Feature Fusion Based SVM Classifier for Protein Subcellular Localization Prediction

Julia Rahman[1,*], Md. Nazrul Islam Mondal[1], Md. Khaled Ben Islam[1,2], Md. Al Mehedi Hasan[1]

[1]Department of Computer Science & Engineering, Rajshahi University of Engineering & Technology, Rajshahi, Bangladesh

[2]Department of Computer Science & Engineering, Pabna University of Science & Technology, Pabna, Bangladesh

Summary

For the importance of protein subcellular localization in different branch of life science and drug discovery, researchers have focused their attentions on protein subcellular localization prediction. Effective representation of features from protein sequences plays most vital role in protein subcellular localization prediction specially in case of machine learning technique. Single feature representation like pseudo amino acid composition (PseAAC), physiochemical property model (PPM), amino acid index distribution (AAID) contains insufficient information from protein sequences. To deal with such problem, we have proposed two feature fusion representations AAIDPAAC and PPMPAAC to work with Support Vector Machine classifier, which fused PseAAC with PPM and AAID accordingly. We have evaluated performance for both single and fused feature representation of Gram-negative bacterial dataset. We have got at least 3% more actual accuracy by AAIDPAAC and 2% more locative accuracy by PPMPAAC than single feature representation.

1 Introduction

Knowledge of protein's subcellular localization plays an important role in its function prediction, helps to understand its role in biological processes, identification of drug targets and genome annotation. Not only these, but also there are many biological implication of localization prediction. Because of the vital roles of protein's location, subcellular localization prediction becomes one of the key tasks in molecular research. Traditional biochemical experiments for protein subcellular localization prediction are expensive and face uncertain time boundary to meet the research demands. Additionally, in the post-genomic era, an exponentially growing number of protein sequences are available for the large-scale proteome projects. As a result, computational methods are required as alternative choice to predict protein subcellular localization automatically and accurately.

The main concern of developing computational method for subcellular localization prediction is to collect information from protein sequence as much as possible to construct informative features. The selection of enrich feature is vital for any classification system. Informative feature representations provide abundant discriminant information to improve prediction accuracy [1]. Up to now numerous predictors have been developed for identifying subcellular localization of proteins with different features which can deal with single-location and multi-location proteins. Cai et al. have used well known representation amino acid composition (AAC) in which feature vector has been constructed by calculating occurrence frequency of

[*] To whom correspondence should be addressed. Email: juliacse06@gmail.com

doi:10.2390/biecoll-jib-2016-288

20 native amino acid residues [2]. Park and Kanehisa have proposed to apply SVM classifier onto five features based on amino acid compositions, amino acid pair compositions and gapped amino acid compositions [3]. 400 dimension vector called Dipeptide composition (DC) has been used for protein subcellular localization prediction where each value of vector represents the occurrence of 2 consecutive residues from protein sequence string [4]. Pseudo-Amino Acid Composition (PseAAC) feature has included not only protein sequence composition but also its order information to predict subcellular location [5-7]. Li et al. have improved the prediction accuracy by combining Gene Ontology (GO) annotations and Amphiphilic pseudo amino acid composition (APseAAC) [8]. A number of classifiers were developed based on Gene Ontology Annotation [8-11]. Xumi et al. have used fusion feature extraction methods based on pseudo amino acid composition (PseAAC), physiochemical property model (PPM), amino acid index distribution (AAID) and N-terminal signal to improve the accuracy [12]. S. Wang and S. Liu have proposed two fusion feature representations by integrating PSSM with DipC and PseAAC, respectively as features and KNN as classifier to predict sub-nuclear location of protein [1].

With informative feature representation researchers also used different machine learning approaches as classifiers like K nearest neighbour (KNN) [1], multi-label K nearest neighbour (ML-KNN) [12], artificial neural network [2], fuzzy-K nearest neighbour (Fuzzy-KNN)[4], support vector machine (SVM) [3, 9, 13].

We have found that, classifiers with single feature representations failed to show significant accuracy. Other researchers, mentioned above also found the similar trend. A closer look at the single feature representations reveal that they contain insufficient discrimination information of protein sequence like as Amino Acid Composition (AAC) feature [1], which leads to classifier confusion and thus worse performance. For constructing informative representation of feature, in our paper, we have proposed two effective feature fusion representations by combining two single feature representations respectively and then apply SVM classifier as a predictor. Though, theoretically, all classification approaches will suffer for insufficiency of information in feature representation, but SVM like classifier expose it more evidently. Moreover, SVM not only provides unique and global solution but also gives potential solution to subcellular localization prediction [3, 9, 13, 21]. At first, we find out Pseudo-Amino Acid Composition (PseAAC), Physicochemical Properties Model (PPM) and amino acid index distribution (AAID) feature representation from protein sequence and then fused both PseAAC and AAID to form a new representation AAIDPAAC and also take account both PPM and PseAAC to construct another representation PPMPAAC. SVM classifier has applied on all of the single and fused features. We have observed that the actual accuracy and locative accuracy of feature fusion representation is greater than single feature representation.

2 Dataset and Feature Extraction Methods

2.1 Dataset

In this paper, we have used Gram-negative bacterial dataset [13, 14] because of its importance in drug discovery. Proteins of this dataset have ≤ 25% pairwise sequence similarity to any other proteins in a same subcellular location. This dataset contains 1,392 unique proteins and 1,456 total proteins with 8 different subcellular locations. We have only considered protein sequences with 20 standard amino acids and excluded sequences containing B, Z, X, J symbols because of their ambiguity [15]. After pre-processing this dataset, we have got the following statistics:

Table 1: Proteins in Gram Negative Dataset after preprocessing.

No. of Unique protein	No. of Locations	No. of 1-locative proteins	No. of 2-locative Proteins	Total proteins
1390	8	1326	64	1454

2.2 Feature Extraction Methods

Features are extracted from protein sequences. Protein sequences are expressed by the 20 native amino acid residues: P = (p_1, p_2, ……….. , p_N), where pi ϵ A and A = 20 amino acids. Initially, for feature extraction, we have used PseAAC, PPM, AAID schemes.

2.2.1 Pseudo-Amino Acid Composition (PseAAC)

Pseudo-Amino Acid Composition [12, 16] feature extraction method not only based on sequence composition but also its sequence order information. The feature vector is expressed as:

$$F = [v_1, v_2, \ldots, v_{20}, v_{20+1}, \ldots\ldots, v_{20+\eta}\,] \quad (\eta < N) \tag{1}$$

Here, first 20 components represent the occurrence frequency of one of the 20 native amino acid and rest of the components represents the protein position information.

The sequence order η-correlation factor of each residue in protein P is defined as:

$$\delta_\eta = \frac{1}{N-\eta}\sum_{i=1}^{N-\eta}\Omega_{i,\,i+\eta} \;; \qquad (\eta < N) \tag{2}$$

Here, if the value of η is 1 then δ_1 is defined as first-tier correlation factor which reflects the sequence order correlation between all the most contiguous residues along a protein chain. If η=2 then the second-tier correlation factor δ_2 reflects the sequence order correlation between all the second most contiguous residues. Again for η=3, the third-tier correlation factor δ_3 that reflects the sequence order correlation between all the 3rd most contiguous residues and so forth.

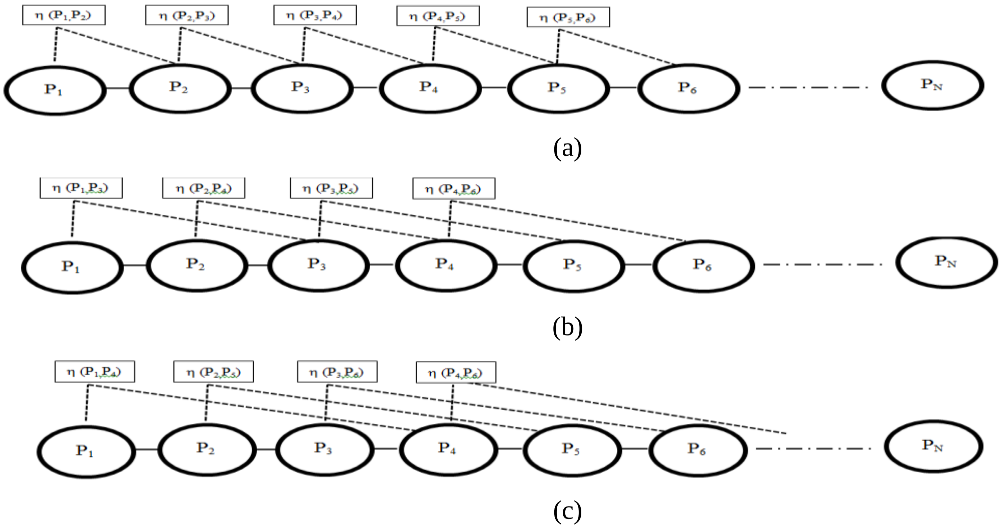

Figure 1: (a) the first-tier, (b) the second-tier, and (c) third-tier sequence order correlation mode along a protein sequence.

$\Omega_{i,\,i+\eta}$ is coupling factor for η most adjacent residues expressed as:

$$\Omega_{i,\, i+\eta} = \frac{1}{3}\left\{\left[\, S_1(P_{i+\eta}) - S_1(P_i)\right]^2 + \left[\, S_2(P_{i+\eta}) - S_2(P_i)\right]^2 + \left[T(P_{i+\eta}) - T(P_i)\right]^2\right\} \quad (3)$$

Here, S_1, S_2 and T are the value of normalized hydrophobicity, hydrophilicity and the side chain mass for amino acid residues. P_i and $P_{i+\eta}$ are i^{th} and $(i+\eta)^{th}$ amino acid residue respectively.

If the η correlation factor is calculated then PseAAC feature vector can be represented as:

$$P_u = \begin{cases} \dfrac{f_u}{\sum_{i=1}^{20} f_i + \omega \sum_{k=1}^{\eta} \delta_k} & (1 \le u \le 20) \\ \dfrac{\omega \delta_{u-20}}{\sum_{i=1}^{20} f_i + \omega \sum_{k=1}^{\eta} \delta_k} & (20 + 1 \le u \le 20 + \eta) \end{cases} \quad (4)$$

Where, weight factor $\omega = 0.05$, f_i is the occurrence frequency of the residues in this protein sequence. The value of η is L_{min}- 1, where L_{min} is the length of the shortest protein sequence in the dataset.

2.2.2 Physicochemical Properties Model (PPM)

In Physicochemical Properties Model, amino acid residues in all proteins are divided into neutral, hydrophobic and polar groups according to their seven physicochemical properties. The seven physicochemical properties are hydrophobicity, normalized Vander Waals volume, polarity, polarizability, charge, secondary structures and solvent accessibility.

According to the seven properties of amino acids residues, compute occurrence frequency of the residues manifested as polar, neutral and hydrophobic in a protein sequence. The calculation formula is [12]:

$$f_{i,\,polar} = \frac{n_{polar}}{N} \quad (5)$$

$$f_{i,neutral} = \frac{n_{neutral}}{N} \quad (6)$$

$$f_{i,\,hydrophobic} = \frac{n_{hydrophobic}}{N} \quad (7)$$

Where, i =1, 2, … , 7 (seven properties), $f_{i,\,j}$ is the frequency of amino acid characterized by polar / neutral / hydrophobic. n_j represents the total number of polar / neutral / hydrophobic characters present in protein sequence. N is the length of the protein sequences.

However, according to this feature extraction model, it creates 21 dimension feature vector for each protein sequence. The distribution situation of amino acids properties are showed in Table 2.

Table 2: Distribution situation of amino acid properties [12]

Property	Polar	neutral	hydrophobic
hydrophobicity	RKEDQN	GASTPHY	CLVIMFW
normalized Vander Waals	GASCTPD	NVEQIL	MHKFRYW
Polarity	LIFWCMVY	PATGS	HQRKNED
Polarizability	GASDT	CPNVEQIL	KMHFRYW
Charge	KR	ANCQGHILMFPSTWYN	DE
secondary structures	EALMQKRH	VIYCWFT	GNPSD
solvent accessibility	ALFCGINW	RKQEND	MPSTHY

2.2.3 Amino Acid Index Distribution (AAID)

Amino Acid Index Distribution (AAID) [12,17] considers the physicochemical value and order of amino acids appeared in the protein sequence to express the protein sequence. In this kind of model, the feature vector of the protein sequence can be represented by the following formula:

$$F_{AAID} = [x_1, x_2, \dots, x_{20};\ y_1, y_2, \dots, y_{20};\ z_1, z_2, \dots\dots, z_{20}] \tag{8}$$

In this case, the first 20 dimension of vector F_{AAID} is the combination of statistical information and physicochemical values. Let, $R_1, R_2, \dots\dots, R_{20}$ represent the 20 natural amino acids A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, and Y respectively and J_i $(i = 1,2, \dots, 20)$ be the amino acid index value of the 20 natural amino acids R_i . Here amino acid index is a set of 20 numerical values representing any of the different physicochemical properties of the 20 amino acids.

By considering the physicochemical property of the amino acid, we can define the feature x_i of amino acid

$$x_i = J_i f_i, \qquad i = 1,2, \dots\dots., 20 \tag{9}$$

where f_i is the frequency of the amino acids R_i in the protein sequence, J_i is the physicochemical values of the amino acids R_i which are as follows: $J_1 = 0.486$, $J_2 = 0.2$, $J_3 = 0.288$, $J_4 = 0.538$, $J_5 = 0.318$, $J_6 = 0.12$, $J_7 = 0.4$, $J_8 = 0.37$, $J_9 = 0.402$, $J_{10} = 0.42$, $J_{11} = 0.417$, $J_{12} = 0.193$, $J_{13} = 0.208$, $J_{14} = 0.418$, $J_{15} = 0.262$, $J_{16} = 0.2$, $J_{17} = 0.272$, $J_{18} = 0.379$, $J_{19} = 0.462$, $J_{20} = 0.161$.

The following 20 dimension feature vectors is 2-order centre distance information, it does not only includes the statistical information and physicochemical values, but also contains position information. The formula is as follows:

$$y_i = \sum_{j=1}^{N_{R_i}} \left(\frac{P_{i,j} - \overline{P_i}}{T} J_i \right)^2 \qquad i = 1,2, \dots\dots, 20 \tag{10}$$

where, N_{R_i} is the total number of amino acid R_i appearing in the protein sequence P, $P_{i,j}$ is the jth position of the amino acid R_i in the sequence, and $\overline{P_i}$ is the mean of the position of amino acid R_i.

Now feature y_i contains the physicochemical information, statistical information and the sequence-order information of amino acids R_i, but it still does not distinguish the protein pairs in some cases. To solve this problem, the 3rd order centre distance z_i of amino acid R_i was introduced, which is defined as

$$z_i = \sum_{j=1}^{N_{R_i}} \left(\frac{P_{i,j} - \overline{P_i}}{T} J_i \right)^3 \qquad i = 1,2, \dots\dots, 20 \tag{11}$$

The 3rd order centre distance information is almost same as 2nd order centre distance except the order number. In this paper, we used the first 40 dimension vectors.

3 Implementation

Firstly we have introduced two feature fusion representation and then the classifier which is used to predict subcellular localization. The procedure of protein subcellular localization prediction is shown by the following Figure 2:

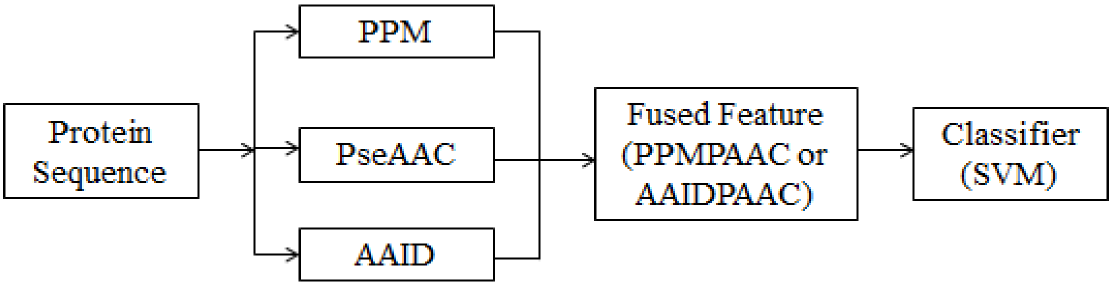

Figure 2: Flowchart of the prediction procedure

3.1 Feature Fusion

Whereas AAID as well as PPM reflect information about physicochemical property attributes of protein sequence, on the other side, PseAAC feature representation takes into account both sequence order and length information about sequence which are not presented by AAID and PPM. As a result, based on the concept of combining sequence related information with physicochemical property for getting rich feature without redundancy we apply feature fusion techniques to produce AAIDPAAC and PPMPAAC by combining PseAAC with AAID and PPM respectively.

To produce fused feature sets we use parallel feature fusion strategy which is formed complex feature vector. Let consider two different feature vectors A and B of same sample ξ with m and n dimension accordingly. If the dimension of two feature vector is same then no problem is occurred. Only, Complex vector λ= A+iB (i is an imaginary unit) is used to represent the combination of A and B and is named parallel combined feature of ξ [18]. But if $m \neq n$, that means feature vector A as well as B different length of features then zero padding is occurred with lower dimension feature space until m = n. For example, if $A = (a_1; a_2; a_3)^T$ and $B = (b_1; b_2)^T$ then the dimension 3≠2. For this reason zero padding is occurred and B turns into $B = (b_1; b_2; 0)^T$. The resultant fused feature is $C = (a_1 + ib_1; a_2 + ib_2; a_3 + i0)^T$. The dimension of parallel fused representation max{dimension of A; dimension of B}. Both A and B belongs to feature space α and β but parallel combined feature space is a unitary space because of the imaginary part. In unitary space, the measurement (norm) can be introduced as follows [18]:

$$\|Z\| = \sqrt{Z^H Z} = \sqrt{\sum_{j=1}^{n}(a_j^2 + b_j^2)} \tag{12}$$

Where, $Z = (a_1 + ib_1, a_2 + ib_2, \ldots, a_n + ib_n)^T$

Following this procedure we develop two fused feature vectors AAIDPAAC and PPMPAAC.

3.2 Support Vector Machine (SVM)

In this work, we use one-against-rest based Support Vector Machine (SVM) in which decision about each subcellular location is taken separately. In this case, each location is transformed to two class labels, either +1 (support the location), or -1 (does not support the location). SVM tries to find the optimal hyper plane in the feature space with maximum margin between the two classes.

For each target protein P_t, decision about j^th location will be made by SVM as:

$$S_j(x^t) = \sum_{i=1}^{n} \alpha_i y_i K(x^i, x^t) + b \tag{13}$$

where, $K(x^i, x^t) = exp(-\gamma\ ||x^i - x^t||^2)$

$y_i \in (+1, -1)$,

$\gamma = \frac{1}{2\sigma^2}$

σ is the width of the function,

α_i is the lagrange multipliers.

The subcellular location of target protein P_t will be predicted as:

$$Loc(x^t) = \bigcup_{j=1}^{d} \{j: S_j(x^t) > 0\} \tag{14}$$

If $Loc(x^t) = \emptyset$, then the number of subcellular locations is set to one and the location is given by:

$$Loc(x^t) = arg_{j} \overset{d}{=} {}_1 minf_j \tag{15}$$

where, f_j is the functional value returned by each SVM classifier. This idea of avoiding zero prediction is adapted from Wan et al. work [9].

4 Result and Discussion

In statistical prediction, two of the most crucial issues are what metrics should be used and what kind of test strategy should be followed. For test strategy selection, we adopt the technique used in [19, 20], and that technique was also followed by other researchers, mentioned in the both works.

4.1 Metrics in Multi-label Systems

In biological context, a protein may exist in more than one location, so, our prediction problem is a multi-label problem. To evaluate the anticipated performance, here we adopted two well defined and widely used metrics from the work [9]: actual accuracy and locative accuracy.

Actual accuracy (AA) is defined as the exact match of classifier's predicted labels with the actual labels of a target protein.

$$AA = \frac{\sum_{i=1}^{N} \Delta[M(P_i),\ L\ (P_i)]}{N_{AA}} * 100 \tag{16}$$

Where, $\Delta[M(P_i), L\ (P_i)] = \begin{cases} 1,\ if\ \ M(P_i)\ \equiv L(P_i) \\ 0,\ otherwise \end{cases}$

$L(P_i)$ represents true label set for ith protein

$M(P_i)$ represents predicted label set for ith protein

N_{AA} is total number of unique proteins

On the other hand, if the predicted label $M(P_i)$ of target protein P_i matches with any label of the true label set $L(P_i)$ then the accuracy is considered as *Locative Accuracy*. It can be defined as-

$$LA = \frac{\sum_{i=1}^{N} \Delta|M(P_i) \cap L\ (P_i)|}{N_{LA}}\ * 100 \tag{17}$$

N_{LA} is total number of locative proteins

4.2 Cross-Validation and Success Rate

In most of the statistical prediction problem, for examining the predictor's strength, independent dataset test, K-fold cross validation and jackknife test are widely used. As mentioned in both [19, 20], only the jackknife test always provide a unique result for a given benchmark dataset but the computational cost is very high compared to the other techniques.

In this study, however, to reduce the computational cost, we adopted the 8-fold cross-validation method, as done by many researchers with support vector machine. In this case, in each pass, seven portions were used as training data, remaining set was used as test data and this process repeats until all the proteins goes to test set. 8-fold was selected as there are only 8 proteins at nucleoid location and more than 8 proteins in other locations in our Gram-negative bacterial dataset. We have ensured that each fold contains at least one protein sequence of each class.

As SVM with RBF kernel was our prediction engine and parameter selection may bias the predictor hence, we need to select optimal combination of γ and regularization coefficient c for obtaining highest accuracy. For achieving better accuracy, we select the parameters from $\gamma = \{2^{-8}, 2^{-7}, \ldots, 2^{0}, \ldots, 2^{8}\}$ and $c = \{2^{-2}, 2^{-1}, 2^{0}, 2^{1}, 2^{2}\}$. Since AAIDPAAC feature outperforms other features with SVM, the corresponding search space in Gram Negative dataset are shown in Figure 3.

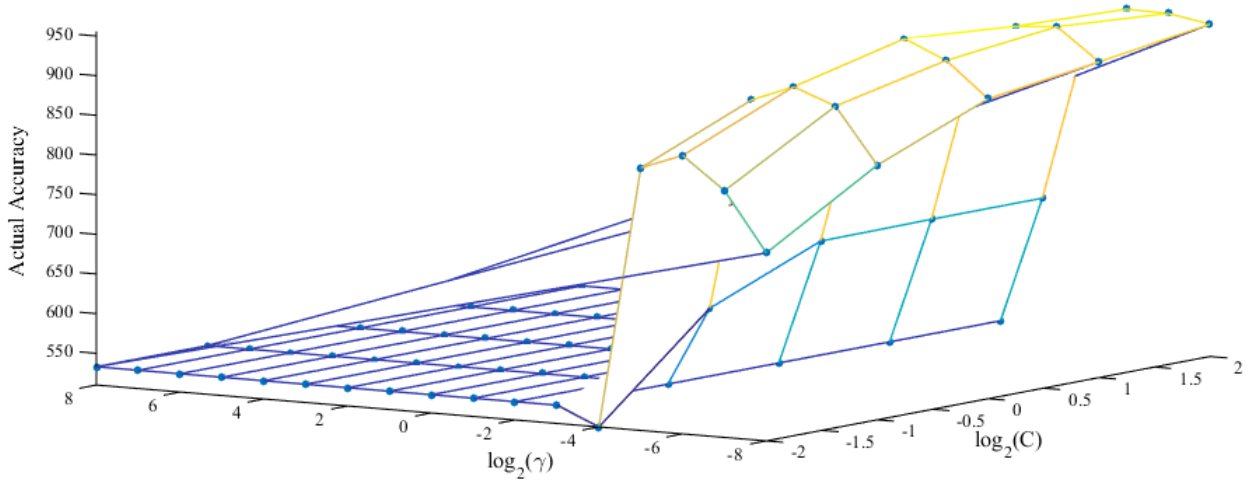

Figure 3: Tuning SVM with AAIDPAAC in Gram Negative Dataset

Here, the experimental result of single features and fused feature sets is given bellow:

Table 3: Performance for different features using SVM

Features	Actual Accuracy (AA) (%)	Locative Accuracy (LA) (%)
PseAAC	65.04	68.5
PPM	61.3	68.8
AAID	65.4	69.25
AAIDPAAC	68.7	70.7
PPMPAAC	67.99	71.05

It can be seen from the following graph Figure 4 that the best actual accuracy 68.7% is achieved by AAIDPAAC which is 3.5%, 7% and 3% greater than PseAAC, PPM and AAID respectively. AAIDPAAC as the fuse form of AAID and PseAAC feature representation has achieved this best performance for $\gamma = 2^{-6}$, c= 2^0. The second highest actual accuracy 67.99% is also achieved by feature fusion representation PPMPAAC.

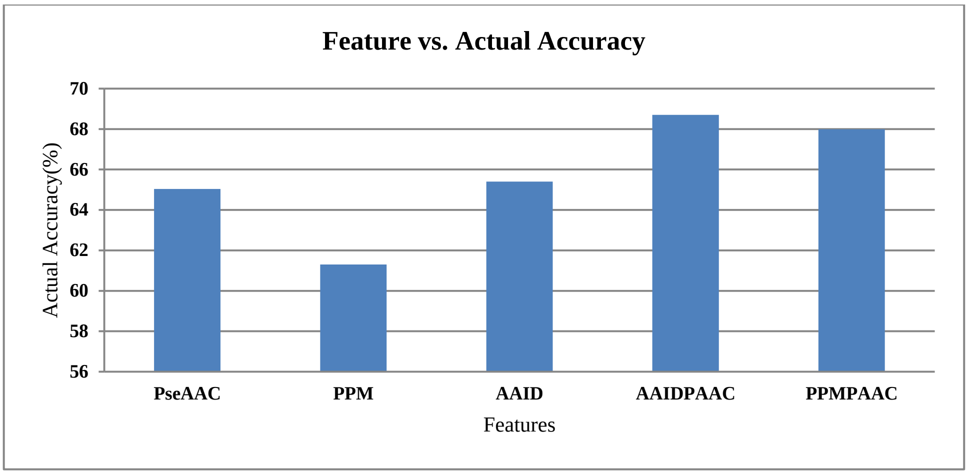

Figure 4: Actual Accuracy (%) for different Features

In the following Figure 5, we observe that best locative accuracy 71.05% is achieved by PPMPAAC which is also a feature fusion representation. The locative accuracy of PPMPAAC is 2.5%, 2.2% and 2% heigher than PseAAC, PPM and AAID respectively.

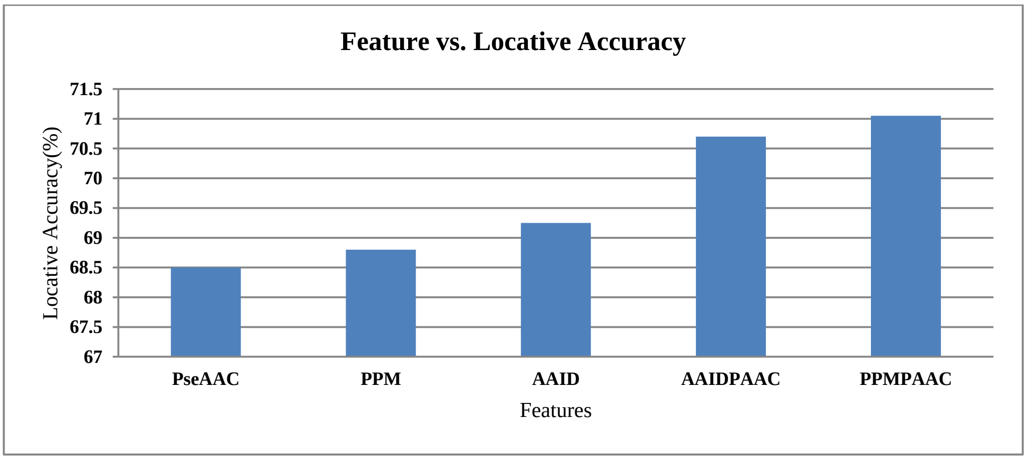

Figure 5: Locative Accuracy (%) for different Features

The feature fusion representations perform better actual and locative accuracy than single features as illustrated in Figure 4 and Figure 5.

5 Conclusion

In this paper, we present an approach to improve the accuracy for protein subcellular localization prediction. In protein subcellular localization prediction using machine learning technique, informative feature extraction methods from protein sequence mostly affect the performance. If inappropriate, noisy and less informative feature extraction methods are selected for classification then the accuracy is decreased instead of increasing. In this paper, we have used SVM, one of the widely used machine learning techniques with five distinct features for subcellular localization prediction. Among them, three is single feature representations and two is feature fusion representations. From the result, we have seen that feature fusion representation performs better than single features. In this paper, we have evaluated the performance on Gram negative bacterial dataset. The actual and locative accuracy using fusion representation such as AAIDPAAC and PPMPAAC are at least 3% and 2% higher respectively in comparison to single feature representation. Nevertheless, still it is a challenge to achieve higher accuracy by using more efficient methods and it is the important part of our future work.

References

[1] S. Wang and S. Liu. Protein sub-Nuclear localization based on effective fusion representations and dimension reduction algorithm LDA. International Journal of Molecular Sciences, 16:30343–30361, Dec. 2015.

[2] Y.-D. Cai, X.-J. Liu, and K.-C. Chou. Artificial neural network model for predicting protein subcellular location. Computers and Chemistry, 26:179–182, Jan. 2002.

[3] K.-J. Park and M. Kanehisa. Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs. Bioinformatics, 19(13):1656–1663, 2003.

[4] Y. Huang and Y. Li. Prediction of protein subcellular locations using fuzzy k-NN method. Bioinformatics, 20(1):21–28, 2004.

[5] H.-B. Shen and K.-C. Chou. Nuc-PLoc: a new web-server for predicting protein subnuclear localization by fusing PseAA composition and PsePSSM. Protein Engineering, Design & Selection, 20(11):561–567, 2007.

[6] L. Li, S. Yu, W. Xiao, Y. Li, M. Li, L. Huang, X. Zheng, S. Zhou, and H. Yang. Prediction of bacterial protein subcellular localization by incorporating various features into Chou's PseAAC and a backward feature selection approach. Biochimie, 104:100–107, Jun. 2014.

[7] C. Huang and J. Yuan. Using radial basis function on the general form of Chou's pseudo amino acid composition and PSSM to predict subcellular locations of proteins with both single and multiple sites. BioSystems, 113(1):50– 57, Apr. 2013.

[8] L. Li, Y. Zhang, L. Zou, C. Li, B. Yu, X. Zheng, and Y. Zhou. An Ensemble Classifier for Eukaryotic Protein Subcellular Location Prediction Using Gene Ontology Categories and Amino Acid Hydrophobicity. PLoS ONE, 7(1), Jan. 2012.

[9] S. Wan, M.-W. Mak, and S.-Y. Kung. mGOASVM: Multi-label protein subcellular localization based on gene ontology and support vector machines. BMC Bioinformatics, 2012.

[10] C.-S. Yu, C.-W. Cheng, W.-C. Su, K.-C. Chang, S.-W. Huang, J.-K. Hwang, and C.-H. Lu. CELLO2GO: A Web Server for Protein subCELlular LOcalization Prediction with Functional Gene Ontology Annotation. PLOS ONE, 9(6), Jun. 2014.

[11] X. Wang, J. Zhang, and G.-Z. Li. Multi-location gram-positive and gram-negative bacterial protein subcellular localization using gene ontology and multi-label classifier ensemble. BMC Bioinformatics, 16(12), 2015.

[12] X. Qu, Y. Chen, S. Qiao, D. Wang, and Q. Zhao. Predicting the subcellular localization of proteins with multiple sites based on multiple features fusion. presented at the ICIC, 2014, :456–465.

[13] A. Dehzangi, R. Heffernan, A. Sharma, J. Lyons, K. Paliwal, and A. Sattar. Gram-Positive and Gram-Negative Protein Subcellular Localization by Incorporating Evolutionary-based Descriptors into Chou's General PseAAC. Journal of Theoretical Biology, :284–294, 2015.

[14] X. Xiao, Z.-C. Wu, and K.-C. Chou. A Multi-Label Classifier for Predicting the Subcellular Localization of Gram-Negative Bacterial Proteins with Both Single and Multiple Sites. PLoS ONE, 6(6), Jun. 2011.

[15] IUPAC — IUB Commission on Biochemical Nomenclature, A One-Letter Notation for Amino Acid Sequences (Definitive Rules). Pure Appl. Chem., 34(4):639–646, 1971.

[16] K.-C. Chou. Prediction of Protein Cellular Attributes Using Pseudo-Amino Acid Composition. PROTEINS: Structure, Function, and Genetics, 43:246–255, 2001.

[17] S.-W. Zhang, L.-Y. Hao, and T.-H. Zhang. Prediction of Protein–Protein Interaction with Pairwise Kernel Support Vector Machine. International Journal of Molecular Sciences, 15:3220–3233, Feb. 2014.

[18] J. Yanga, J. Yanga, D. Zhangb, and J. Lua. Feature fusion: parallel strategy vs. serial strategy. Pattern Recognition, 36:1369 – 1381, 2003.

[19] W.-R. Qiu, B.-Q. Sun, X. Xiao, Z.-C. Xu, and K.-C. Chou. iPTM-mLys: identifying multiple lysine PTM sites and their different types. Bioinformatics, 32(20):3116–3123, 2016.

[20] Y. Xu, J. Ding, L.-Y. Wu, and K.-C. Chou. iSNO-PseAAC: Predict Cysteine S-Nitrosylation Sites in Proteins by Incorporating Position Specific Amino Acid Propensity into Pseudo Amino Acid Composition. PLOS ONE, 8(2), Feb. 2013.

[21] L. Li, H. Kuang, Y. Zhang, Y. Zhou, K. Wang, and Y. Wan. Prediction of eukaryotic protein subcellular multilocalisation with a combined KNN-SVM ensemble classifier. Journal of Computational Biology and Bioinformatics Research, 3(2):15–24, Feb. 2011.