Di-peptide+ CTD +ACD with FCBF feature subset selection (33 features)	87%	57%	0.465	71	79	76%	0.81	0.08	89%	63%	0.539	75	81	80%	0.86	0.25

Table 3c: Performance evaluation summary of classifiers against features for the 5 fold cross-validation test at all levels of hierarchy for dataset ASN_G_1677. Performance evaluation of classification of proteins between Nucleus, Cytosol and others at Level 2.

Classifier	**Adaboost**								**Bagging**							
Feature	Sp	Sn	Mcc	Ppv	Npv	Acc	ROC	Time in Sec. – to build the model	Sp	Sn	Mcc	Ppv	Npv	Acc	ROC	Time in Sec. – to build the model
Di-peptide+ CTD +ACD (5151 features)	0%	100%	<0	100	0	99%	0.35	10.2	0%	100%	<0	100	<0	100%	0.3	7.59
Di-peptide+ CTD +ACD with FCBF feature subset selection (11 features)	50%	100%	0.498	100	50	100%	0.5	0.02	0%	100%	<0	100	<0	100%	0.3	0.02

Table 3d: Performance evaluation summary of classifiers against features for the 5 fold cross-validation test at all levels of hierarchy for dataset ASN_G_1677. Performance evaluation of classification of proteins between extra-cellular and golgi at Level 3.

Classifier	**Adaboost**								**Bagging**							
Feature	Sp	Sn	Mcc	Ppv	Npv	Acc	ROC	Time in Sec. – to build the model	Sp	Sn	Mcc	Ppv	Npv	Acc	ROC	Time in Sec. – to build the model
Di-peptide+ CTD +ACD (5151 features)	74%	47%	0.208	54	68	63%	0.68	8.72	77%	50%	0.279	59	70	66%	0.72	23.42
Di-peptide+ CTD +ACD with FCBF feature subset selection (26 features)	77%	48%	0.264	58	69	66%	0.7	0.05	76%	59%	0.349	62	74	69%	0.75	0.34

(Sp– pecificity, Sn–Sensitivity, Acc–Accuracy, Mcc–Mathews correlation coefficient , Ppv–Positive predictive value, Npv–Negative predictive value, ROC–Receiver operating characteristic)

Table 3e: Performance evaluation summary of classifiers against features for the 5 fold cross-validation test at all levels of hierarchy for dataset ASN_G_1677. Performance evaluation of classification of proteins between Nucleus and Cytosol at Level 3.

Classifier	**Adaboost**								**Bagging**							
Feature	Sp	Sn	Mcc	Ppv	Npv	Acc	ROC	Time in Sec. – to build the model	Sp	Sn	Mcc	Ppv	Npv	Acc	ROC	Time in Sec. – to build the model
Di-peptide+ CTD +ACD (5151 features)	52%	97%	0.583	93	76	91%	0.91	29.53	52%	99%	0.638	93	87	92%	0.94	65.28
Di-peptide+ CTD +ACD with FCBF feature subset selection (75 features)	64%	93%	0.544	94	58	89%	0.9	0.36	56%	97%	0.587	93	72	91%	0.93	0.88

Table 4: Performance evaluation summary of classifiers against features for the independent test at all levels of hierarchy for dataset ASN_G_1677. Performance evaluation of classification of proteins between non membrane and trans membrane pathway at Level 0.

Classifier	**Adaboost**								**Bagging**							
Feature	Sp	Sn	Mcc	Ppv	Npv	Acc	ROC	Time in Sec. – to build the model	Sp	Sn	Mcc	Ppv	Npv	Acc	ROC	Time in Sec. – to build the model
Di-peptide+ CTD +ACD (5151 features)	82%	71%	0.528	74	79	77%	0.84	25.17	82%	74%	0.558	75	81	78%	0.87	78.3
Di-peptide+ CTD +ACD with FCBF feature subset selection (77 features)	80%	71%	0.514	72	79	76%	0.85	0.36	86%	73%	0.599	79	82	81%	0.87	1.22

Table 4a: Performance evaluation summary of classifiers against features for the independent test at all levels of hierarchy for dataset ASN_G_1677. Performance evaluation of classification of proteins between secretory and non-secretory pathway at Level 1.

Classifier	**Adaboost**								**Bagging**							
Feature	Sp	Sn	Mcc	Ppv	Npv	Acc	ROC	Time in Sec. – to build the model	Sp	Sn	Mcc	Ppv	Npv	Acc	ROC	Time in Sec. – to build the model
Di-peptide+ CTD +ACD (5151 features)	0%	100%	<0	97	<0	97%	0.48	10.41	0%	100%	<0	97	<0	97%	0.5	10.63
Di-peptide+ CTD +ACD with FCBF feature subset selection (32 features)	13%	100%	0.241	98	50	97%	0.94	0.03	0%	100%	<0	97	<0	97%	0.5	0.03

Table 4b: Performance evaluation summary of classifiers against features for the independent test at all levels of hierarchy for dataset ASN_G_1677. Performance evaluation of classification of proteins between ER and others at Level 2.

Classifier	**Adaboost**								**Bagging**							
Feature	Sp	Sn	Mcc	Ppv	Npv	Acc	ROC	Time in Sec. – to build the model	Sp	Sn	Mcc	Ppv	Npv	Acc	ROC	Time in Sec. – to build the model
Di-peptide+ CTD +ACD (5151 features)	83%	48%	0.326	61	74	70%	0.77	13.78	85%	49%	0.369	65	75	72%	0.8	40.28
Di-peptide+ CTD +ACD with FCBF feature subset selection (33 features)	91%	50%	0.457	76	76	76%	0.83	0.08	88%	53%	0.442	71	77	75%	0.83	0.25

Table 4c: Performance evaluation summary of classifiers against features for the independent test at all levels of hierarchy for dataset ASN_G_1677. Performance evaluation of classification of proteins between Nucleus, Cytosol and others at Level 2.

Classifier	**Adaboost**								**Bagging**							
Feature	Sp	Sn	Mcc	Ppv	Npv	Acc	ROC	Time in Sec. – to build the model	Sp	Sn	Mcc	Ppv	Npv	Acc	ROC	Time in Sec. – to build the model
Di-peptide+ CTD +ACD (5151 features)	0%	100%	<0	99	<0	99%	0.5	9.89	0%	100%	<0	99	<0	99%	0.5	7.22
Di-peptide+ CTD +ACD with FCBF feature subset selection (11 features)	0%	100%	<0	99	<0	99%	0.5	0.02	0%	100%	<0	99	<0	99%	0.5	0

Table 4d: Performance evaluation summary of classifiers against features for the independent test at all levels of hierarchy for dataset ASN_G_1677. Performance evaluation of classification of proteins between extra-cellular and golgi at Level 3.

Classifier	**Adaboost**								**Bagging**							
Feature	Sp	Sn	Mcc	Ppv	Npv	Acc	ROC	Time in Sec. – to build the model	Sp	Sn	Mcc	Ppv	Npv	Acc	ROC	Time in Sec. – to build the model
Di-peptide+ CTD +ACD (5151 features)	80%	38%	0.203	57	65	63%	0.67	8.7	82%	42%	0.255	61	67	65%	0.68	24.17
Di-peptide+ CTD +ACD with FCBF feature subset selection (26 features)	74%	44%	0.183	53	66	62%	0.68	0.06	75%	49%	0.245	57	68	64%	0.67	0.14

(Sp–Specificity, Sn–Sensitivity, Acc–Accuracy, Mcc–Mathews correlation coefficient , Ppv–Positive predictive value, Npv–Negative predictive value, ROC–Receiver operating characteristic)

Table 4e: Performance evaluation summary of classifiers against features for the independent test at all levels of hierarchy for dataset ASN_G_1677. Performance evaluation of classification of proteins between Nucleus and Cytosol at Level 3.

LocTree2 5 fold cross validation with SVM		**Our method 5 fold cross validation**		
Levels	Ppv	Sequence Feature	Adaboost -Ppv	Bagging-Ppv
Level 0 - Non membrane and Trans-membrane pathway	**90%**	Di-peptide+ CTD +ACD (5151 features)	**92%**	**92%**
		Di-peptide+ CTD +ACD with FCBF feature subset selection (75 features)	**93%**	**93%**
Level 1 - Secretory and Non secretory pathway	**83%**	Di-peptide+ CTD +ACD (5151 features)	78%	**81%**
		Di-peptide+ CTD +ACD with FCBF feature subset selection (77 features)	79%	**81%**
Level 2 - ER and Others	75%	Di-peptide+ CTD +ACD (5151 features)	98%	98%
		Di-peptide+ CTD +ACD with FCBF feature subset selection (32 features)	99%	98%
Level 2 - Nucleus, Cytosol and others	**75%**	Di-peptide+ CTD +ACD (5151 features)	66%	71%
		Di-peptide+ CTD +ACD with FCBF feature subset selection (33 features)	71%	**75%**
Level 3 - Extra-cellular and golgi	80%	Di-peptide+ CTD +ACD (5151 features)	100%	100%
		Di-peptide+ CTD +ACD with FCBF feature subset selection (11 features)	100%	100%
Level 3 - Nucleus and Cytosol	**67%**	Di-peptide+ CTD +ACD (5151 features)	54%	59%
		Di-peptide+ CTD +ACD with FCBF feature subset selection (26 features)	58%	**62%**

(SVM–Support Vector Machine, Ppv–Positive predictive value)

Table 5: Comparison of the 5 fold cross-validation results with the published results of LocTree2.

have been implemented using standard machine learning algorithms with parallel architecture for classification [53-56]. Here a novel system of ensemble learners, using hierarchical architecture with features extracted directly from full length protein sequences, with and without feature subset selection was tested. Test results, at the non-membrane pathway of hierarchy show that the prediction accuracy can be significantly improved by using the classifier Bagging and FCBF feature subset selection with significant reduction in time for model building. Accuracy above 90% using bagging on independent data tests indicates that the native protein localization is imprinted onto the protein sequence for each compartment. Sequence features experimented share a common composition and explicitly utilizes intrinsic correlation between proteins that share these common features. Additionally, this hierarchical structure has provided insights into the sorting process, such as the accurate distinction between the intracellular and secretory pathway. Our study supports the hypothesis reported by Nakashima and Nishawa [15].

In the future, it should be possible to extend the classification to any level in the hierarchy using these sequence features and with the location descriptions provided by the gene ontology consortium (GO) [21]. This method can predict the final location of the protein as well as the mechanism of localization. Our findings may contribute to the development of clinical strategies related to drug design. We observed that, as one descends the hierarchical path, the prediction accuracy progressively decreases as the classification task complexity increases. The best scoring decisions are at the top, and the worst are at the bottom. Major problem with this type of hierarchical model is its inability to correct a prediction mistake made at the top node.

References

1. Alberta B, Johnson A, Lewis J, Raff M, Roberts K, et al. (2007) Molecular Biology of the Cell. Garland Science: New York.
2. Nishikawa K, Kubota Y, Ooi T (1983) Classification of proteins into groups based on amino acid composition and other characters. J Biochem 94: 997-1007.
3. Bork P, Eisenhaber F (1998) Wanted: subcellular localization of proteins based on sequence. Trends Cell Biol 8: 169-170.
4. Drawid A, Gerstein M (2000) A Bayesian system integrating expression data with sequence patterns for localizing proteins: comprehensive application to the yeast genome. J Mol Biol 301: 1059-1075.
5. Bork P, Koonin EV (1998) Predicting functions from protein sequences-where are the bottlenecks? 18: 313-318.
6. Rusch SL, Kendall D (1995) Protein transport via amino-terminal targeting sequences: common themes in diverse systems. Molecular Membrane Biology 12: 295-307.
7. Horton P, Nakai K (1996) A probabilistic classification system for predicting the cellular localization sites of proteins. Fourth International Conference on Intelligent Systems for Molecular Biology 109-115.
8. Lodish H, Berk A, Zipursky (2000) S.L. Molecular Cell Biology. (5thedn). Sinauer Associates.
9. Cooper GM, Hausman RE (2009) The Cell: A Molecular Approach. (5thedn). Sinauer Associates, Inc.
10. Reinhardt A, Hubbard T (1998) Using neural networks for prediction of the subcellular location of proteins. Nucl Acids Res 26: 2230-2236.
11. Emanuelsson O, Nielsen H, Brunak S, Heijne GV (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J Mol Biol 300: 1005-1016.
12. Nakai K (2001) Prediction of in vivo fates of proteins in the era of genomics and proteomics. J Struct Biol 134: 103-116.
13. Chou KC, Cai YD (2003) Prediction and classification of protein subcellular location-sequence-order effect and pseudo amino acid composition. J Cell Biochem 90: 1250-1260.
14. Tantoso E, Li KB (2008) AAindexLoc: predicting subcellular localization of proteins based on a new representation of sequences using amino acid indices 35: 345-353.
15. Nakashima H, Nishikawa K (1994) Discrimination of intracellular and extracellular proteins using amino acid composition and residue pair frequencies. J Mol Biol 238: 54-61.
16. Nair R, Rost B (2005) Mimicking Cellular Sorting Improves Prediction of Subcellular Localization. J Mol Biol 348: 85-100.
17. Goldberg T, Hamp T, Rost B (2012) LocTree2 predicts localization for all domains of life. Bioinformatics 28: i458-458i465.
18. Pierleoni A, Martelli PL, Fariselli P, Casadio R (2007) eSLDB: eukaryotic subcellular localization database. Nucleic Acids Res 35: D208-212.
19. Pierleoni A, Martelli PL, Fariselli P, Casadio R (2006) BaCelLo: a balanced subcellular localization predictor. Bioinformatics 22: e408-416.
20. Lin T, Murphy RF, Joseph ZB (2011) Discriminative Motif Finding For Predicting Protein Subcellular Localization. IEEE/ACM Trans Comput Biol Bioinform 8: 441-451.
21. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25: 25-29.
22. Mika S, Rost B (2003) UniqueProt: Creating representative protein sequence sets. Nucleic Acids Res 31: 3789-3791.
23. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403-410.
24. Rost B (1999) Twilight zone of protein sequence alignments. Protein Eng 12: 85-94.
25. Sander C, Schneider R (1991) Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins 9: 56-68.
26. UniProt Consortium1 (2012) Reorganizing the protein space at the Universal Protein Resource (UniProt). Nucleic Acids Res 40: D71-75.
27. Ding Y, Cai Y, Zhang G, Xu W (2004) The influence of dipeptide composition on protein thermostability. FEBS Lett 569: 284-288.
28. Dubchak I, Muchnik I, Holbrook SR, Kim SH (1995) Prediction of protein folding class using global description of amino acid sequence. Proc Natl Acad Sci U S A 92: 8700-8704.
29. Cai CZ, Han LY, Ji ZL, Chen X, Chen YZ (2003) SVM-Prot: Web-based support vector machine software for functional classification of a protein from its primary sequence. Nucleic Acids Res 31: 3692-3697.
30. Lin HH, Han LY, Zhang HL, Zheng CJ, Xie B, et al. (2006) Prediction of the Functional Class of Metal-Binding Proteins from Sequence Derived Physicochemical Properties by Support Vector Machine Approach. BMC Bioinformatics 7(Suppl 5): S13.
31. Han LY, Cai CZ, Lo SL, Chung MC, Chen YZ (2004) Prediction of RNA-binding proteins from primary sequence by a support vector machine approach. RNA 10: 355-368.
32. Han LY, Zheng CJ, Xie B, Jia J, Ma XH, et al. (2007) Support vector machines approach for predicting druggable proteins: recent progress in its exploration and investigation of its usefulness. Drug Discov Today 12: 304-313.
33. Rao HB, Zhu F, GB Yang, Z.R Li, YZ Chen (2011) Update of PROFEAT: a Web Server for Computing Structural and Physicochemical Features of Proteins and Peptides from Amino Acid Sequence. Nucleic Acids Res.
34. Tobler, WR (1970) A computer movie simulating urban growth in the Detroit region. Economic Geography, 46: 234-240.
35. Loftin, Colin, Sally, K Ward (1981) Spatial autocorrelation models for Galton's problem. Cross-Cultural Research 16: 105-141.
36. Feng ZP, Zhang CT (2000) Prediction of membrane protein types based on the hydrophobic index of amino acids. J Protein Chem 19: 269-275.
37. Lin Z, Pan XM (2001) Accurate prediction of protein secondary structural content. J Protein Chem 20: 217-220.
38. Horne DS (1988) Prediction of protein helix content from an autocorrelation analysis of sequence hydrophobicities. Biopolymers 27: 451-477.
39. Sokal RR, Thomson BA (2006) Population structure inferred by local spatial autocorrelation: an example from an Amerindian tribal population. Am J Phys Anthropol 129: 121-131.
40. Blum LA, Langley P (1997) Selection of relevant features and examples in machine learning. Artificial Intelligence 97: 245-271.
41. Dash M, Liu H (1997) Feature selection for classifications. Intelligent Data Analysis 1: 131-156.
42. Kohavi R, John GH (1997) Wrappers for feature subset selection. Artificial Intelligence 97: 273-324.
43. Yu L, Liu H (2003) Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution. In Proceedings of The Twentieth International Conference on Machine Leaning (ICML-03).
44. Liu H, Hussain F, Tan CL, Dash M (2002) "Discretization: An Enabling Technique". Journal of Data Mining and Knowledge Discovery 6: 393-423.
45. Pengyi, Yang, Yee, Yang H, Bing B (2010) A review of ensemble methods in bioinformatics. Current Bioinformatics 5: 296-308.
46. Breiman L (1996) Bagging predictors. Machine Learning 26: 123-140.
47. Freund Y, Schapire RE (1996) Experiments with a new boosting algorithm. Proceedings of the Thirteenth International Conference on Machine Learning.

48. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, et al. (2009) The WEKA Data Mining Software: An Update. ACM SIGKDD Explorations Newsletter 1: 10-18.

49. Matthews BW (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta 405: 442-451.

50. Altman DG, Bland JM (1994) Diagnostic tests 2: Predictive values. BMJ 309: 102.

51. Spackman KA (1989) Signal detection theory: Valuable tools for evaluating inductive learning. Proceedings of the Sixth International Workshop on Machine Learning, Morgan Kaufmann Publishers Inc.

52. Bendtsen JD, Nielsen H, von Heijne G, Brunak S (2004) Improved prediction of signal peptides: SignalP 3.0. J Mol Biol 340: 783-795.

53. Cherian BS, Nair AS (2010) Protein location prediction using atomic composition and global features of the amino acid sequence. Biochem Biophys Res Commun 391: 1670-1674.

54. Yu Su EC, Chiu HS, Lo A, Kang Hwang J, Yi Sung T, et al. (2007) Protein subcellular localization prediction based on compartment-specific features and structure conservation. BMC Bioinformatics 8: 330.

55. Blum T, Briesemeister S, Kohlbacher O (2009) MultiLoc2: integrating phylogeny and Gene Ontology terms improves subcellular protein localization prediction. BMC Bioinformatics 10: 274.

56. Emanuelsson OL, Nielsen H, Brunak S, von Heijne G (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J Mol Biol 300: 1005-1016.