International Journal of
Molecular Sciences

MDPI

Communication

Predicting Apoptosis Protein Subcellular Locations based on the Protein Overlapping Property Matrix and Tri-Gram Encoding

Yang Yang [1,†], Huiwen Zheng [2,†], Chunhua Wang [3,†], Wanyue Xiao [1] and Taigang Liu [3,4,*]

[1] AIEN Institute, Shanghai Ocean University, Shanghai 201306, China; 1591117@st.shou.edu.cn (Y.Y.); poppy_xiaowanyue@163.com (W.X.)

[2] College of Sciences & Engineering, University of Tasmania, 7001 Tasmania, Australia; huiwenz@utas.edu.au

[3] College of Information Technology, Shanghai Ocean University, Shanghai 201306, China; wchshou@163.com

[4] Key Laboratory of Fisheries Information Ministry of Agriculture, Shanghai 201306, China

* Correspondence: tgliu@shou.edu.cn; Tel.: +86-21-61900624

† These authors contributed equally to this work.

Received: 1 April 2019; Accepted: 8 May 2019; Published: 11 May 2019

Abstract: To reveal the working pattern of programmed cell death, knowledge of the subcellular location of apoptosis proteins is essential. Besides the costly and time-consuming method of experimental determination, research into computational locating schemes, focusing mainly on the innovation of representation techniques on protein sequences and the selection of classification algorithms, has become popular in recent decades. In this study, a novel tri-gram encoding model is proposed, which is based on using the protein overlapping property matrix (POPM) for predicting apoptosis protein subcellular location. Next, a 1000-dimensional feature vector is built to represent a protein. Finally, with the help of support vector machine-recursive feature elimination (SVM-RFE), we select the optimal features and put them into a support vector machine (SVM) classifier for predictions. The results of jackknife tests on two benchmark datasets demonstrate that our proposed method can achieve satisfactory prediction performance level with less computing capacity required and could work as a promising tool to predict the subcellular locations of apoptosis proteins.

Keywords: tri-gram; protein overlapping property matrix; subcellular location; support vector machine; recursive feature elimination

1. Introduction

Apoptosis, as a form of programmed cell death occurring in multicellular organisms, is vital for balancing cell proliferation and death. Malfunction of apoptosis can trigger undesirable maladies including cancer, autoimmune disease, ischemic damage, and neurodegenerative disease [1]. It has been certified that the apoptosis proteins' functions are closely connected with their subcellular locations [2]. Hence, assigning subcellular locations for apoptosis proteins is a crucial step to understanding their working mechanisms. However, conventional experimental methods to determine subcellular locations are usually time-consuming. The rapid development of high-throughput sequencing techniques has accelerated the demand for reliable and precise computational methods to locate apoptosis proteins' subcellular positions from their primary sequences.

From the machine learning perspective, the identification of protein subcellular locations is usually described as a multi-class classification problem. Researchers have made great efforts in this field. These methods focus mainly on two aspects: (1) the construction of protein sequence encoding schemes and feature extraction; and (2) the design of a classification algorithm. There exist multiple machine learning techniques to estimate protein subcellular positions, such as covariant discriminant [2], fuzzy k-nearest neighbor [3,4], support vector machine (SVM) [5–8], and ensemble classifier [9,10]. Among these, SVM is widely used for its robust prediction performance. In addition, a series of feature extraction schemes has been developed to transform protein sequences into fixed-length numeric vectors, including amino acid composition (AAC) [11], pseudo-amino acid composition (PseAAC) [12–16], grouped weight encoding [17], wavelet coefficients [6], distance frequency [5], position-specific scoring matrix (PSSM) profile [18–22], and fusion of multi-view features [23,24].

Recently, SVM has demonstrated promising performance with a fast processing speed and has become the most popular classifier for researchers. The main difference among various SVM-based methods is the feature encoding schemes used. Among these, sequence representation models based on the PSSM profile have shown the most conspicuous improvements in the prediction accuracy aspect [25,26]. The PSSM profile of each sequence is usually achieved by executing the Position Specific Iterated BLAST (PSI-BLAST) program [27] against a specific protein database, e.g., NCBI's non-redundant (NR) database or Swiss-Prot. Although the PSSM profile can provide important identifiable information for the prediction of protein subcellular location, several inherent limitations of alignment-based programs still exist. First, the PSI-BLAST technique is generally time-consuming and occupies more memory and has limited ability to deal with a large scale of sequence data. In addition, it is a challenge to construct a multiple sequence alignment, which is a type of NP-hard problem. Moreover, the accuracy of sequence alignments declines significantly due to the limited number of homologous sequences in the existing recognized database [28]. These issues have pushed us to design a more effective and alignment-free feature encoding method.

In a previous study, a tri-gram encoding scheme was used to transform the PSSM profiles of proteins into 8000-dimensional feature vectors [18]. However, this method has two main shortcomings: (1) obtaining the PSSM profiles is usually time-consuming; and (2) high dimension data is more likely to cause the curse of dimensionality and costs too many computing resources. On the other hand, the physicochemical properties of amino acids are generally considered to affect the structure and function of proteins. In this study, we present an improved tri-gram encoding technique based on the protein overlapping property matrix (POPM) to lower the threshold of computing capability. The use of the scheme is as follows: First, 20 amino acids are divided into ten overlapping sets, where each group represents a distinctive physicochemical property. Second, each residue is encoded by using a 10-dimensional binary vector based on its physicochemical property. Third, the POPM of a protein consists of these row vectors related to amino acids at the corresponding positions in the sequence. Fourth, tri-gram encoding technique transforms the POPM into a 1000-dimensional feature vector. Finally, the optimal features after recursive feature selection are put into an SVM classifier to make predictions. The datasets used in this study and the source code for implementing the algorithm are freely available to the academic community at https://github.com/taigangliu/POPM-trigram.

2. Results and Discussion

2.1. Effects of Top K Features

After computing tri-gram features, a 1000-dimensional feature vector for each protein was obtained. Then, we acquired a ranked list of these features on the basis of their importance with the help of SVM-RFE. To find the ideal dimensions, the overall accuracies for the top K features were calculated by using jackknife cross-validation, where K = 10, 20, 30, ..., 300. We set the maximum value of K to be 300 because the prediction accuracies decline after reaching their peak points.

Figure 1 shows the values on ZW225 and CL317 datasets with different top K features corresponding to their accuracies. It is clear that the overall accuracy (OA) for the ZW225 dataset reached the highest level when K climbed to 120. Besides, the CL317 dataset also obtained a favorable accuracy at this point. Therefore, we selected the top 120 features to represent a protein in the following study. Table 1 illustrates the performance of our method on two datasets by performing jackknife tests. As shown in the table, the accuracies of ZW225 and CL317 datasets reached relatively high levels of 98.2% and

96.2% respectively. Among these subcellular locations, the specificity (Spec) values were more than 98%, and the Matthews correlation coefficient (MCC) values were more than 92% for the two datasets. Notably, only the sensitivity (Sens) value of the secreted (Secr) location on the CL317 dataset was slightly lower than in the other locations and so was the accuracy of the mitochondrial (Mito) location on the ZW225 dataset. This may be due to the limited numbers of Mito and Secr proteins on the two datasets. Namely, the training sample size has an important influence on the accuracy.

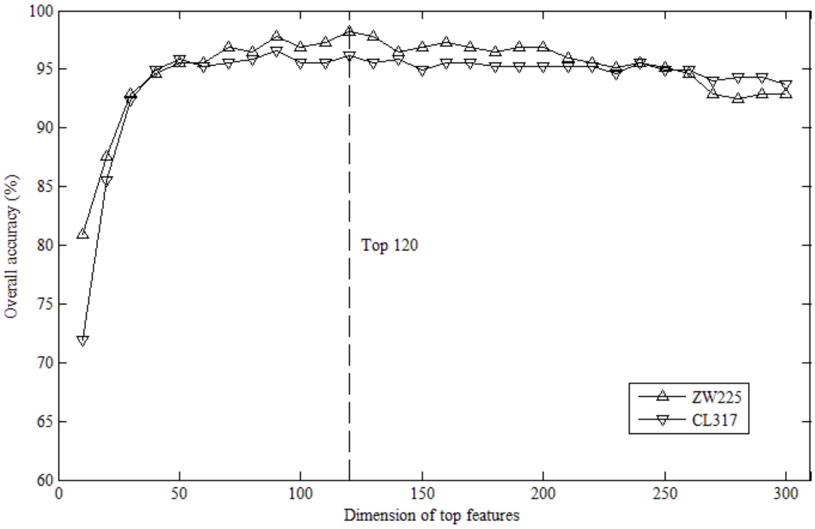

Figure 1. The graph illustrates the efficiency of various top features on the two datasets' overall accuracies.

Table 1. Results for the two datasets by jackknife tests.

Dataset	Location [1]	Sens (%)	Spec (%)	MCC	OA (%)
ZW225	Cyto	100	98.1	0.970	98.2
	Memb	96.6	100	0.972	
	Mito	96.0	100	0.977	
	Nucl	100	99.5	0.985	
CL317	Cyto	99.1	98.5	0.973	96.2
	Memb	92.7	98.9	0.923	
	Mito	97.1	99.3	0.951	
	Secr	88.2	100	0.936	
	Nucl	96.2	99.2	0.954	
	Endo	95.7	99.3	0.950	

[1] For short, cytoplasmic, membrane, mitochondrial, secreted, nuclear and endoplasmic reticulum are called Cyto, Memb, Mito, Secr, Nucl, and Endo, respectively.

2.2. Performance Comparison with Existing Methods

To assess the performance of our method objectively, we compared it with results from other existing methods based on the same datasets. The detailed outcomes of jackknife tests are reported in Tables 2 and 3 where the Sens of each class and the OA are chosen as performance indexes. The comparison results of Spec and MCC of different methods on the two datasets are listed in Tables S1–S4 in the Supplementary Materials.

Table 2. Performance comparison of different methods on the ZW225 dataset.

Method	Sens for Each Class (%)				OA (%)
	Cyto	Memb	Mito	Nucl	
EBGW_SVM [17]	90.0	93.3	60.0	63.4	83.1
DF_SVM [5]	87.1	92.1	64.0	73.2	84.0
PSSM_AC [8]	82.9	92.1	68.0	78.0	84.0
ID_SVM [15]	92.9	91.0	68.0	73.2	85.8
Auto_Cova [12]	81.3	93.3	85.7	84.6	87.1
EN_FKNN [9]	94.3	94.4	60.0	80.5	88.0
Tri-gram PSSM [18]	97.1	98.9	96.0	97.6	97.8
PsePSSM-DCCA-LFDA [25]	100	98.9	100	100	99.6
Our method	100	96.6	96.0	100	98.2

Table 3. Performance comparison of different methods on the CL317 dataset.

Method	Sens for Each Class (%)						OA (%)
	Cyto	Memb	Mito	Secr	Nucl	Endo	
ID [14]	81.3	81.8	85.3	88.2	82.7	83.0	82.7
ID_SVM [15]	91.1	89.1	79.4	58.8	73.1	87.2	84.2
DF_SVM [5]	92.9	85.5	76.5	76.5	93.6	86.5	88.0
PseAAC_SVM [13]	93.8	90.9	85.3	76.5	90.4	95.7	91.1
PSSM-AC [8]	93.8	90.9	91.2	82.4	86.5	95.7	91.5
APSLAP [10]	99.1	89.1	85.3	88.2	84.3	95.8	92.4
Tri-gram PSSM [18]	98.2	96.4	94.1	82.4	96.2	95.7	95.9
PsePSSM-DCCA-LFDA [25]	99.1	100	100	100	100	100	99.7
Our method	99.1	92.7	97.1	88.2	96.2	95.7	96.2

Based on Table 2, our method had an outstanding overall performance (98.2% in OA) on the ZW225 dataset, which was an improvement of over 10% compared with other methods such as Auto_Cova [12] and EN_FKNN [9]. Noticeably, the prediction accuracies of both Cyto proteins and Nucl proteins achieved 100%. Also, the Sens values of Memb and Mito reached relatively high prediction levels, with 96.6% and 96% respectively, which performed better than many other methods, including EBGW_SVM [17], DF_SVM [5], PSSM_AC [8], and ID_SVM [15]. In general, for the ZW225 dataset, our proposed method achieved a pleasing level.

In Table 3, for the CL317 dataset, our method generated a relatively high OA (96.2%) and achieved a remarkably enhanced performance for the subcellular locations of Cyto and Mito with 99.1% and 97.1% respectively. Compared with the previous study of the tri-gram PSSM algorithm [18], the proposed method not only improves the prediction accuracy but also largely reduces the computing costs. Admittedly, the PsePSSM-DCCA-LFDA [25] method performs excellently in every aspect, reaching 100% in almost all performance indexes for both datasets. This means that the combination of those three proven techniques—PsePSSM, DCCA, and LFDA—is effective for predicting protein subcellular locations. However, generating the PSSM profiles of query proteins by PSI-BLAST program is usually time-consuming and memory-consuming, which may limit its application with large-scale sequence data. To illustrate this point, the longest sequence (ID: Q68749, length: 3037) and the shortest sequence (ID: O43715, length: 76) of the datasets were selected to test the time required of this method. Remarkably, in our laboratory environment (Intel Xeon CPU E5620 @ 2.40GHz, 16 4-core processors, 16G RAM), it took 8334 seconds and 471 seconds to generate the PSSM profiles of two proteins (i.e., Q68749 and O43715) respectively. This result also indicates that the longer a sequence is, the more time it will take to process it. However, the required time for obtaining POPMs of two proteins is less than 1 second, which suggests that our method provides a convenient and fast way to extract features solely from amino acid sequences. The results also show that this relatively efficient method can achieve a favorable prediction accuracy as well.

In conclusion, our method not only greatly reduces the computational complexity but also obtains a comparable performance for predicting apoptosis protein subcellular locations. This significant progress is attributed to the powerful feature encoding scheme based on the tri-gram computed from POPM and SVM-RFE applied to select optimal features.

3. Materials and Methods

3.1. Datasets

Two benchmark datasets were employed to examine the validity of the proposed method: the ZW225 dataset [17] and the CL317 dataset [14,15]. The ZW225 dataset consists of 225 apoptosis proteins with 41 nuclear proteins, 70 cytoplasmic proteins, 25 mitochondrial proteins, and 89 membrane proteins. There are six types of subcellular locations in the CL317 dataset, including 112 cytoplasmic proteins, 47 endoplasmic reticulum proteins, 55 membrane proteins, 34 mitochondrial proteins, 52 nuclear proteins, and 17 secreted proteins. For short, cytoplasmic, membrane, mitochondrial, secreted, nuclear and endoplasmic reticulum are called Cyto, Memb, Mito, Secr, Nucl, and Endo, respectively.

3.2. Feature Extraction from POPM

It is demonstrated that, when properly represented, the amino acid composition of protein sequences contains the information necessary to delineate the structure and function of proteins. The physicochemical properties of amino acids encoded in protein sequences are believed to be important and original discriminatory information for predicting protein subcellular locations.

In this work, Taylor's overlapping physicochemical properties were adopted due to their successful application in catalytic residue prediction [29] and phosphorylation site identification [30]. The ten properties are listed in Table 4 [31]. Each amino acid residue was encoded using a 10-dimensional binary vector based on its physicochemical properties where the dimensions of the corresponding properties were set to 1 and the remaining positions were 0, e.g., A (0000100010) and V (0000110100). Thus, the POPM of a protein is defined as an $L \times 10$ binary encoding matrix, which is denoted as $[M_{i,j}]$, where $i = 1, 2, \ldots, L$ denotes the position in the sequence and $j = 1, 2, \ldots, 10$ denotes a physicochemical property.

Table 4. Amino acid groups based on Taylor's overlapping properties.

Physicochemical Properties	Amino Acid Residues
Polar	N, Q, S, D, E, C, T, K, R, H, Y, W
Positive	K, H, R
Negative	D, E
Charged	K, H, R, D, E
Hydrophobic	A, G, C, T, I, V, L, K, H, F, Y, W, M
Aliphatic	I, V, L
Aromatic	F, Y, W, H
Small	P, N, D, T, C, A, G, S, V
Tiny	A, S, G, C
Proline	P

Then, the tri-gram encoding technique based on POPM was adopted to represent sequence samples, which reflected local interactions among three adjacent amino acids. Tri-gram features were generated using the following formula:

$$gram(x, y, z) = \frac{1}{L-2} \sum_{i=1}^{L-2} M_{i,x} \times M_{i+1,y} \times M_{i+2,z} (1 \le x, y, z \le 10). \quad (1)$$

Hence, the total number of tri-gram features extracted from POPM was 1000.

3.3. Support Vector Machine

SVM is a powerful machine learning model, which has been widely used for many protein prediction tasks in the field of computational biology [32–37]. Noticeably, this technique can construct an optimum hyperplane in a high-dimensional space to achieve precise linear classification. Besides, SVM can attain a non-linear classification with the use of a kernel trick. In this study, we chose the LIBSVM package [38] to help the SVM classifier work better. Among four in-built kernels provided by the LIBSVM package, i.e., linear, polynomial, radial basis function, and Gaussian, we adopted the linear kernel for this work, since it takes parameter optimization into account.

3.4. Feature Selection by SVM-RFE

The contrasting dimensions of the feature vector can lead to obvious difference of efficiency in machine learning. The smaller the dimension, the less information it contains, which is not sufficient to identify the classification of the sample. In other words, the higher the dimension, the more information redundancy is involved, which not only greatly increases the computational complexity, but also affects the prediction accuracy. Therefore, we used support vector machine-recursive feature elimination (SVM-RFE) to select the appropriate features. This useful technique was originally applied for cancer classification [39] and then applied to predict functional attributes of proteins [40,41]. Firstly, we constructed a feature matrix of all the feature vectors of proteins according to each dataset, where each row represented a sample and each column corresponded to a feature. Next, all features were be ranked based on their importance through the SVM-RFE algorithm. Finally, the top *K* features ranked in the list were selected to represent a protein sequence.

3.5. Performance Evaluation

For statistical prediction, there are three types of cross-validation methods: the independent dataset test, the sub-sampling test, and the jackknife test [42,43]. In this research, the jackknife test was adopted to evaluate the performance of predictors because of its objectivity and rigorousness. During the jackknife test, each protein sequence in the dataset was picked out successively as a test sample, while the rest of protein sequences played the role of training samples.

To objectively assess the performance of our method, four standard performance indexes were reported, including Sens, Spec, and MCC for each subcellular location, and the OA [44,45]. They were defined using the following formulae:

$$Sens_j = \frac{TP_j}{TP_j + FN_j} = \frac{TP_j}{|C_j|}, \tag{2}$$

$$Spec_j = \frac{TN_j}{TN_j + FP_j} = \frac{TN_j}{\sum_{k \neq j} |C_k|}, \tag{3}$$

$$MCC_j = \frac{TP_j TN_j - FP_j FN_j}{\sqrt{(TP_j + FP_j)(TP_j + FN_j)(TN_j + FP_j)(TN_j + FN_j)}}, \tag{4}$$

$$OA = \frac{\sum_j TP_j}{\sum_j |C_j|}. \tag{5}$$

here, TP_j, TN_j, FP_j, FN_j, and $|C_j|$ indicate the number of true positives, true negatives, false positives, false negatives, and proteins in the subcellular location C_j, respectively.

4. Conclusions

In this study, we focused on the design of a high-efficiency feature extraction technique for the prediction of the subcellular locations of apoptosis proteins. Firstly, a tri-gram encoding scheme based on POPM was introduced to transform the sequences of query proteins into 1000-dimensional feature vectors. Then, 120 optimal features selected by the SVM-RFE algorithm were input into a SVM prediction engine to perform the classification. The comparison with other existing models very strongly suggested that the proposed method is not encumbered by the limitations of alignment-based methods and could work as a very cost-effective tool for predicting subcellular location of apoptosis proteins. Due to the generality of this method, it is promising as an application for other protein classification problems in the future.

Supplementary Materials: Supplementary materials can be found at http://www.mdpi.com/1422-0067/20/9/2344/s1.

Author Contributions: The manuscript was written through contributions of all authors. T.L. designed the study and revised the paper. Y.Y., H.Z. and W.X. analyzed the data and wrote the first draft of paper. C.W. implemented the algorithm. All authors read and approved the final manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, grant numbers 11601324 and 11701363.

Acknowledgments: The authors would like to thank Xiaoguang Bao for his pertinent suggestions.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Steller, H. Mechanisms and genes of cellular suicide. *Science* **1995**, *267*, 1445–1449. [CrossRef] [PubMed]
2. Zhou, G.P.; Doctor, K. Subcellular location prediction of apoptosis proteins. *Proteins* **2003**, *50*, 44–48. [CrossRef]
3. Ding, Y.S.; Zhang, T.L. 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.* **2008**, *29*, 1887–1892. [CrossRef]
4. Jiang, X.; Wei, R.; Zhang, T.; Gu, Q. Using the concept of Chou's pseudo amino acid composition to predict apoptosis proteins subcellular location: An approach by approximate entropy. *Protein Pept. Lett.* **2008**, *15*, 392–396. [CrossRef] [PubMed]
5. Zhang, L.; Liao, B.; Li, D.; Zhu, W. A novel representation for apoptosis protein subcellular localization prediction using support vector machine. *J. Theor. Biol.* **2009**, *259*, 361–365. [CrossRef]
6. Qiu, J.D.; Luo, S.H.; Huang, J.H.; Sun, X.Y.; Liang, R.P. Predicting subcellular location of apoptosis proteins based on wavelet transform and support vector machine. *Amino Acids* **2010**, *38*, 1201–1208. [CrossRef] [PubMed]
7. Huang, J.; Shi, F. Support vector machines for predicting apoptosis proteins types. *Acta Biotheor.* **2005**, *53*, 39–47. [CrossRef]
8. Liu, T.; Zheng, X.; Wang, C.; Wang, J. Prediction of subcellular location of apoptosis proteins using pseudo amino acid composition: An approach from auto covariance transformation. *Protein Pept. Lett.* **2010**, *17*, 1263–1269. [CrossRef]
9. Gu, Q.; Ding, Y.S.; Jiang, X.Y.; Zhang, T.L. Prediction of subcellular location apoptosis proteins with ensemble classifier and feature selection. *Amino Acids* **2010**, *38*, 975–983. [CrossRef]
10. Saravanan, V.; Lakshmi, P.T. APSLAP: An adaptive boosting technique for predicting subcellular localization of apoptosis protein. *Acta Biotheor.* **2013**, *61*, 481–497. [CrossRef]
11. Zhou, X.B.; Chen, C.; Li, Z.C.; Zou, X.Y. Improved prediction of subcellular location for apoptosis proteins by the dual-layer support vector machine. *Amino Acids* **2008**, *35*, 383–388. [CrossRef]
12. Yu, X.; Zheng, X.; Liu, T.; Dou, Y.; Wang, J. Predicting subcellular location of apoptosis proteins with pseudo amino acid composition: Approach from amino acid substitution matrix and auto covariance transformation. *Amino Acids* **2012**, *42*, 1619–1625. [CrossRef]
13. Lin, H.; Wang, H.; Ding, H.; Chen, Y.L.; Li, Q.Z. Prediction of subcellular localization of apoptosis protein using Chou's pseudo amino acid composition. *Acta Biotheor.* **2009**, *57*, 321–330. [CrossRef]
14. Chen, Y.L.; Li, Q.Z. Prediction of the subcellular location of apoptosis proteins. *J. Theor. Biol.* **2007**, *245*, 775–783. [CrossRef]

15. Chen, Y.L.; Li, Q.Z. Prediction of apoptosis protein subcellular location using improved hybrid approach and pseudo-amino acid composition. *J. Theor. Biol.* **2007**, *248*, 377–381. [CrossRef] [PubMed]
16. Liao, B.; Jiang, J.B.; Zeng, Q.G.; Zhu, W. Predicting apoptosis protein subcellular location with PseAAC by incorporating tripeptide composition. *Protein Pept. Lett.* **2011**, *18*, 1086–1092. [CrossRef]
17. Zhang, Z.H.; Wang, Z.H.; Zhang, Z.R.; Wang, Y.X. A novel method for apoptosis protein subcellular localization prediction combining encoding based on grouped weight and support vector machine. *FEBS Lett.* **2006**, *580*, 6169–6174. [CrossRef]
18. Liu, T.; Tao, P.; Li, X.; Qin, Y.; Wang, C. Prediction of subcellular location of apoptosis proteins combining tri-gram encoding based on PSSM and recursive feature elimination. *J. Theor. Biol.* **2015**, *366*, 8–12. [CrossRef]
19. Zhang, S.; Liang, Y. Predicting apoptosis protein subcellular localization by integrating auto-cross correlation and PSSM into Chou's PseAAC. *J. Theor. Biol.* **2018**, *457*, 163–169. [CrossRef] [PubMed]
20. Yu, B.; Li, S.; Qiu, W.Y.; Chen, C.; Chen, R.X.; Wang, L.; Wang, M.H.; Zhang, Y. Accurate prediction of subcellular location of apoptosis proteins combining Chou's PseAAC and PsePSSM based on wavelet denoising. *Oncotarget* **2017**, *8*, 107640–107665. [CrossRef]
21. Xiang, Q.; Liao, B.; Li, X.; Xu, H.; Chen, J.; Shi, Z.; Dai, Q.; Yao, Y. Subcellular localization prediction of apoptosis proteins based on evolutionary information and support vector machine. *Artif. Intell. Med.* **2017**, *78*, 41–46. [CrossRef] [PubMed]
22. Liang, Y.; Liu, S.; Zhang, S. Detrended cross-correlation coefficient: Application to predict apoptosis protein subcellular localization. *Math. Biosci.* **2016**, *282*, 61–67. [CrossRef] [PubMed]
23. Zhang, S.; Zhang, T.; Liu, C. Prediction of apoptosis protein subcellular localization via heterogeneous features and hierarchical extreme learning machine. *SAR QSAR Environ. Res.* **2019**, *30*, 209–228. [CrossRef] [PubMed]
24. Li, B.; Cai, L.; Liao, B.; Fu, X.; Bing, P.; Yang, J. Prediction of Protein Subcellular Localization Based on Fusion of Multi-view Features. *Molecules* **2019**, *24*, 919. [CrossRef]
25. Yu, B.; Li, S.; Qiu, W.; Wang, M.; Du, J.; Zhang, Y.; Chen, X. Prediction of subcellular location of apoptosis proteins by incorporating PsePSSM and DCCA coefficient based on LFDA dimensionality reduction. *BMC Genom.* **2018**, *19*, 478. [CrossRef]
26. Liang, Y.; Zhang, S. Prediction of Apoptosis Protein's Subcellular Localization by Fusing Two Different Descriptors Based on Evolutionary Information. *Acta Biotheor.* **2018**, *66*, 61–78. [CrossRef]
27. Altschul, S.F.; Madden, T.L.; Schaffer, A.A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D.J. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **1997**, *25*, 3389–3402. [CrossRef]
28. Zielezinski, A.; Vinga, S.; Almeida, J.; Karlowski, W.M. Alignment-free sequence comparison: Benefits, applications, and tools. *Genome Biol.* **2017**, *18*, 186. [CrossRef] [PubMed]
29. Dou, Y.; Zheng, X.; Yang, J.; Wang, J. Prediction of catalytic residues based on an overlapping amino acid classification. *Amino Acids* **2010**, *39*, 1353–1361. [CrossRef] [PubMed]
30. Dou, Y.; Yao, B.; Zhang, C. PhosphoSVM: Prediction of phosphorylation sites by integrating various protein sequence attributes with a support vector machine. *Amino Acids* **2014**, *46*, 1459–1469. [CrossRef]
31. Taylor, W.R. The classification of amino acid conservation. *J. Theor. Biol.* **1986**, *119*, 205–218. [CrossRef]
32. Manavalan, B.; Basith, S.; Shin, T.H.; Wei, L.; Lee, G. mAHTPred: A sequence-based meta-predictor for improving the prediction of anti-hypertensive peptides using effective feature representation. *Bioinformatics* **2018**. [CrossRef] [PubMed]
33. Wei, L.; Luan, S.; Nagai, L.A.E.; Su, R.; Zou, Q. Exploring sequence-based features for the improved prediction of DNA N4-methylcytosine sites in multiple species. *Bioinformatics* **2018**. [CrossRef] [PubMed]
34. Manavalan, B.; Basith, S.; Shin, T.H.; Choi, S.; Kim, M.O.; Lee, G. MLACP: Machine-learning-based prediction of anticancer peptides. *Oncotarget* **2017**, *8*, 77121–77136. [CrossRef] [PubMed]
35. Wei, L.; Chen, H.; Su, R. M6APred-EL: A Sequence-Based Predictor for Identifying N6-methyladenosine Sites Using Ensemble Learning. *Mol. Ther. Nucleic Acids* **2018**, *12*, 635–644. [CrossRef] [PubMed]
36. Manavalan, B.; Shin, T.H.; Lee, G. DHSpred: Support-vector-machine-based human DNase I hypersensitive sites prediction using the optimal features selected by random forest. *Oncotarget* **2018**, *9*, 1944–1956. [CrossRef]

37. Wei, L.; Zhou, C.; Chen, H.; Song, J.; Su, R. ACPred-FL: A sequence-based predictor using effective feature representation to improve the prediction of anti-cancer peptides. *Bioinformatics* **2018**, *34*, 4007–4016. [CrossRef] [PubMed]
38. Chang, C.C.; Lin, C.J. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.* **2011**, *2*, 27. [CrossRef]
39. Guyon, I.; Weston, J.; Barnhill, S.; Vapnik, V. Gene Selection for Cancer Classification using Support Vector Machines. *Mach. Learn* **2002**, *46*, 389–422. [CrossRef]
40. Li, L.; Yu, S.; Xiao, W.; Li, Y.; Huang, L.; Zheng, X.; Zhou, S.; Yang, H. Sequence-based identification of recombination spots using pseudo nucleic acid representation and recursive feature extraction by linear kernel SVM. *BMC Bioinform.* **2014**, *15*, 340. [CrossRef]
41. Li, L.; Yu, S.; Xiao, W.; Li, Y.; Hu, W.; Huang, L.; Zheng, X.; Zhou, S.; Yang, H. Protein submitochondrial localization from integrated sequence representation and SVM-based backward feature extraction. *Mol. Biosyst.* **2015**, *11*, 170–177. [CrossRef] [PubMed]
42. Su, R.; Hu, J.; Zou, Q.; Manavalan, B.; Wei, L. Empirical comparison and analysis of web-based cell-penetrating peptide prediction tools. *Brief. Bioinform.* **2019**. [CrossRef] [PubMed]
43. Basith, S.; Manavalan, B.; Shin, T.H.; Lee, G. iGHBP: Computational identification of growth hormone binding proteins from sequences using extremely randomised tree. *Comput. Struct. Biotechnol. J.* **2018**, *16*, 412–420. [CrossRef]
44. Qu, K.; Han, K.; Wu, S.; Wang, G.; Wei, L. Identification of DNA-Binding Proteins Using Mixed Feature Representation Methods. *Molecules* **2017**, *22*, 1602. [CrossRef] [PubMed]
45. Wei, L.; Xing, P.; Su, R.; Shi, G.; Ma, Z.S.; Zou, Q. CPPred-RF: A Sequence-based Predictor for Identifying Cell-Penetrating Peptides and Their Uptake Efficiency. *J. Proteome Res.* **2017**, *16*, 2044–2053. [CrossRef]

© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).