Journal of Physics: Conference Series

PAPER • OPEN ACCESS

Implementation protein sequence segmentation in AAC and DC as protein descriptors for improving a classification performance of acetylation prediction

To cite this article: A Rizqiana *et al* 2021 *J. Phys.: Conf. Ser.* **1751** 012031

Implementation protein sequence segmentation in AAC and DC as protein descriptors for improving a classification performance of acetylation prediction

A Rizqiana[1,a], M R Faisal[2,b], F R Lumbanraja[3,c]

[1] Undergraduate School of Computer Science, Faculty of Mathematics and Natural Sciences, Lambung Mangkurat University, Banjarbaru, Indonesia
[2] Department of Computer Science, Faculty of Mathematics and Natural Sciences, Lambung Mangkurat University, Banjarbaru, Indonesia
[3] Department of Computer Science, Faculty of Mathematics and Natural Sciences, Lampung University, Lampung, Indonesia

email: rizqiananisa@gmail,com[a], reza.faisal@ulm.ac.id[b], favorisen.lumbanraja@fmipa.unila.ac.id[c]

Abstract. Post-Translational Modification (PTM) identification is carried out to determine the position of the PTM in protein. Acetylation in the lysine protein is one of the many types of PTM that play an important role in biological processes. In existing research, identification of lysine acetylation was developed by computational methods, using several available protein descriptors along with classification methods. Research on protein classification usually only uses the length of the protein sequence to describe the state of the whole protein, not its local state. Knowing the local state of the protein sequence will have a good effect on the classification results. To find out the situation, the protein sequence segmentation approach is done by adjacent and overlapped segments. Adjacent and overlapped segments divide the length of the protein into several segments, then numerical features will be calculated, so that information about the protein is also obtained locally. Calculation of numerical features using the Amino Acid Composition and Dipeptide Composition descriptors, then the data is classified with Support Vector Machine. The experimental results show that protein segmentation increases the performance of protein classification by 0.7-2.5%. Segmentation using adjacent and overlapped segments provides improved performance. In this research, it was found that protein segmentation affected the performance of protein classification, especially in overlapped segments.

Keyword: lysine acetylation, sequence segmentation, Amino Acid Composition, Dipeptide Composition, protein classification, Support Vector Machine

1. Introduction

Post-Translational Modifications (PTM) are chemical modifications of a polypeptide chain that occur after DNA has been transcribed into RNA and translated into protein. These chemical modifications of a polypeptide chain after its biosynthesis extends the range of amino acid structures and properties, and consequently, diversifies structures and functions of proteins [1] The most common posttranslational modifications include phosphorylation, sulfation, disulfide formation, N-methylation, O-methylation, S-methylation, N-acetylation, hydroxylation, glycosylation, ADP-ribosylation, prenylation, biotinylation, lipoylation, and phosphopantetheine tethering [2] Acetylation is one of the major post-translational protein modifications in the cell, with multiple effects on protein and metabolism. Acetylation in the lysine protein is a reaction that can be reversed enzymatically through a mechanism that is tightly regulated and dependent on metabolism [3] identification of protein acetylated sites through traditional experiment methods is time-consuming and laborious. Those methods are not suitable to identify a large number of acetylated sites quickly. Therefore, computational methods are still very valuable to accelerate lysine-acetylated site finding [4] In recent years, a number of computational methods for lysine PTM identification have been developed. These computational methods show variations and differences in the algorithm and feature extraction techniques. These methods need to be reviewed to find out the best computational method for PTM prediction [5] PTM used in this research is PTM acetylation on lysine protein.

One of the computational methods is classification. Basically, classification is a 2-step process; the first step is supervised learning for the sake of the predefined class label for training data set, the second step is classification accuracy evaluation [6] There are many methods that can be used in classification, including Support Vector Machine. Support Vector Machine is a method that works by defining the boundary between two data classes with the maximum distance from the closest data, the maximum distance obtained from the best hyperplane in the input space obtained by measure the hyperplane margin, which is the distance from the hyperplane to the closest point of each data class [7] SVM supports linear and non-linear classification. When data cannot be separated in a straight line, non-linear SVM is used, which is the kernel function. The research of [8] uses the Support Vector Machine method with 3 kernels, which are linear, Gaussian, and polynomial, and conclude that the Gaussian kernel has the best performance results with the highest accuracy than polynomial and linear kernels.

Before classified, protein data must be extracted first into structured data. The structural alignment of a pair of proteins can be defined with the use of a concept of protein descriptors. The protein descriptors are local substructures of protein molecules, which allow us to divide the original problem into a set of subproblems and, consequently, to propose a more efficient algorithmic solution [9] One of the protein descriptors is in the R package, which is protr. The protr is used to produce various numerical representation schemes of proteins from the amino acid sequence, in which they are eight descriptor groups composed of 22 types of commonly used descriptors [10] Several studies that using protein descriptors from R package have been conducted. The first, research by Lumbanraja, *et al.* [8] on identifying PTM lysine acetylation has been used as a combination of protein descriptors from protr and bioseq packages, namely CTD, Hydrophobicity, AAIndex, and APAAC. This research obtained the accuracy of up to 97.52% using the gaussian kernel of the SVM method. Other studies, Lumbanraja, *et al.* [11] also using CTD, AAIndex, pseudoAAC, and QSO from protr and bioseq package on the arginine methylation. They get an accuracy of 98.08%.

One common thing to do in protein identification research is only the length of the sequence of proteins used as input to the protein descriptor, which means that the output will only describe the state of the whole protein. Segmentation of protein sequences input into protein descriptors will provide information from the segment itself. The numerical representation of the entire protein and its segments will provide new information from the protein sequence. From the things stated, there is a gap between the previous research and this research. The previous research only used the entire sequence to be classified, but in this study, we also used the additional segment to obtain an overview of the entire sequence and additional segment. Faisal *et al.* [12] found a simple approach in dealing with the problem of the classification of these protein sequences. The approach is to build protein features from all proteins, adjacent segments, and overlapped segments rather than just the whole protein. From this study, it was found an increase in the accuracy of predictions.

The identification of lysine protein acetylation with computational methods still needs to be reviewed. This study will use lysine protein acetylation data, where the data also be segmented into adjacent and overlapped segments, then extracted with the protein descriptors from protr package, which are Amino Acid Composition and Dipeptide Composition. The SVM method was used in classifying the data using the gaussian kernel. This study will measure the performance of the protein descriptors and the methods mentioned, and to find out and test the use of adjacent and overlapped segments on protein sequences.

2. Data and Methods

2.1 Data

Protein is an amino acid polymer. All organisms use the same 20 amino acids as the building blocks of a protein molecule [13] PTMs are chemical modifications of a polypeptide chain that occur after DNA has been transcribed into RNA and translated into protein [1] One type of PTM is acetylation that occurs in the lysine protein. Lysine acetylation data used in this study came from the research of Huang *et al.* [14] which consist of data in the negative and the positive classes with a sequence length of 21. The total data of lysine protein sequence data is presented in Table 1.

Table 1. Total Data Lysine Protein.

Positive	Negative	Total
14.407	8.704	23.111

The data is still not clean and balanced, so it needs to be preprocessed first. Preprocessing is done in three steps, that is cleansing non-amino acid data, data redundancy, and data imbalance. Cleaning non-amino acid data is the cleaning of protein sequences that are not included in amino acids. The protein sequence, which contains the letter X, cannot be read and cannot be processed during feature extraction because it is not part of the protein sequence data, so the data must be deleted or cleaned. Cleaning this non-amino acid data for example removes the protein sequence contained in the letter X, “RALPRQDTVIKHYQRPAXXXX”. After cleaning, the protein sequence data redundancy, which at this stage will delete the protein sequences that have similarities using the implementation of the Skipredundant program. Skipredundant is carried out at 10% for each negative data and positive data with a gap extension penalty of 10,0 and a gap open penalty of 0,5. The final step is the imbalance data. data results from the previous preprocessing step there was an imbalance between negative data and positive data. To get the maximum classification results, the data needs to be balanced by removing some of the data in more data so that the two classes of data have a balanced amount. The preprocessing process has been conducted by Lumbanraja *et al.* [8] and we use this dataset for this study. The number of processed protein lysine sequence data is shown in Table 2.

Table 2. Total data after preprocessing.

Positive	Negative	Total
8.701	8.701	17.402

2.2 Adjacent Segment and Overlapped Segment

Adjacent and Overlapped Segment are built so that protein features have complete information, that is global information and local information. Adjacent segments are sequence divided into several segments with the same length. The first segment is calculated from the beginning of the sequence, then followed by the second segment and so on, or can be said *k* segments [12]. For example, given a protein sequence *s* as shown below: RKDAAEHTLTAYKAAQDIANS. If the sequence divide into *k* segments where *k* = 2, then the generated segments are shown in Figure 1.

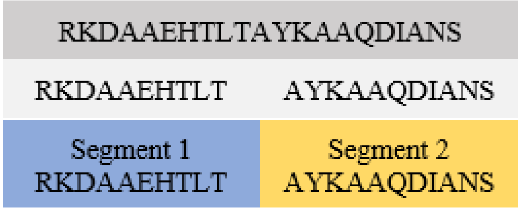


Figure 1. Adjacent Segment

The first adjacent segment is taken from the beginning of the original segment, which is RKDAAEHTLT. The second adjacent segment is taken from the last amino acid of the first segment, which is AYKAAQDIANS.

Overlapped segments are developed by adjacent segments and local information between two adjacent segments. An overlapped segment is the union of the half from the end of the first segment and a half from the beginning of the second segment. For example, an overlapped segment is generated as shown in Figure 2.

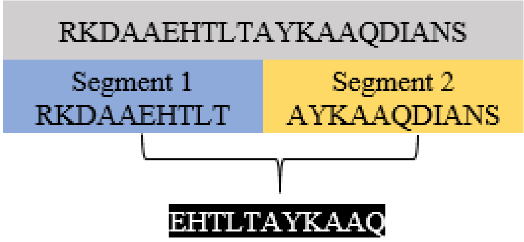


Figure 2. Overlapped Segment

The overlapped segment is created by using a half of first and a half of second adjacent segment, which is EHTLTAYKAAQ.

2.3 Protein Descriptor

The protein descriptors are local substructures of protein molecules, which allow us to divide the original problem into a set of subproblems and, consequently, to propose a more efficient algorithmic solution [9] Protr is currently the most comprehensive, flexible, and integrated open-source toolkit for protein sequence-derived structural and physicochemical descriptor computation. The descriptors are generally divided into eight groups. Amino Acid Composition and Dipeptide Composition are protein descriptors contained in the protr package. Number of these descriptors represented in Table 3 [10].

Table 3. Number of descriptors.

Descriptors	Number
Amino Acid Composition	20
Dipeptide Composition	400

2.3.1 Amino Acid Composition

Amino Acid Composition (AAC) describes the fractions of each type of amino acid in protein sequences. AAC gives 20 dimensions, defined as:

$$f(r) = \frac{N_r}{N} \quad r = 1,2,3\ldots,20 \tag{1}$$

where Nr is the number of the amino acid type r and N is the length of the sequence. AAC is implemented with the extractAAC() function of the protr package to extract features from a protein sequence [10]

2.3.2 Dipeptide Composition (DC)

Dipeptide Composition (DC) gives 400 dimensions, defined as:

$$f(r,s) = \frac{N_{rs}}{N-2} \quad r,s = 1,2,\ldots,20 \tag{2}$$

where Nrs is the number of the amino acid type r and s. DC implemented with the extractDC() function of the protr package to extract features from a protein sequence [10].

2.4 Support Vector Machine

Support vector machine (SVM) is a classification method that was first introduced by Vapnik in 1998. Basically, this method works by defining the boundary between two classes with the maximum distance from the closest data. To get the maximum distance between classes, a hyperplane must be formed at the input space obtained by measuring the hyperplane's margin and looking for the maximum point. Margin is the distance between the hyperplane and the closest point of each class. This closest point is called support vector machine [7] SVM can work in linear and non-linear cases. For the case of non-linear SVM used with kernel functions, including gaussian, signoid, linear, and polynomial. In this study, SVM used with gaussian kernel because, in the previous research [8], gaussian obtain the highest accuracy. Gaussian kernel in SVM is represented as follows [15]:

$$K(x,y) = exp\left(\frac{-||x-y||^2}{2.\sigma^2}\right) \tag{3}$$

2.5 k-Fold Cross Validation

Different cross-validation methods are available in the literature for sample selection as a training data set. The k-fold cross-validation method subdivided actual samples into k equal sized subsamples. Each subsample is taken as the validation data for testing the classification model and repeat the process k times. The advantage of this method is over repeated random sub-sampling as training and validation for each for validation at least once. Here k is the unfixed parameter will be chosen by the user [16]

2.6 Confusion Matrix

A confusion matrix contains information about actual and predicted classifications done by a classification system. For a binary class problem, a matrix is a square of 2x2 as shown in Table 4 [17]

Table 4. Actual and Predicted Table 2x2.

	Predicted Positive	**Predicted Negative**
Actual Positive	TP (number of True Positive)	FN (number of False Negative)
Actual Negative	FP (number of False Positive)	TN (number of True Negative)

The acronym TP, FN, FP, and TN of the confusion matrix cells refers to the following:
TP = true positive, the number of positive cases that are correctly identified as positive,
FN = false negative, the number of positive cases that are misclassified as negative cases,
FP = false positive, the number of negative cases that are incorrectly identified as positive cases,
TN = true negative, the number of negative cases that are correctly identified as negative cases
Table 5 presents fundamental evaluation metrics [17]s.

Table 5 Evaluation Metrics.

Measure	Formula
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
Sensitivity (or Recall)	$\frac{TP}{TP+FN}$
Specificity	$\frac{TN}{TN+FP}$
Mathews Correlation Coefficient	$\frac{TPxTN-FPxFN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$

3. Methodology

The steps taken in this research are shown in Figure 3.

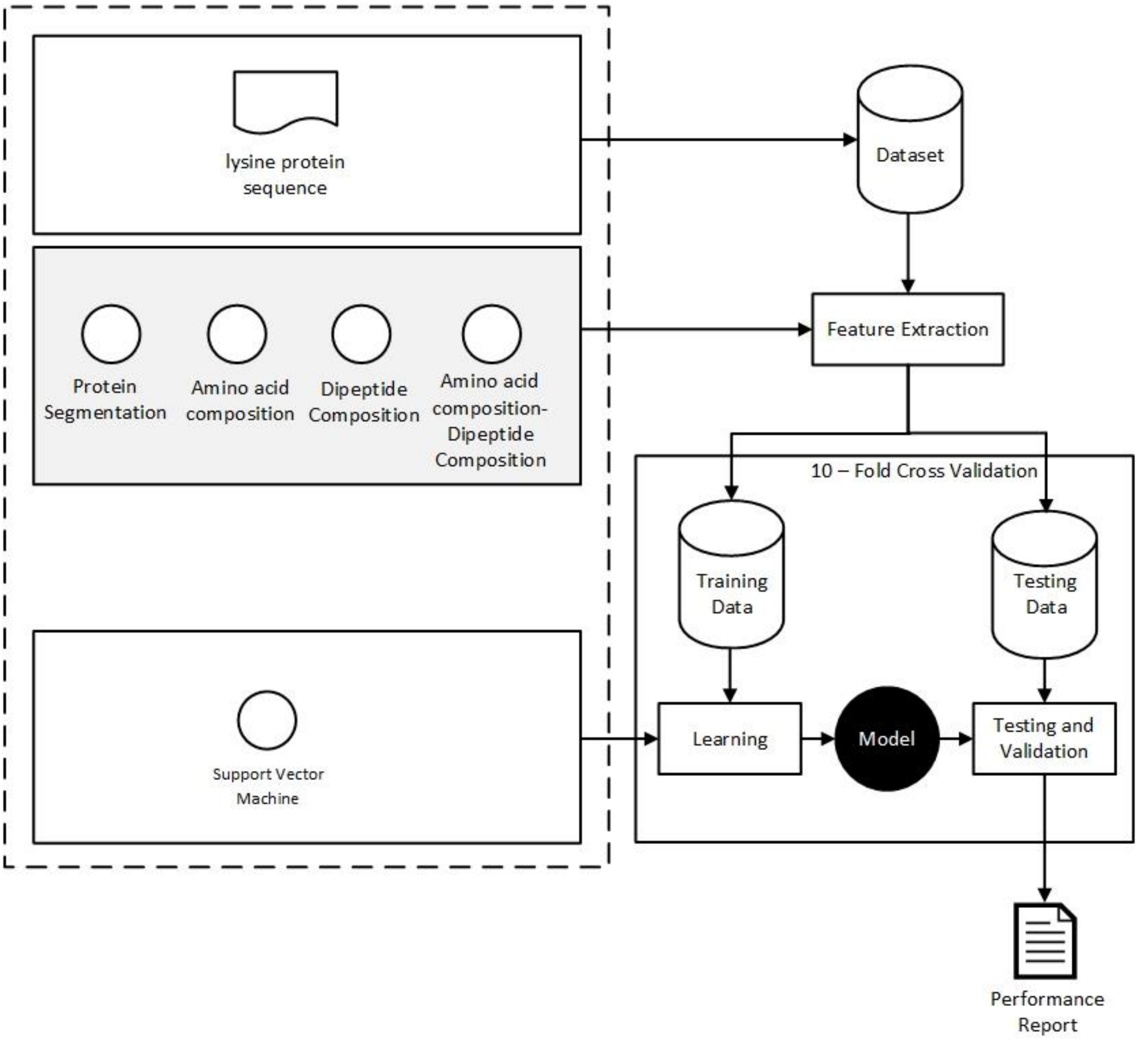


Figure 3. Research Flowchart

The flowchart explained as follows:

3.1 Dataset

Data used are lysine protein sequences that consisted of lysine acetylation and non-acetylation classes. Each of the classes collects 8701 data with a length of 21. This data was divided into training and testing data using k-Fold Cross Validation where each of iteration divides the data into 10 folds, 1 fold was used as a testing data and 9 folds were used as training data. from the amount of the data, the number of 1 fold is 1740 and the number of 9 fold is 15662.

3.2 Feature Extraction

The first step, feature extraction divide data into several segments with adjacent segment and overlapped segment. Sequence segmentation is illustrated in Figure 4, where the original sequence divides into two segments and represented by concatenating all feature representations. Then, feature extraction change the original sequence, adjacen, and overlapped segments into numerical features using the Amino Acid Composition (AAC) and Dipeptide Composition (DC) protein descriptors from the protr package [10]

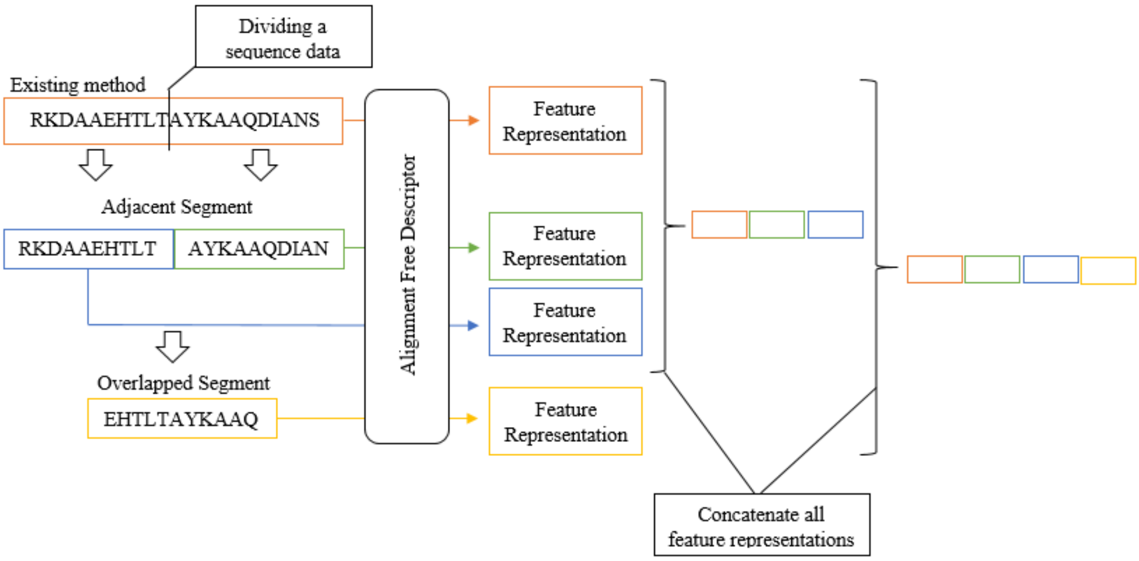


Figure 4. Sequence Segmentation for Feature Representation

3.3 Separate Training and Test Data

k-Fold Cross Validation is used to divide the number of the dataset as much as k parts and in this study using k = 10. Then, the data will be used as test data and training data with a ratio of 1:9 and repeated 10 times. 10-fold cross validation is illustrated in Figure 5.

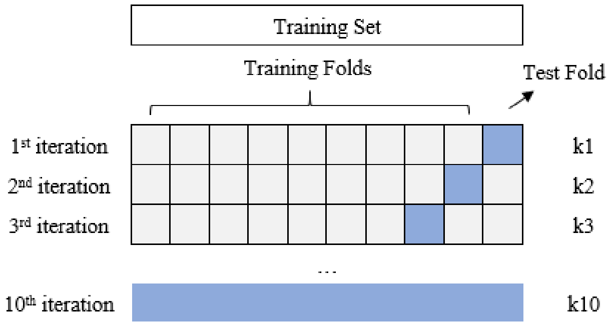


Figure 5. 10-fold cross validation

3.4 Classification

Classification models are created using the Support Vector Machine method with gaussian kernel only. We used e1071 package [18] from R to create a classification model.

3.5 Evaluation

After being classified, the model will be tested by confusion matrix calculation, which is calculating the performance of accuracy, specificity, sensitivity, and Mathews Correlation Coefficient (MCC). The confusion matrix is used with the caret package [19] and MCC with the mltools package [20]

4. Experiment and Result

Acetylation classification on lysine protein with the support vector machine method gaussian kernel was carried out with several experiments, where the data used was built into three kinds of data. The data is original sequence data, sequence data built with adjacent segments, and sequence data built with overlapped segments. Original sequence data is data that is not segmented. Data sequences are built with adjacent segments using *k* = 2, so that the number of features to be built is multiplied by three. For sequence data built with overlapped segments with adjacent segment *k* = 2, the overlapped segment is 1, so the number of features to be built is multiplied by four. Each of these data will be extracted features with Amino Acid Composition (AAC) Dipeptide Composition (DC), and a combination of both to find out how the influence given by adjacent and overlapped segments on the performance of feature extraction. In addition, the results that will be obtained will also provide knowledge about how the performance of the two feature extractions used. The number of features resulting from the three feature extractions is illustrated in Table 6.

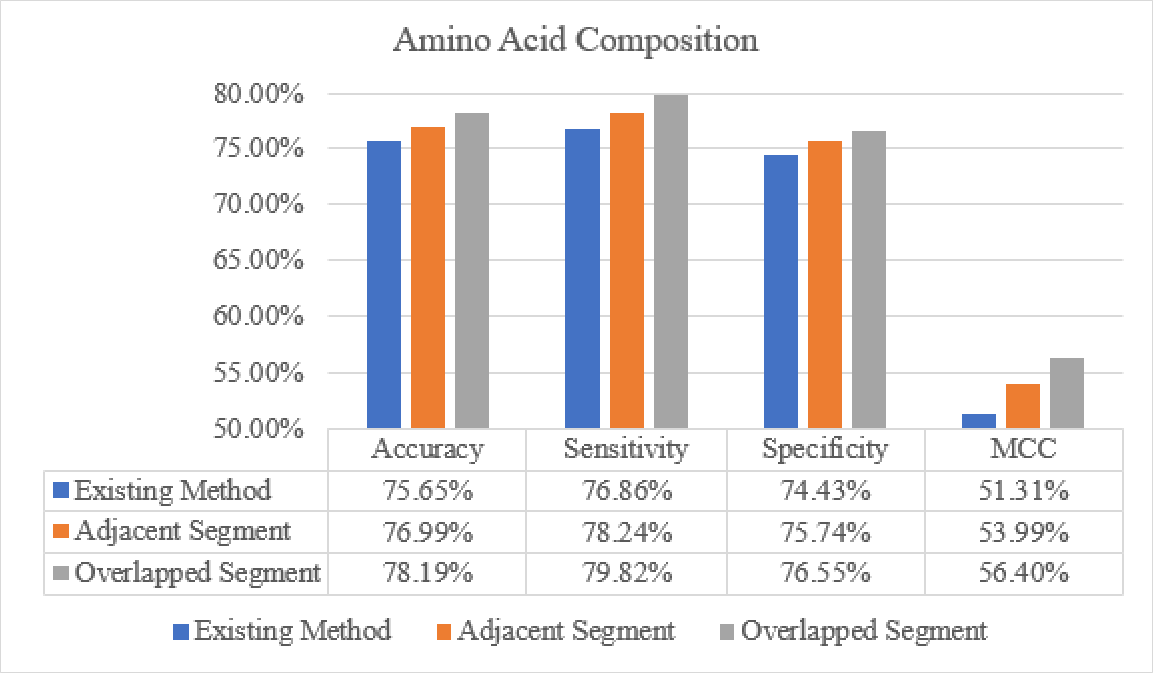
Table 6 Number of Features.

	Original Sequence	Adjacent Segment	Overlapped Segment
Amino Acid Composition (AAC)	20	60	80
Dipeptide Composition (DC)	400	1200	1600
AAC-DC	420	1260	1680

The nine types of data are each classified using the Gaussian kernel SVM method with 10-fold cross-validation and tested by calculating accuracy, specificity, sensitivity, and MCC. Figure 6 displays the results of AAC descriptor testing on original data, adjacent segments, and overlapped segments. The results of the experiment with AAC protein descriptor obtained accuracy in the original sequence by 75.65%, then accuracy increased in the adjacent segment to 76.99%, and rose again in the overlapped segment which was 78.19%. The same cases, sensitivity, specificity, and MCC also increased. Improved accuracy also occurred in experiments with DC feature extraction shown in Figure 7.

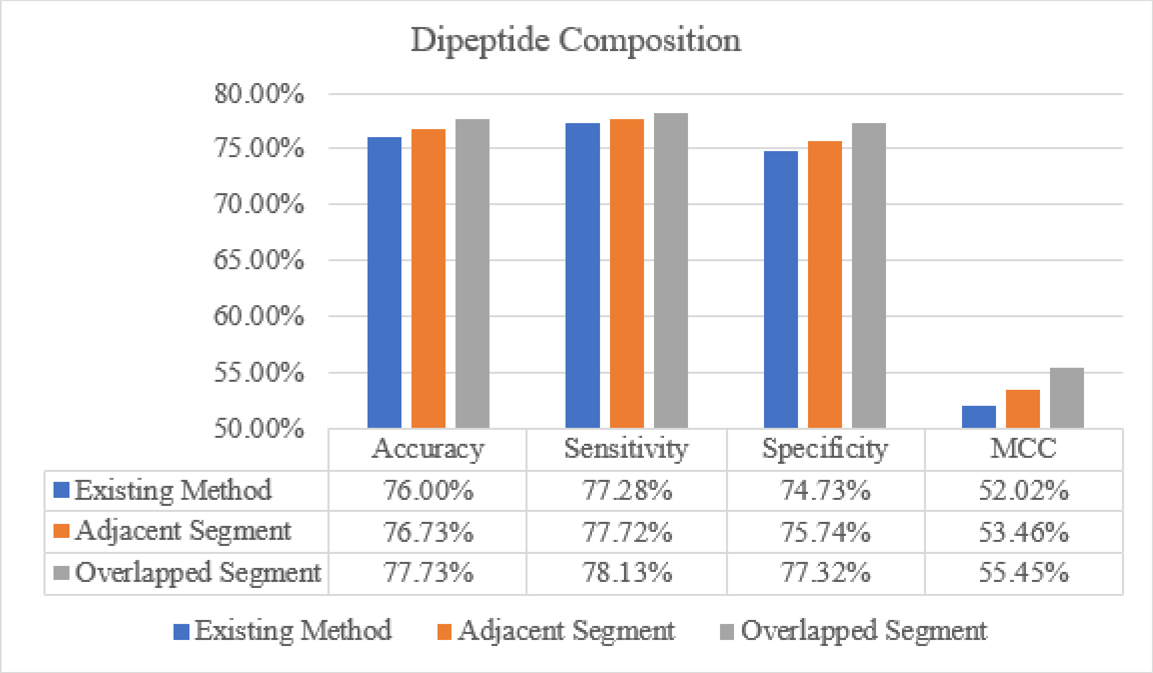
Increased accuracy of DC feature extraction obtained 76% accuracy in the original sequence, 76.73% in the adjacent segment, and 77.73% in the overlapped segment. Improvements also occurred in testing sensitivity, specificity, and MCC. The results of AAC-DC protein descriptor testing on the original data, adjacent segments, and overlapped segments are shown in Figure 8.

The combination of AAC and DC protein descriptors affects the accuracy of the original sequence data. In addition, increased accuracy also occurs when the combined feature extraction is performed on adjacent and overlapped segment data. The original data obtained an accuracy of 76.20%, an adjacent segment of 77.33%, and an overlapped segment of 78.08%. Measurement of sensitivity, specificity, and MCC also increased. The combination of AAC and DC protein descriptors affects the accuracy of the original sequence data. In addition, increased accuracy also occurs when the combined feature extraction is performed on adjacent and overlapped segment data. The original data obtained an accuracy of 76.20%, and the adjacent segment of 77.33%, and an overlapped segment of 78.08%. Measurement of sensitivity, specificity, and MCC also increased.



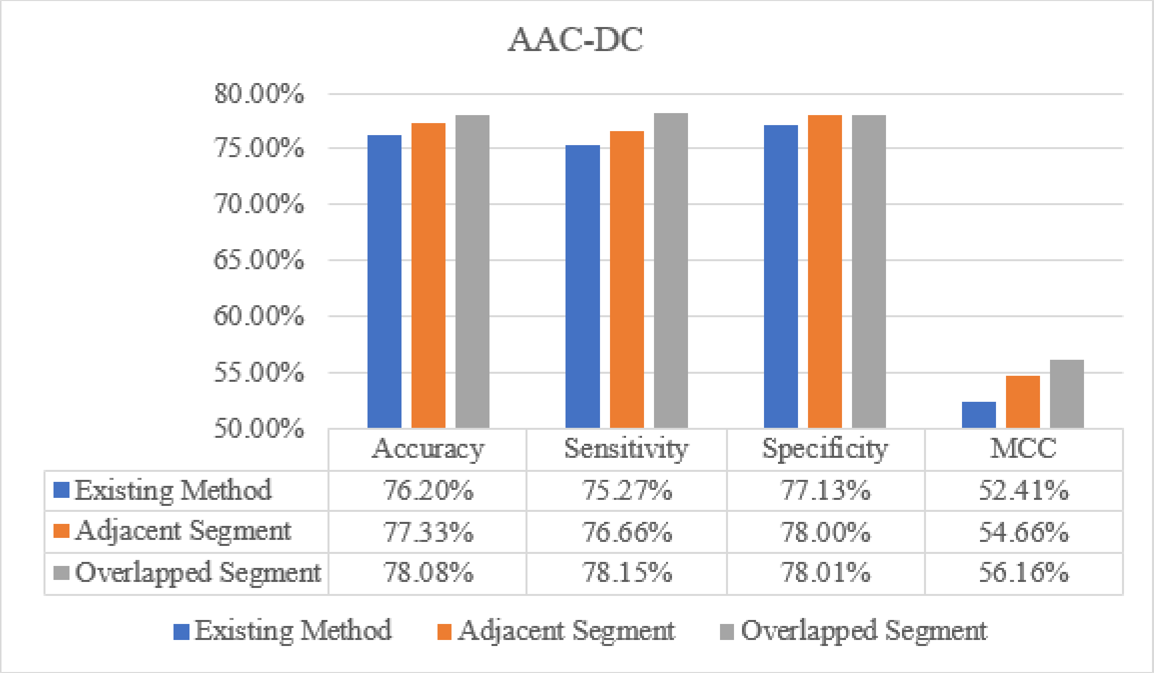
	Accuracy	Sensitivity	Specificity	MCC
🟦 Existing Method	75.65%	76.86%	74.43%	51.31%
🟧 Adjacent Segment	76.99%	78.24%	75.74%	53.99%
⬜ Overlapped Segment	78.19%	79.82%	76.55%	56.40%

Figure 6. Testing Result of AAC Protein Descriptor



	Accuracy	Sensitivity	Specificity	MCC
🟦 Existing Method	76.00%	77.28%	74.73%	52.02%
🟧 Adjacent Segment	76.73%	77.72%	75.74%	53.46%
⬜ Overlapped Segment	77.73%	78.13%	77.32%	55.45%

Figure 7. Testing Result of DC Protein Descriptor



	Accuracy	Sensitivity	Specificity	MCC
🟦 Existing Method	76.20%	75.27%	77.13%	52.41%
🟧 Adjacent Segment	77.33%	76.66%	78.00%	54.66%
▨ Overlapped Segment	78.08%	78.15%	78.01%	56.16%

Figure 8. Testing Result of AAC-DC Protein Descriptor

Accuracy provides an assessment of the accuracy of the model in classifying, sensitivity provides an assessment of how often a model can correctly predict positive data, specificity gives an assessment of how often a model can correctly predict negative data, and MCC provides an assessment of the quality measures of binary classification. From the experiments that have been described, segmentation gives an increase in the test value for each descriptor protein that is tested with SVM. Overall, the protein segmentation in the form of an overlapped segment provides a significant increase in accuracy performance, as well as the adjacent segment. Before being segmented, which is when using the existing method, the combination descriptor AAC and DC gave the best performance, which was 76.20%. However, after being segmented, the performance of all descriptors increased by 0.7-2.5%, and the best performance of the segmented data was obtained from the AAC descriptor which was 78.19%. Not only accuracy, but protein segmentation also improves sensitivity, specificity, and MCC performance. The best sensitivity performance was obtained from the AAC descriptor (79.82%), the best specificity was obtained from the combination of AAC and DC descriptors (78,01%), and the best MCC was obtained from the AAC descriptors (56,40%), and all of the best performance was obtained from the use of protein segmentation, namely the overlapped segment. So, it can be said that this protein segmentation technique has a better performance than the existing methods.

5. Conclusion

Amino Acid Composition (AAC) and Dipeptide Composition (DC) are two of the 21 types of protein descriptors in the protr package used to extract protein sequences into numerical data. AAC, DC, and a combination of both achieve good classification accuracy by testing on three types of data. In the original data, the combination of AAC and DC achieves the highest accuracy compared to other protein descriptors, which is 76.20%. In the adjacent segment and overlapped segment data, the highest accuracy was obtained in the AAC protein descriptor, respectively 76.99% and 78.19%. Based on this, the AAC-DC combination works best on the original data. But if the protein sequence is segmented by adjacent and overlapped segments, the protein descriptor that works best is the Amino Acid Composition (AAC). Protein segmentation has a good effect on the performance of each descriptor, where this effect can be seen from a significant increase in accuracy in all protein descriptors. The intended segmentation is the adjacent segment and the overlapped segment. Overall, the highest accuracy was obtained from the AAC protein descriptor on the overlapped segment data of 78.19%. In addition to accuracy, segmentation also influences the calculation of specificity, sensitivity, and MCC were the values obtained increase, which means that the model constructed predicts data more precisely that before.

In this study, we only focus on solving the problem of classifying protein sequences using the Amino Acid Composition and Dipeptide Composition protein descriptors by using data segmentation approaches, namely adjacent segments and overlapped segments. Adjacent and overlapped segments give users the freedom to determine the parameter *k* and in this study use *k* = 2. For future research, we suggest adjacent segment and overlapped segment approaches to use the larger parameter k, for example, three and so on.

Acknowledgments

The experiment in this research was provided by Computer Lab at the Department of Computer Science Lambung Mangkurat University. This research is a collaboration between the Department of Computer Science Lambung Mangkurat University and Lampung University.

Reference

[1] Uversky VN 2013 Posttranslational Modification. Vol. 5, Brenner's Encyclopedia of Genetics: Second Edition. Elsevier Inc. p. 425–430.

[2] Green KD and Garneau-tsodikova S. 2010 5.15 Posttranslational Modification of Proteins. Comprehensive Natural Products II. *Elsevier Inc.* p. 433–68.

[3] Drazic A, Myklebust LM, Ree R, and Arnesen T 2016 Biochimica et Biophysica Acta The world of protein acetylation. *BBA - Proteins Proteomics*, **1864**(10). p. 1372–401

[4] Hou T, Zheng G, Zhang P, Jia J, Li J, Xie L, Wei C, and Li Y 2014 LAceP: Lysine acetylation site prediction using logistic regression classifiers. *PLoS One*, **9**(2).

[5] Chen Z, Liu X, Li F, Li C, Marquez-Lago T, Leier A, Akutsu T, Webb GI, Xu D, Smith AI, Li L, Chou KC, and Song J 2019 Large-scale comparative assessment of computational predictors for lysine post-translational modification sites. *Brief Bioinform*, **20**(6). p. 2267–90.

[6] S.Neelamegam and E.Ramaraj 2013 Classification Algorithm in Data mining: An Overview. *Int J Netw Trends Technol* 2013, **3**(5).

[7] Rizal RA, Girsang IS and Prasetiyo SA 2019 Klasifikasi Wajah Menggunakan Support Vector Machine (SVM). *Ris dan E-Jurnal Manaj Inform Komput*, **3**(2). p. 1–5.

[8] Lumbanraja FR, Silalahi EDP, Kurniawan D and Syarif A 2019 Prediksi Posisi Asetilasi pada Protein Lisin menggunakan Support Vector Machine. In: Seminar Nasional Sains, Matematika, Informatika, dan Aplikasinya. Lampung: FMIPA Universitas Lampung. p. 95–104.

[9] Antczak M, Kasprzak M, Lukasiak P, and Blazewicz J 2016 Structural alignment of protein descriptors - a combinatorial model. *BMC Bioinformatics*, **17**(1). p. 1–14.

[10] Xiao N, Cao DS, Zhu MF, and Xu QS 2015 Protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences. *Bioinformatics*, **31**(11). p. 1857–9.

[11] Lumbanraja FR, Mudyaningsih W, Hermanto B, and Syarif A 2019 Implementasi Metode Random Forest untuk Prediksi Posisi Metilasi pada Sekuens Protein. In: Seminar Nasional Sains, Matematika, Informatika, dan Aplikasinya. Lampung: FMIPA Universitas Lampung. p.

105–12.
[12] Faisal MR, Abapihi B, Nguyen NG, Purnama B, Delimayanti MK, Phan D, Lumbanraja FR, Kubo M, and Satou K 2018 Improving Protein Sequence Classification Performance Using Adjacent and Overlapped Segments on Existing Protein Descriptors. *J Biomed Sci Eng*, **11**(06). p. 126–43.
[13] Azhar M 2016 *Biomolekul sel*. Padang: UNP Press. p. 125–130.
[14] Huang K, Su M, Kao H, Hsieh Y, Jhong J, Cheng K, Huang H, and Lee T 2016 dbPTM 2016 : 10-year anniversary of a resource for post-translational modification of proteins. *Nucleic Acids Res*, **44**. p. 435–46.
[15] Mase J, Furqon MT, and Rahayudi B 2018 Penerapan Algoritme Support Vector Machine (SVM) Pada Pengklasifikasian Penyakit Kucing. *J Pegembangan Teknol Inf dan Ilmu Komput*, **2**(10). p. 3648–54.
[16] Raju KS, Murty MR, Rao MV, and Satapathy SC 2018 Support Vector Machine with K-fold Cross Validation Model for Software Fault Prediction. *Int J Pure Appl Math*, **118**(20). p. 321–34.
[17] Bekkar M, Djemaa HK, and Alitouche TA 2013 Evaluation Measures for Models Assessment over Imbalanced Data Sets. *J Inf Eng Appl*, **3**(10). p. 27–39
[18] Meyer D, Dimitriadou E, Hornik K, Weingessel A and Leisch F 2019 e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071).
[19] Kuhn M 2020 caret: Classification and Regression Training.
[20] Gorman B. 2018 mltools: Machine Learning Tools.