Acta Biotheor (2013) 61:481–497
DOI 10.1007/s10441-013-9197-1

ORIGINAL ARTICLE

APSLAP: An Adaptive Boosting Technique for Predicting Subcellular Localization of Apoptosis Protein

Vijayakumar Saravanan · P. T. V. Lakshmi

Received: 30 January 2013 / Accepted: 16 August 2013 / Published online: 28 August 2013

Abstract Apoptotic proteins play key roles in understanding the mechanism of programmed cell death. Knowledge about the subcellular localization of apoptotic protein is constructive in understanding the mechanism of programmed cell death, determining the functional characterization of the protein, screening candidates in drug design, and selecting protein for relevant studies. It is also proclaimed that the information required for determining the subcellular localization of protein resides in their corresponding amino acid sequence. In this work, a new biological feature, class pattern frequency of physiochemical descriptor, was effectively used in accordance with the amino acid composition, protein similarity measure, CTD (composition, translation, and distribution) of physiochemical descriptors, and sequence similarity to predict the subcellular localization of apoptosis protein. AdaBoost with the weak learner as Random-Forest was designed for the five modules and prediction is made based on the weighted voting system. Bench mark dataset of 317 apoptosis proteins were subjected to prediction by our system and the accuracy was found to be 100.0 and 92.4 %, and 90.1 % for self-consistency test, jack-knife test, and tenfold cross validation test respectively, which is 0.9 % higher than that of other existing methods. Beside this, the independent data (N151 and ZW98) set prediction resulted in the accuracy of 90.7 and 87.7 %, respectively. These results show that the protein feature represented by a combined feature vector along with AdaBoost algorithm holds well in effective prediction of subcellular localization of apoptosis proteins. The user friendly web interface "APSLAP" has

Electronic supplementary material The online version of this article (doi:10.1007/s10441-013-9197-1) contains supplementary material, which is available to authorized users.

V. Saravanan · P. T. V. Lakshmi (✉)
Centre for Bioinformatics, School of Life Sciences, Pondicherry University, RK Nagar, Kalapet, Pondicherry 605014, India
e-mail: lakanna@bicpu.edu.in

V. Saravanan
e-mail: brsaran@bicpu.edu.in

been constructed, which is freely available at http://apslap.bicpu.edu.in and it is anticipated that this tool will play a significant role in determining the specific role of apoptosis proteins with reliability.

Keywords AdaBoost · Apoptosis protein · Jack-knife test · Physio-chemical parameres · Random forest · Subcellular localization · Web-server

1 Introduction

The term apoptosis is generally used to describe the morphologically distinct form of cell death (Kerr et al. 1972). The process of apoptosis is extremely complex and involves various cascades of molecular events, including protein signaling. Thus, apoptotic proteins have a distinct role in the development and the homeostasis of an organism (Raff 1998). Studies on apoptosis suggest that the alterations in the cell survival and or blocking the apoptosis contribute to the pathogenesis of a number of human diseases including cancer, autoimmune disease, neurodegenerative disorders, and acquired immunodeficiency syndrome (Thompson 1995). An insight on to proteins involved in apoptotic process could therefore help in understanding the mechanism and provide various therapeutic targets. It is also evident that the knowledge about the subcellular localization of apoptosis protein (SLAP) aids in understanding the underlying principle behind the apoptosis mechanism (Schulz et al. 1999; Suzuki et al. 2000). Although the unknown localization of an apoptosis protein can be determined by various experimental techniques, they are both expensive and time consuming. Moreover, with the rapid increase of protein sequences in databank with unknown function, it is essential to develop a fast and reliable computational method to predict the SLAP from their native amino acid residues.

When compared to the immense work on protein subcellular localization prediction (Chou and Shen 2007), studies on apoptosis protein subcellular localization predictions were limited. It may be due to the availability of limited number of experimentally proved apoptosis protein in the database. On behalf of that, various methods have been reported on prediction of SLAP. Initial attempt by Zhou and Doctor (2003) adopted covariant discriminant based on amino acid composition to predict the subcellular localization of 98 apoptosis proteins and achieved 72.5 % accuracy through jack-knife test. With the same dataset Huang and Shi (2005) used support vector machine (SVM) and sqrt-amino acid composition, Bulashevska and Eils (2006) used single Bayesian classifier, and Chen and Li (2004) adopted measure of diversity algorithm. They all achieved an overall accuracy of 90.8, 85.7, and 90.8 %, respectively, with the jack-knife test. Zhang et al. (2006) introduced a new dataset ZW225 and used weighted SVM to achieve 83.1 % accuracy in jack-knife test. Meanwhile, Chen and Li (2007) constructed a new dataset of 317 apoptotic proteins covering six subcellular compartments and achieved 82.7 and 84.2 % accuracy by adopting increment of diversity and increment of diversity by SVM, respectively. This dataset of 317 proteins were further used by others to evaluate their methods. Ding and Zhang (2008), by using an ensemble classifier, Fuzzy K- nearest neighbor (FKNN), achieved an overall accuracy of 90.9 %. Lin et al. (2009) achieved 91.1 % accuracy by wavelet transformation method. With the new mode of pseudo amino acid composition Kandaswamy et al. (2010) achieved 90.3 % accuracy. On other hand, Gu et al. (2010) by using ensemble classifier along with FKNN achieved a maximum of 91.5 % accuracy, which was revealed to be highest among all methods including the recent attempt by Yu et al. (2012), that employed amino acid substitution and auto covariance method to yield 90.0 % accuracy, and Liao et al. (2011), that employed pseudo amino acid composition and tri-peptides to yield 91.2 % accuracy, through the jack-knife test. Thus on an overview from all the above mentioned methods, the representative feature vectors were solely derived from the amino acid sequence of the protein, suggesting that the SLAP is predictable to a substantial extent, if a good vector representation of protein could be made in hands with a powerful mathematical method for prediction. It is also expected that improved feature representation from amino acid properties along with a potential mathematical method could have a significant impact in achieving better prediction accuracy of SLAP.

There are several methods exist to represent biological sequences (Yau et al. 2008; Yu et al. 2010, 2011, 2013; Deng et al. 2011), in this work priority is given to the protein global sequence feature unlike the N-terminal or C–terminal regions alone (Matsuda et al. 2005; Tantoso and Li 2008). A new biological feature, class pattern frequency of physiochemical descriptor, was effectively used in accordance with the amino acid composition, protein similarity measure, CTD (composition, translation, and distribution) of physiochemical descriptors, sequence similarity, and signal peptide prediction to predict the subcellular localization of apoptosis protein. Ensemble learner AdaBoost (Freund and Schapire 1997) algorithm was employed, wherein jack-knife test, tenfold cross validation, and independent dataset test were carried out on CL317 to evaluate our method.

2 Materials and Methods

In this study, the dataset constructed by Chen and Li (2007) (CL317), which contained 317 proteins classified into six compartments as Cytoplasm, Mitochondrion, Nucleus, Membrane, Secreted, and Endoplasmic reticulum each containing 112, 34, 52, 55, 17, and 47 proteins, respectively were used. Independent data set was generated from UNIPROT database (UniProt release 2011–2012-Dec 14) by same method as described in Chen and Li (2007). Further, the proteins in the training set CL317 were removed and the remaining sequences were subjected to protein culling program PISCES (Wang and Dunbrack 2003) to remove sequences that had sequence similarity of >40 % within the dataset as well as within the training dataset CL317, in order to avoid bias in the prediction result. The final independent dataset (N151) resulted in a total of 151 proteins (Table 1) (*supplementary S1*). Apart from the newly developed independent set, dataset ZD98 constructed by Zhou and Doctor (2003) consisting of 98 apoptosis proteins classified into four classes was also used as independent dataset to evaluate the proposed method.

2.1 Feature Vector Representation

The input features used for prediction comprises of amino acid composition (AA), class pattern frequency of physiochemical descriptor, PSM (protein similarity measure), CTD (composition, translation, and distribution) of physiochemical descriptors, hybrid vector, signal peptide prediction and sequence similarity. The feature vector calculation for amino acid composition (AA_C), CTD of physiochemical descriptors (CTD), and protein similarity measure (S_m) were calculated as described by Chou (1995), Dubchak et al. (1995), and Carr et al. (2010), respectively.

2.2 Amino Acid Composition

Amino acid composition calculates the percentage of twenty standard amino acid occurrences in a protein. The protein sequence is represented by a 20-D feature vector. If index *k* was used to represent the amino acid in the protein sequence, then the feature vector AA_C was calculated as follows:

$$AA_C = [AA_1..AA_k..AA_{20}] \rightarrow 20\boldsymbol{D} \tag{1}$$

$$AA_k = \frac{\sum k}{L_p} \tag{2}$$

where, AA_K was the percentage of amino acid *k* occurring in a protein and L_p was the length of the protein.

2.3 Protein Similarity Measure

Characterization of protein relatedness using feature vector as reported by Carr et al. (2010) was adopted to represent the protein as 60-D vector. Three measures compositional, centroidal, and distributional were used to construct the feature vector. For each parameter first a theoretical mean and theoretical variance were calculated. Then the feature vector was computed by

$$M_i = (P_i - mP_i)/\sqrt{vP_i} \tag{3}$$

where M denotes measure; $i = 1$, 2, and 3; P_i, the parameter value; mP_i, the theoretical mean; and vP_i, the theoretical variance. The parameter value, theoretical

Table 1 Independent dataset

Location	Number of proteins
Cytoplasm	54
Secreted	7
Nucleus	27
Mitochondria	29
Endoplasmic reticulum	10
Membrane	24
Total	151

mean, and theoretical variance were calculated by same method as described by Carr et al. (2010). The three measures were calculated for all the standard 20 amino acid and the resulting feature vector S_m is represented as 60-D vector as shown below,

$$S_m = \left[M_{1j}, M_{2j}, M_{3j}\right] \rightarrow 60\boldsymbol{D}; \quad j = 1, 2, 3 \ldots 20 \tag{4}$$

2.4 CTD of Physiochemical Descriptors

The protein sequences were encoded as described in the calculation of class pattern frequency of physiochemical descriptor section. The compositional (C_s), transitional (T_{xy}), and distribution (D_s), of the physiochemical parameters were calculated as follows,

$$C_S = n_S/L_p; \quad s = 1, 2, 3 \tag{5}$$

where, n_S is the number of s in the encoded sequence and L_p is the length of the protein sequence.

$$T_{xy} = \frac{n_{xy} + n_{yx}}{L_p - 1}; \quad xy = [12], [13], [23] \tag{6}$$

where, n_{xy} is number of dipeptide encoded as “xy” and “yx” respectively; and L_p is the length of the protein sequence.

For each descriptor in Table 2, five distributions were assigned; position percentage of first residue occurrence in the entire sequence, position percentage of 25, 50, 75, and 100 % residue occurrence in the entire sequence. So, the distribution D_x for the descriptor E_i is calculated as follows,

$$E_i 1 D_x = \frac{P_1}{L_p}; \tag{7}$$

$$E_i 25 D_x = \frac{P_{25}}{L_p}; \tag{8}$$

$$E_i 50 D_x = \frac{P_{50}}{L_p}; \tag{9}$$

$$E_i 75 D_x 1 = \frac{P_{75}}{L_p}; \tag{10}$$

$$E_i 100 D_x 1 = \frac{P_{100}}{L_p}; \quad i = 1, 2, \ldots, 7; \quad x = 1, 2, 3 \tag{11}$$

where P_1, P_{25}, P_{50}, P_{75}, and P_{100}we position of first occurrence of x, position of 25 % occurrence of x, position of 50 % occurrence of x, position of 75 % occurrence of x, and position of 100 % occurrence of x, respectively; and L_p is the length of the protein sequence. The composition, translational, and distributional values were calculated for all the seven descriptors and the resulting feature vector CTD was represented as

Table 2 Details of the physiochemical descriptor proposed by Dubchak et al. (1995)

Physiochemical property	Class one	Class two	Class three
Hydrophobicity	Polar	Neutral	Hydrophobicity
	R, K, E, D, Q, N	G, A, S, T, P, H, Y	C, L, V, I, M, F, W
Normalized van der Waals volume	0–2.78	2.95–4.0	4.03–8.08
	G, A, S, T, P, D, C	N, V, E, Q, I, L	M, H, K, F, R, Y, W
Polarity	4.9–6.2	8.0–9.2	10.4–13.0
	L, I, F, W, C, M, V, Y	P, A, T, G, S	H, Q, R, K, N, E, D
Polarizability	0–1.08	0.128–0.186	0.219–0.409
	G, A, S, D, T	C, P, N, V, E, Q, I, L	K, M, H, F, R, Y, W
Charge	Positive	Neutral	Negative
	K, R	A, N, C, Q, G, H, I, L, M, F, P, S, T, W, Y, V	D, E
Secondary structure	Helix	Strand	Coil
	E, A, L, M, Q, K, R, H	V, I, Y, C, W, F, T	G, N, P, S, D
Solvent accessibility	Buried	Exposed	Intermediate
	A, L, F, C, G, I, V, W	P, K, Q, E, N, D	M, R, S, T, H, Y

$$CTD = \left[Ci_{s[1,2,3]}, Ti_{xy[12,23,13]}, Ei_{[1,2,3,4,5]}\right] \rightarrow 147\boldsymbol{D}; \quad i = 1,2,3,4,5,6,7 \tag{12}$$

2.5 Class Pattern Frequency of Physiochemical Descriptor

To derive the class pattern frequency of physiochemical descriptor, the descriptors developed by Dubchak et al. (1995) were utilized (Table 2). Since Hydrophobicity, Normalized van der Waals volume, Polarity, Polarizability, Charge, Secondary structure, and Solvent accessibility parameters were widely used (Binkowski et al. 2003; Saravanan and Lakshmi 2013; Shen and Chou 2006) to represent the physiochemical nature of the protein, all the seven physiochemical parameters were considered. Twenty standard amino acids were first divided into three classes according to the corresponding physiochemical attribute. Each amino acid was then encoded by 1, 2 or 3, according to the class it belongs to. The numbers of possible pattern of length 3 for three classes were determined using, $\mathrm{n}^{\mathrm{r}} = 3^3 = 27$ where ‘*n*’ represents the number of classes and ‘*r*’ represents the pattern length. The 27 class-pattern for the single descriptor would then be

$$P_c = \left\{ \begin{array}{l} (111),(112),(113),(121),(122), \\ (123),(131),(132),(133),(211), \\ (212),(213),(222),(223),(231), \\ (232),(233),(311),(312),(313), \\ (321),(322),(323),(331),(332),(333) \end{array} \right\} \tag{13}$$

If index ‘i’ is used to represent the pattern in the set P_C and D_1, D_2, D_3 … D_7 to represent the seven descriptors, then the feature vector C_P of 189-D (27 patterns per descriptor, so 27X7 = 189) could be calculated as follows,

$$C_p = \left[D_1 P_{C1} \ldots D_1 P_{C27}, D_i P_{Cj} \ldots D_7 P_{C27}\right]; \quad i = 1, 2, ..7; j = 1, 2, ..27 \rightarrow 189\boldsymbol{D} \tag{14}$$

$$P_{Ci} = \sum i / \left({}^{L_P}/_3\right) \tag{15}$$

where, P_{Ci} is frequency of pattern i. and L_P is the length of the protein sequence. Similarly, dipeptides and tetra peptides were also calculated.

2.6 Hybrid vector

Hybrid vector is the combination of all the above mentioned feature vectors and it is represented as *Hyb*,

$$Hyb = [AA_C, C_P, S_m, CTD] \tag{16}$$

2.7 Algorithm

AdaBoost proposed by Freund and Schapire (1997) was used to construct a strong classifiers based on the linear combination of simple weak classifiers. The main property of AdaBoost is that it converges to the logarithm of likelihood ratio with good generalization property (Schapire and Singer 1999). Initially the AdaBoost picks the learner that classifies more data accurately, and then to increase the significance, the misclassified samples were re-weighted. The re-weighting was done for *T* rounds; until it finds an optimal base classifier h_t and ensemble it to form a strong classifier. The AdaBoost algorithm is as follows (Freund and Schapire 1996).

Let D, be the training set

$$D = (x_1, y_1), \ldots, (x_m, y_m); \quad x_i \in X, y_i \in \{-1, 1\} \tag{17}$$

Initialize weights $D_1(i) = 1/m$

For $t = 1, \ldots T$:

1. Call weak Learner, which returns the weak classifier h_t: $X \rightarrow \{1, 1\}$ with minimum error with respect to distribution D_t;
2. Choose $\alpha_t \in R$,
3. Update

$$D_{t+1}(i) = D_t(i) \exp\left(-\frac{\alpha_t y_i h_t(x_i)}{z_t}\right) \tag{18}$$

where Z_t is a normalization factor chosen so that D_{t+1} is a distribution;

Output the Strong Classifier:

$$H(x) = sign\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right) \tag{19}$$

AdaBoost generally applies to handle two-class problems. Since this work deals with proteins of six locations, it becomes a multiclass problem. In fact, there are two general methods available to handle the multiclass problem “one-versus-other” and "all-versus-all" (Ding and Dubchak 2001). In this work "All-versus-All" method was employed to overcome the multiclass problem. Since, weak classifier plays an important role in the performance of AdaBoost a special decision tree, Random Forest (Breiman 2001) was selected as a weak classifier. Seven modules were designed Fig. 1, in which five AdaBoost modules were designed for each Amino acid composition, class pattern frequency of physiochemical descriptor, protein similarity measure, CTD of physiochemical descriptor, and hybrid feature, and were named as ADA_AA, ADA_CPF, ADA_PSM, ADA_CTD and ADA_HYB, respectively. Two other modules such as signal peptide prediction (Petersen et al. 2011) and BLAST (basic local alignment search tool) (Altschul et al. 1990) were named as SP and BLAST, respectively. Weighted voting scheme was employed to derive the prediction from these modules. Since "All-versus-All" strategy was adopted, two-way classifiers were trained between all the classes. For instance, a four class problem will have six two-way classifiers, $K(K–1)/2$, where K is the total number of classes. In this case there were 15 two-way classifiers, as $K = 6$. When query sequence was tested against these classifiers, the one that belong to particular class will get the maximum vote. Thus, the maximum possible vote for a class will be $K - 1$. Since there was more chance of getting a tie when vote alone was considered, the sum of probability value (Pw) from the AdaBoost predictor was taken into account to evaluate the majority. Further these values were normalized to 1 to calculate the final confident score of a class (location). SignalP 4.0 (Petersen et al. 2011) program was used to predict the signal peptide if any in the query sequence. The weight for the signal peptide module (*SP*) was calculated as follows,

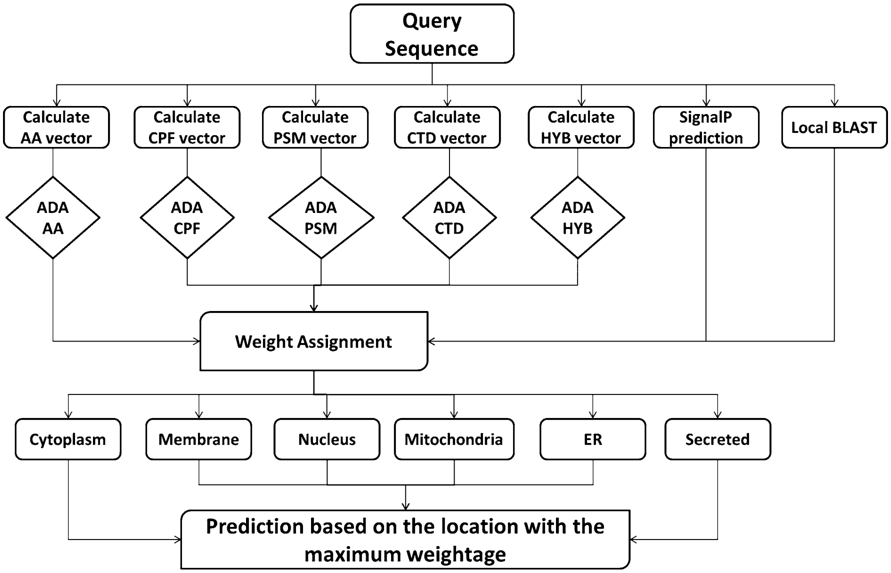

Fig. 1 Architecture of APSLAP

```
If (Secreted){
    SP = 1 for Secretory class
    SP = 0 for all other class
} else {
    SP = 0 for Secretory class
    SP = 1 for all other class
}
```

The D-cutoff for SignalP network was set to 0.6, in order to weigh the prediction with higher confidence. For the BLAST module, the query sequence was subjected to Smith–Waterman sequence alignment (Altschul et al. 1990) with custom made eukaryotic protein database of decided single subcellular location of cytoplasm, secretory, membrane, nucleus, endoplasmic reticulum, and mitochondrion, respectively. The identity score of the query sequence to each location was then normalized to 1 to calculate the final confident score of a class (location).

$$B_x = \frac{identity}{length\ of\ query\ sequence} \tag{20}$$

where B_x is the BLAST score, and x is the six different classes.

The final prediction was made with respect to the confident score of each of the six classes.

$$C_{Score(x)} = \left(\sum P_{Wax} + SP_x + B_x/7\right) \tag{21}$$

$$P_L = max\left\{\begin{matrix} c_{score(cytoplasm)} \\ c_{score(secretory)} \\ c_{score(ER)} \\ c_{score(membrane)} \\ c_{score(mitochondria)} \\ c_{score(nucleus)} \end{matrix}\right\} \tag{22}$$

where C_{score} is the confident score of respective x classes (subcellular location), a the five different AdaBoost modules, and P_L the final predicted location.

2.8 Evaluation methods

Three different evaluation methods have been employed, self-consistency test, jackknife test, and independent data set test, to evaluate our method. Since AdaBoost is considered to be sensitive to the noisy data and likely to over fit, compared to other algorithms (Dietterich 2000), tenfold cross validation test was also carried-out to evaluate the performance of AdaBoost modules. A detailed receiver operator characteristic curve (ROC) and its corresponding area under the curve for each of the five modules were computed and evaluated. The overall accuracy O_A, sensitivity S_{ex}, specificity S_{px}, and Mathew's correlation coefficient MCC_x were calculated as follows,

$$S_{ex} = TP_x/(TP_x + FN_x) \tag{23}$$

$$S_{px} = TN_x/(TN_x + FP_x) \tag{24}$$

$$O_A = \Sigma_x TP_x/N \tag{25}$$

$$MCC_x = \frac{TP_x TN_x - FP_x FN_x}{\sqrt{(TP_x + FP_x)(TP_x + FN_x)(TN_x + FP_x)(TN_x + FN_x)}} \tag{26}$$

where $\mathrm{TP_x}$ is the total number of *x*th subcellular location correctly reported as positive, $\mathrm{TN_x}$ is the total number of other subcellular location correctly reported, $\mathrm{FN_x}$ is the total number of *x*th subcellular location recognized as other subcellular location, $\mathrm{FP_x}$ is the total number of other subcellular location reported as *x*th subcellular location, and *N* is the total number of proteins sequences.

All the calculations and coding were done using PERL script. WEKA (Hall et al. 2009) java classes were used for the implementation of AdaBoost, generating AUC values, and ROC curves.

3 Results and Discussion

The protein sequences of the dataset CL317 and independent data set were first converted into numerical values (*supplementary S2*) of corresponding feature vector, as described in the method section. Self-consistency, jack-knife test, and tenfold cross-validation test were performed to each of the five AdaBoost modules (Table 3). Since support vector machine (SVM) and AdaBoost is considered successful classification technique (Zhang and Gu 2006) and also to compare the performance, self-consistency and jack-knife tests were also carried out using SVM (radial basis function kernel) as classifier on CL317 dataset (Table 4). The self-consistency test revealed that after training, the AdaBoost and SVM was capable of discriminating the relation between the subcellular locations through various features. However, AdaBoost achieved 100 % accuracy in all modules except AA module, whereas SVM suffered achieving 100 % in PSM and CTD modules. Initially, the new feature representation, ADA_CPF with di, tri, and tetra peptide representation was subjected to jack-knife cross validation on CL317 dataset, which resulted in over-all accuracy of 83.4, 86.9, and 84.7 %, respectively. Also, a tenfold cross-validation test was performed on ADA_CPF with di, tri, and tetra-peptide achieved an over-all accuracy of 78.7, 81.1, and 80.9 %, respectively. Since the overall accuracy of jack-knife as well as tenfold cross-validation tests were in favor of ADA_CPF tri-peptide representation, for the ADA_CPF module tri-peptide representation were considered and di and tetra peptide were ruled out (this was not carried out with SVM). In jack-knife test, model was created with the $N - 1$. sets of proteins for all the five modules and the prediction was made to predict the *N*th protein. Each protein data was left from the model once and predicted, which resulted in 92.4 % (291/315 = 0.924) accuracy for our method based on AdaBoost and 90.1 % (284/315 = 0.901) based on SVM. The accuracy of the jack-knife test for the different modules using AdaBoost and SVM (Tables 3, 4) reveals that, though the accuracy of SVM_AA and SVM_CPF was significantly more than ADAB_AA and ADAB_CPF, the accuracy of other modules were low. Since the proposed method derives final prediction based on the results from all modules, the imbalanced performance of SVM_PSM and SVM_CTD modules influenced the SVM-based prediction system to lag behind the AdaBoost-based one on jack-knife accuracy with and without SP and BLAST modules of the proposed prediction system. Subsequently, the accuracies of AdaBoost based modules were consistent on jack-knife test (Table 3). Hence for the final prediction system, AdaBoost-based modules were considered instead of SVM-based modules. Our method, without the signal peptide and BLAST module, achieved 90.15 % (AdaBoost based) accuracy and was perhaps higher than all other methods, except Gu et al. (2010). With the signal peptide module and BLAST module our method (AdaBoost based) reached an overall accuracy of 92.4 %, which is significantly higher than that of others as well as Gu et al. (2010) method (Table 5). The sensitivity of our method for the location cytoplasm and secretory was higher than that of any others. On comparing Mathew's correlation coefficient (Table 6), which is the balanced measure to evaluate the correlation coefficient between the observed and predicted binary classifications, our method performed better than that of the other existing methods with higher accuracy. Dataset constructed by Zhou and Doctor (2003) (ZD98), which contained 98 proteins, were also subjected to jack-knife cross validation and the results were compared with other methods (Table 7). The overall accuracy on Z98 dataset reached 94.9 %, with 100 % accuracy in predicting the membrane proteins. This suggests that the proposed method was consistent on different dataset and competing with other methods.

Since AdaBoost is considered to be sensitive to the noisy data and likely to be over fit (Dietterich 2000), a tenfold cross-validation was performed (Table 3). The accuracy of different modules on tenfold cross-validation was greater than 80 %, which signifies the consistent performance of proposed modules. Among five different AdaBoost modules, ADA_HYB accuracy was higher (87.7 %), whereas proposed method with SP and BLAST module achieved 90.2 % accuracy. Further, a detailed ROC curve analysis was performed and corresponding AUC values were calculated to validate the AdaBoost modules. The ROC plots (sensitivity vs. 1-specificity graph) of five different modules (*supplementary S3*) for each class (six

Table 3 Performance of different Ada-boost modules on CL317 dataset

Module	Self-consistency (%)	Jack-knife (%)	tenfold cross validation
ADAB_AA	98.9	84.7	80.1
ADAB_CPF	100.0	86.9	81.1
ADAB_PSM	100.0	89.8	85.7
ADAB_CTD	100.0	86.3	80.5
ADAB_HYB	100.0	89.8	87.7
Prediction system without BLAST and SP module	100.0	90.2	88.4
Prediction system with BLAST and SP module	100.0	92.4	90.2

Table 4 Performance of different SVM modules on CL317 dataset

Module	Self-consistency (%)	Jack-knife (%)
SVM_AA	100.0	92.1
SVM_CPF	100.0	90.5
SVM_PSM	98.9	79.4
SVM_CTD	98.9	76.6
SVM_HYB	100.0	87.3
Prediction system without BLAST and SP module	100.0	87.7
Prediction system with BLAST and SP module	100.0	90.1

Table 5 Comparison of different methods by jack-knife test on CL317 dataset

Method	C	B	M	S	N	E	Overall accuracy (%)
ID[a]	81.3	81.8	85.3	88.2	82.7	83.0	82.7
ID_SVM[b]	91.1	89.1	79.4	58.8	73.1	87.2	84.2
DF_SVM[c]	92.9	85.5	76.5	76.5	**93.6**	86.5	88.0
FKNN[d]	92.0	89.1	85.3	76.5	92.3	93.7	90.2
FKNN[e]	93.8	**92.7**	82.4	76.5	90.4	93.6	90.9
PseAAC_SVM[f]	93.8	90.9	85.3	76.5	90.4	95.7	91.1
EN_FKNN[g]	98.2	83.6	79.4	82.4	90.4	**97.9**	91.5
Auto_Cova[h]	86.4	90.7	**93.8**	85.7	92.1	93.8	90.0
This Paper	**99.1**	89.1	85.3	**88.2**	84.3	95.8	**92.4**

Highest values are bold faced

C cytoplasm, *B* membrane, *M* mitochondrion, *S* secreted, *N* nucleus, *E* endoplasmic reticulum

[a] Chen and Li (2004)

[b] Chen and Li (2007)

[c] Zhang et al. (2006)

[d] Jiang et al. (2008)

[e] Ding and Zhang (2008)

[f] Lin et al. (2009)

[g] Gu et al. (2010)

[h] Yu et al. (2011)

locations) revealed that, each of the proposed modules performance was not random. In general, a predictor is considered to be random when the AUC values are less than or equal to 0.5 and equals 1 when it is perfectly accurate (Zou et al. 2007). The computed AUC values on different modules (Table 8), which are significantly >0.5 and close to 1, suggest that the proposed modules were not random and close to perfect.

Beside this the new independent dataset (N151) which contained a total of 151 proteins were evaluated. To avoid the bias in the prediction, the independent data set contained no two sequences with more than 40 % similarity within the same class,

Table 6 Comparison of MCC values of recent methods on CL317 dataset

Subcellular location	Mathews correlation coefficient (%)		
	This paper	Auto covariance[a]	EN_FKNN[b]
Cytoplasm	**0.955**	0.823	0.907
Membrane	0.871	0.878	**0.887**
Mitochondria	**0.916**	0.899	0.880
Secreted	**0.936**	0.767	0.902
Nucleus	0.870	**0.896**	0.827
ER	**0.963**	0.899	0.841

Highest values are bold faced

[a] Yu et al. (2011)

[b] Gu et al. (2010)

as well as with the training set. Albeit, the above measures, the accuracy of the N151 was found to be 90.7 %. Z98 dataset, as an independent dataset, were also subjected to prediction through proposed method and achieved 87.7 % accuracy. The sensitivity, specificity, and MCC (Table 9) on independent datasets suggest that the APSLAP was consistent in predicting SLP of apoptosis proteins.

Comparison of jack-knife test results with other earlier methods (as earlier methods adopted jack-knife test in common for evaluation) suggested that, our method to be efficient in accurately predicting the subcellular localization of apoptosis proteins. The reason for the higher accuracy level could be due to the fact

Table 7 Comparison of different methods by jack-knife test on Z98 dataset

Method	C	B	M	O	Overall accuracy %
Covariant[a]	**97.7**	73.3	30.8	25.0	72.5
ID[b]	90.7	90.0	92.3	91.7	90.8
SVM + 20SC[c]	86.0	90.0	**100.0**	**100.0**	90.8
DF_SVM[d]	**97.7**	90.0	92.3	83.3	92.9
ID_SVM[e]	95.3	93.3	84.6	58.3	88.8
FKNN[f]	95.3	96.7	**100.0**	91.7	**95.9**
PseAAC_SVM[g]	95.3	93.3	92.3	83.3	92.9
This Paper	95.3	90.0	**100.0**	91.7	94.9

Highest values are bold faced

C cytoplasm, *B* membrane, *M* mitochondrion, *O* others

[a] Zhou and Doctor (2003)

[b] Chen and Li (2004)

[c] Huang and Shi (2005)

[d] Zhang et al. (2006)

[e] Chen and Li (2007)

[f] Ding and Zhang (2008)

[g] Lin et al. (2009)

Table 8 Area under curve values of each classifier on different AdaBoost modules

Subcellular location	AUC[a] values				
	AA[b]	CPF[c]	CTD[d]	PSM[e]	HYB[f]
Cytoplasm	0.9195	0.9279	0.9444	0.9264	0.9745
Membrane	0.9498	0.9379	0.9551	0.9241	0.9377
Mitochondria	0.9282	0.9179	0.9379	0.9461	0.9897
Secreted	0.7933	0.8115	0.9161	0.9676	0.9795
Nucleus	0.9323	0.9257	0.9245	0.8943	0.9658
ER	0.9828	0.9504	0.9902	0.9324	0.9970

[a] Area under receiver operator characteristic curve

[b] Amino acid

[c] Class pattern frequency

[d] Composition transition and distribution of physio-chemical parameters

[e] Protein similarity measure

[f] Hybrid

Table 9 Performance of APSLAP on independent dataset test

Dataset	Independent dataset test			
	Sensitivity	Specificity	MCC	Overall accuracy (%)
ZD98	0.88	0.96	0.83	87.7
New151	0.90	0.98	0.87	90.7

that protein were represented in various aspect, that included composition, transition, and distribution of physiochemical parameters, protein similarity measure, distribution of physiochemical class patterns, a hybrid representation, information about the signal peptides, and local sequence similarity with the protein with single decided subcellular location using a strong ensemble classifier. The strength of the AdaBoost classifier could be seen from the higher accuracy rate of the proposed method without BLAST and SP modules. The user friendly web interface "APSLAP" was developed and it is freely available at http://apslap.bicpu.edu.in. Apart from predicting the subcellular localization of apoptosis protein, APSLAP also designed to report the possibility of query protein involvement in the process of apoptosis or not. This was achieved with the help of Interpro Scan (Hunter et al. 2012) and basic local alignment search with the UniProtKB. The query protein was first subjected to domain search using Interpro Scan and the gene ontology associated with the biological process "apoptosis" if any for the resulting domain was reported with the corresponding gene ontology identifier. Further basic local alignment search was done on the query protein against UniProtKB (reviewed entries alone), to look for the biological process assigned as 'apoptosis', if any, in the resulting highly similar protein. Moreover, all the supplementary materials and instruction for using the tool could be found at the same web interface.

4 Conclusion

A novel combination of feature vectors was used in this work to predict the subcellular localization of apoptosis protein. Benchmark data set CL317 and new independent dataset (151 proteins) were used to validate the proposed method. The result obtained from jack-knife test was higher than that of all other existing method in predicting SLAP. Also, from the results the newly designed feature, Class Pattern Frequency of Physiochemical Descriptor, was efficient in extracting information from the protein sequence. The experimental results of our combined method with seven different modules were promising. Apart from predicting the subcellular localization of apoptosis proteins, APSLAP also reports the possibility of query protein's involvement in the process of apoptosis or not, which helps in determining the function of the query protein as well. With the growing amount of protein data, it is anticipated that this tool will play a complementary role with other methods in drug design, systems biology, and proteomics studies on apoptosis proteins.

5 Supplement Information

All the supplementary material mentioned in the article is freely available at http://apslap.bicpu.edu.in/supp.php

Acknowledgments Saravanan Vijayakumar is supported by the DBT-BINC junior research fellow. The authors thank Borhan Kazimipour of Shiraz University for suggestions in algorithm design. The authors thank Dr. Archana Pan (Centre for Bioinforamtics, Pondicherry University) and Dr. Sivasathya (Department of Computer Science, Ponidcherry University) for their valuable suggestions and P. Elavarasan, E. Malanco and S. Manikandan of Centre for Bioinformatics, Pondicherry University for their technical support in hosting the tool onto the web server.

Conflict of interest The author declare no conflict of interest.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215(3):403–410. doi:10.1016/S0022-2836(05)80360-2

Binkowski TA, Adamian L, Liang J (2003) Inferring functional relationships of proteins from local sequence and spatial surface patterns. J Mol Biol 332(2):505–526

Breiman L (2001) Random forests. Mach Learn 45(1):5–32

Bulashevska A, Eils R (2006) Predicting protein subcellular locations using hierarchical ensemble of Bayesian classifiers based on Markov chains. BMC Bioinform 7:298. doi:10.1186/1471-2105-7-298

Carr K, Murray E, Armah E, He RL, Yau SS (2010) A rapid method for characterization of protein relatedness using feature vectors. PLoS One 5(3):e9550. doi:10.1371/journal.pone.0009550

Chen Y, Li Q (2004) Prediction of the subcellular location apoptosis proteins using the algorithm of measure of diversity. Acta Sci Nat Univ NeiMongol 25:413–417

Chen YL, Li QZ (2007) Prediction of the subcellular location of apoptosis proteins. J Theor Biol 245(4):775–783. doi:10.1016/j.jtbi.2006.11.010

Chou KC (1995) A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space. Proteins 21(4):319–344. doi:10.1002/prot.340210406

Chou KC, Shen HB (2007) Recent progress in protein subcellular location prediction. Anal Biochem 370(1):1–16. doi:10.1016/j.ab.2007.07.006

Deng M, Yu C, Liang Q, He RL, Yau SS (2011) A novel method of characterizing genetic sequences: genome space with biological distance and applications. PLoS ONE 6(3):e17293. doi:10.1371/journal.pone.0017293

Dietterich TG (2000) An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization. Mach Learn 40(2):139–157

Ding CH, Dubchak I (2001) Multi-class protein fold recognition using support vector machines and neural networks. Bioinformatics 17(4):349–358

Ding Y-S, Zhang T-L (2008) Using Chou's pseudo amino acid composition to predict subcellular localization of apoptosis proteins: an approach with immune genetic algorithm-based ensemble classifier. Pattern Recogn Lett 29(13):1887–1892

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 USA 92(19):8700–8704

Freund Y, Schapire R (1996) Experiments with a new boosting algorithm. In: International conference on machine learning, pp 148–156

Freund Y, Schapire RE (1997) A decision-theoretic generalization of on-line learning and an application to boosting. J Comput Syst Sci 55(1):119–139

Gu Q, Ding YS, Jiang XY, Zhang TL (2010) Prediction of subcellular location apoptosis proteins with ensemble classifier and feature selection. Amino Acids 38(4):975–983. doi:10.1007/s00726-008-0209-4

Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH (2009) The WEKA data mining software: an update. ACM SIGKDD Explor Newslett 11(1):10–18

Huang J, Shi F (2005) Support vector machines for predicting apoptosis proteins types. Acta Biotheor 53(1):39–47. doi:10.1007/s10441-005-7002-5

Hunter S, Jones P, Mitchell A, Apweiler R, Attwood TK, Bateman A, Bernard T, Binns D, Bork P, Burge S, de Castro E, Coggill P, Corbett M, Das U, Daugherty L, Duquenne L, Finn RD, Fraser M, Gough J, Haft D, Hulo N, Kahn D, Kelly E, Letunic I, Lonsdale D, Lopez R, Madera M, Maslen J, McAnulla C, McDowall J, McMenamin C, Mi H, Mutowo-Muellenet P, Mulder N, Natale D, Orengo C, Pesseat S, Punta M, Quinn AF, Rivoire C, Sangrador-Vegas A, Selengut JD, Sigrist CJ, Scheremetjew M, Tate J, Thimmajanarthanan M, Thomas PD, Wu CH, Yeats C, Yong SY (2012) InterPro in 2011: new developments in the family and domain prediction database. Nucleic acids research 40 (Database issue):D306-312. doi:10.1093/nar/gkr948

Jiang X, Wei R, Zhang T, Gu Q (2008) Using the concept of Chou's pseudo amino acid composition to predict apoptosis proteins subcellular location: an approach by approximate entropy. Protein Pept Lett 15:392–396

Kandaswamy KK, Pugalenthi G, Moller S, Hartmann E, Kalies KU, Suganthan PN, Martinetz T (2010) Prediction of apoptosis protein locations with genetic algorithms and support vector machines through a new mode of pseudo amino acid composition. Protein Pept Lett 17(12):1473–1479

Kerr JF, Wyllie AH, Currie AR (1972) Apoptosis: a basic biological phenomenon with wide-ranging implications in tissue kinetics. Br J Cancer 26(4):239–257

Liao B, Jiang JB, Zeng QG, Zhu W (2011) Predicting apoptosis protein subcellular location with PseAAC by incorporating tripeptide composition. Protein Pept Lett 18(11):1086–1092

Lin H, Wang H, Ding H, Chen YL, Li QZ (2009) Prediction of subcellular localization of apoptosis protein using Chou's pseudo amino acid composition. Acta Biotheor 57(3):321–330. doi:10.1007/s10441-008-9067-4

Matsuda S, Vert JP, Saigo H, Ueda N, Toh H, Akutsu T (2005) A novel representation of protein sequences for prediction of subcellular location using support vector machines. Protein Sci Publ Protein Soc 14(11):2804–2813. doi:10.1110/ps.051597405

Petersen TN, Brunak S, von Heijne G, Nielsen H (2011) SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods 8(10):785–786. doi:10.1038/nmeth.1701

Raff M (1998) Cell suicide for beginners. Nature 396(6707):119–122. doi:10.1038/24055

Saravanan V, Lakshmi PT (2013) SCLAP: an adaptive boosting method for predicting subchloroplast localization of plant proteins. OMICS 17(2):106–115. doi:10.1089/omi.2012.0070

Schapire RE, Singer Y (1999) Improved boosting algorithms using confidence-rated predictions. Mach Learn 37(3):297–336

Schulz JB, Weller M, Moskowitz MA (1999) Caspases as treatment targets in stroke and neurodegenerative diseases. Ann Neurol 45(4):421–429

Shen HB, Chou KC (2006) Ensemble classifier for protein fold pattern recognition. Bioinformatics 22(14):1717–1722. doi:10.1093/bioinformatics/btl170

Suzuki M, Youle RJ, Tjandra N (2000) Structure of Bax: coregulation of dimer formation and intracellular localization. Cell 103(4):645–654

Tantoso E, Li KB (2008) AAIndexLoc: predicting subcellular localization of proteins based on a new representation of sequences using amino acid indices. Amino Acids 35(2):345–353. doi:10.1007/s00726-007-0616-y

Thompson CB (1995) Apoptosis in the pathogenesis and treatment of disease. Science 267(5203): 1456–1462

Wang G, Dunbrack RL Jr (2003) PISCES: a protein sequence culling server. Bioinformatics 19(12):1589–1591

Yau SS, Yu C, He R (2008) A protein map and its application. DNA Cell Biol 27(5):241–250. doi:10.1089/dna.2007.0676

Yu C, Liang Q, Yin C, He RL, Yau SS (2010) A novel construction of genome space with biological geometry. DNA Res Int J Rapid Publ Reports Genes Genomes 17(3):155–168. doi:10.1093/dnares/dsq008

Yu C, Cheng SY, He RL, Yau SS (2011) Protein map: an alignment-free sequence comparison method based on various properties of amino acids. Gene 486(1–2):110–118. doi:10.1016/j.gene.2011.07.002

Yu X, Zheng X, Liu T, Dou Y, Wang J (2012) Predicting subcellular location of apoptosis proteins with pseudo amino acid composition: approach from amino acid substitution matrix and auto covariance transformation. Amino Acids 42(5):1619–1625. doi:10.1007/s00726-011-0848-8

Yu C, Deng M, Cheng SY, Yau SC, He RL, Yau SS (2013) Protein space: a natural method for realizing the nature of protein universe. J Theor Biol 318:197–204. doi:10.1016/j.jtbi.2012.11.005

Zhang H, Gu C (2006). Support Vector Machines versus Boosting. Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, USA

Zhang ZH, Wang ZH, Zhang ZR, Wang YX (2006) A novel method for apoptosis protein subcellular localization prediction combining encoding based on grouped weight and support vector machine. FEBS Lett 580(26):6169–6174. doi:10.1016/j.febslet.2006.10.017

Zhou GP, Doctor K (2003) Subcellular location prediction of apoptosis proteins. Proteins 50(1):44–48. doi:10.1002/prot.10251

Zou KH, O'Malley AJ, Mauri L (2007) Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. Circulation 115(5):654–657. doi:10.1161/CIRCULATIONAHA.105.594929