IMPLEMENTATION OF AAC AND DC FEATURE EXTRACTION FOR CLASSIFICATION OF LYSINE PROTEIN ACETYLATION USING THE *SUPPORT* VECTOR MACHINE METHOD

Annisa Rizqiana[1], Mohammad Reza Faisal[2], Favorisen R Lumbanraja[3], Muliadi[4], Rudy Herteno[5]
[1245]Ilmu Komputer FMIPA ULM
[3]Ilmu Komputer FMIPA UNILA
A. Yani St. KM 36 Banjarbaru, South Kalimantan
Email: 1611016220003@ulm.ac.id

Abstract

Post-Translational Modification (PTM) is a change that occurs in the chemical structure of a protein. One type of PTM is acetylation which commonly occurs in lysine proteins where this type of PTM plays an important role in biological processes. Existing research has identified lysine acetylation using computational methods, which is classification. Methods for protein classification have been developed, but much remains to be explored to identify lysine acetylation. Protein classification begins with extracting protein sequences into numerical features with protein descriptors which in this study used Amino Acid Composition (AAC) and Dipeptide Composition (DC). Furthermore, protein classification is carried out using the Support Vector Machine method. Support Vector Machine is a classification method that can be used for protein identification. This study provides the best performance results on the use of the combination of AAC and DC descriptors, which is 76.20%.

Keywords: *lysine acetylation, Amino Acid Composition, Dipeptide Composition, protein classification, Support Vector Machine*

1. INTRODUCTION

Post-Translational Modification (PTM) is a chemical change in protein that may be experienced after translation. [1] explain that Post-Translational Modification diversifies a finite set of amino acids, expanding 20 amino acids into an infinite number of possible residues. One type of PTM that is needs is acetylation. Acetylation is one of the most significant post-translational protein modifications, and plays an important role in a variety of cellular processes, such as cytokine signaling (response to immune stimuli), transcriptional regulation (the way cells regulate the conversion of DNA to RNA), and apoptosis (mechanism of death. programmed cells). Acetylation usually occurs in lysine residues which explains the process of inserting an acetyl group (CH3CO) into the amino acid side chain in protein [2]. Lysine is useful in helping calcium absorption, hormone and collagen formation, and antibodies. The acetylation that occurs in lysine can repair DNA damage, transcription and gene expression.

According to [3], identification of protein acetate sites through traditional (in vitro) experimental methods is time consuming and laborious. So it is not suitable for identifying large quantities of acetylation. Computational theory which is a field of computer science is needed in this problem, where the computational method (in silico) itself is a way to find solutions to problems using a certain algorithm with the

help of a computer. Therefore, computational methods are needed to accelerate the process of identifying and predicting PTM acetylation in lysine protein.

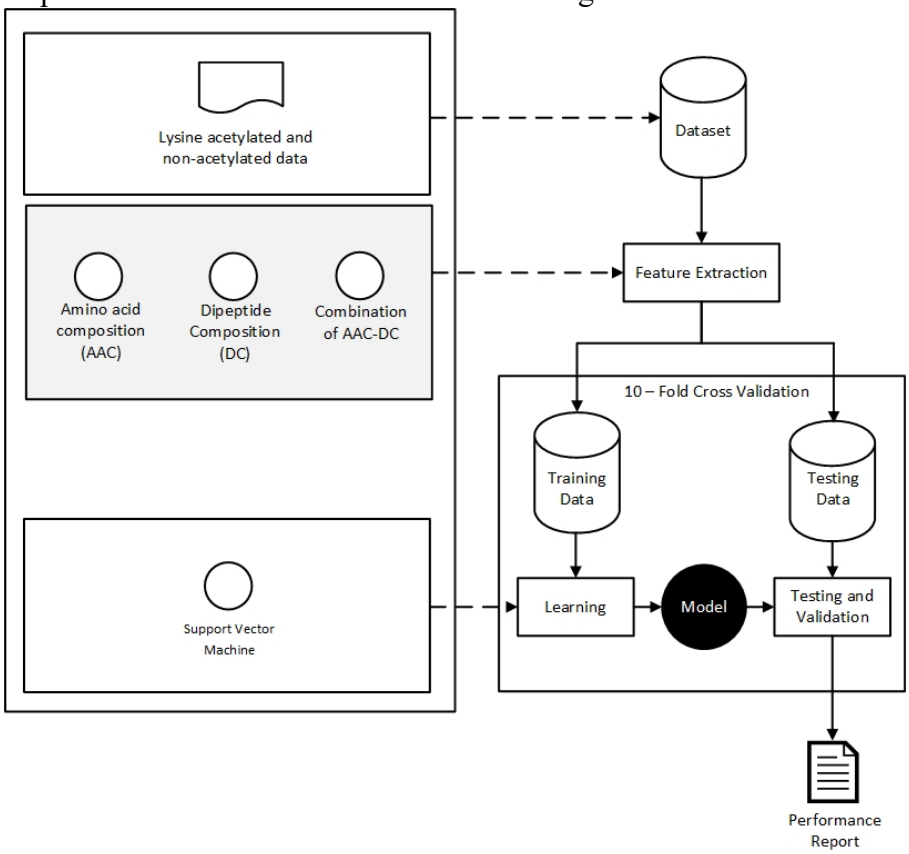
One of the computational methods that can be used is classification. Classification is the process of finding a model or function that can describe the class of data. The data used in the classification must be structured data and not protein. Therefore, prior to classification, the data must be extracted first with feature extraction called Protein Descriptor. Protein Descriptor is a feature extraction that is used to convert protein sequence data into structured data.

Previous research on the identification of PTM in protein sequences was conducted by [4] who identified methylation of arginine proteins using protr protein descriptors, namely a combination of CTD, AAindex, PseudoAAC, and QSO, as well as the Random Forest classification method which resulted in an accuracy of 98.08% but was ineffective because of the imbalance of data each class. In other research, [5] identified acetylation of lysine protein using a combination of protein descriptors, namely CTD, Hydrophobicity, AAindex, and APAAC, as well as the Support Vector Machine classification method with 3 kernels and obtained the best results in the Gaussian kernel with an accuracy value of 97.52%.

There is still a lot of research on the prediction of PTM that can be explored again. In this study, the acetylation of lysine protein sequences was classified using the feature extraction of Amino Acid Composition (AAC), Dipeptide Composition (DC), and a combination of AAC and DC, using the Gaussian kernel Support Vector Machine classification method. Through this research, it is expected to produce a new technique that can identify PTM acetylation in lysine protein.

2. RESEARCH METHODOLOGY

The procedure of this research are shown in Figure 1.



2.1 Dataset

The data used in this study is a dataset of acetylated and non-acetylated lysine proteins. This dataset comes from research [6] which was further preprocessed by [5]. The amount of data obtained from the research of [6] is shown in Table 1.

Table 1 Amount of data before preprocessing

Positive	Negative	Total
14407	8704	23111

The preprocessing carried out by (5) went through 3 stages, namely removing non-amino acid data, data redundancy, and data imbalance. Removing non-amino acid data is deleting protein sequence data containing non-amino acid data such as the letter “X” in the protein sequence "RALPRQDTVIKHYQRPAXXXX". Data deletion that is not amino acid is performed because “X” is not an amino acid, so it will not be read and extracted in the next step, that is feature extraction. Furthermore, data redundancy is also carried out to delete protein sequence data that has similarities. Because there is data imbalance, deleting data from certain classes is done to balance the two data classes. Table 2 shows the amount of data that has been preprocessed and used in this study. An example of the data used in this study is shown in Table 3.

Table 2 Amount of data after preprocessing

Positive	Negative	Total
8701	8701	17402

Table 3 An example of the data

Identifier	Sequence	Class
1A1L1_HUMAN_418	AGFFIWVDLRKYLPKGTFEEE	Negative
1A68_HUMAN_92	EYWDRNTRNVKAQSQTDRVDL	Negative
S6FQI0_9BACI_258	FFDIDTKYYTKELHKAAFVLP	Positive
PPIH_HUMAN_166	NVPTGPNNKPKLPVVISQCGE	Positive

2.2 Feature Extraction

The data is first extracted by feature extraction which is called protein descriptors in the protein sequence. The data was extracted to determine the characteristics of each data class. Protein descriptors used in this study were descriptors of the protr package, namely Amino Acid Composition (AAC), Dipeptide Composition (DC), and AAC-DC combination.

2.2.1 Amino Acid Composition

Amino Acid Coposition (AAC) describes the fraction of each amino acid with the formula:

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

N_r is the number of type r amino acids and N is the number of long protein sequences. AAC is implemented using the extractAAC () function included in the protr package.

2.2.2 Dipeptide Composition

Dipeptide Composition (DC) describes the fraction of each amino acid with the formula:

$$f(rs) = \frac{N_{rs}}{N-2} \quad r, s = 1, 2, 3, \dots, 20$$

N_{rs} is the number of combined amino acids of types r and s. DC is implemented using the extractDC () function found in the protr package.

2.2.3 Combination of AAC-DC

This descriptor is a feature amalgamation of the AAC and DC descriptors.

2.3 k-Fold Cross Validation

k-Fold Cross Validation is a method for dividing sample data into training data and test data into k sections. In this study, the parameter k used is 10, so the data will be divided into 10 parts.

2.4 Classification

After the data is divided into training data and test data, the training data is used as a learning model using the classification method. The classification in this study uses the Support Vector Machine (SVM) method. SVM works by defining the boundaries between two data classes with the maximum distance from the closest data. The constraint in question is a hyperplane (dividing line). For a high dimensional feature space, SVM provides kernel functions, one of which is the Gaussian kernel. The Gaussian kernel formula is as follows:

$$\text{Gaussian} = e^{-\gamma(a-b)^2}$$

In this study, the e1071 package was used to create a classification model using the SVM method.

2.5 Evaluation

After being classified, the model is tested for its performance by calculating the confusion matrix, namely testing the performance of accuracy, sensitivity, specificity, and Mathews Correlation Coefficient (MCC). A confusion matrix is included in the Caret package used in this study. The formula for the performance test is presented in Table 4.

Table 4 Formula of the Confusion Matrix

Kinerja	Formula
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$
Sensitivity	$\frac{TP}{TP + FN}$
Specificity	$\frac{TN}{TN + FP}$
MCC	$\frac{(TP\ x\ TN) - (FP\ x\ FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$

3. RESULTS AND DISCUSSION

3.1 Results

The data obtained were extracted with protein descriptors Amino Acid Composition (AAC) and Dipeptide Composition (DC), as well as a combination of both. Each descriptor produces a different number of features, namely AAC produces 20 features, DC 400 features, and AAC-DC 420 feature combination. The results of feature extraction with the AAC descriptor are presented in Table 5, DC in Table 6, and AAC-DC combination in Table 7.

Table 5 AAC feature extraction results

No	A	R	…	Y	V	Class
1	0.28571	0.04762	…	0.04762	0	Negative
2	0.14286	0	…	0	0.14286	Negative
3	0.04762	0.04762	…	0.04762	0.04762	Negative
…	…	…	…	…	…	…
17400	0.04762	0	…	0.04762	0	Positive
17401	0.04762	0.09524	…	0.04762	0.04762	Positive
17402	0.04762	0.09524	…	0	0.09524	Positive

Table 6 DC feature extraction results

No	AA	RA	…	YV	VV	Class
1	0.1	0	…	0	0	Negative

2	0	0	…	0	0.05	Negative
3	0	0	…	0	0	Negative
…	…	…	…	…	…	…
17400	0	0	…	0	0	Positive
17401	0	0	…	0	0	Positive
17402	0	0.05	…	0	0	Positive

Table 7 Combination of AAC-DC feature extraction results

No	A	R	…	YV	VV	Class
1	0.28571	0.04762	…	0	0	Negative
2	0.14286	0	…	0	0.05	Negative
3	0.04762	0.04762	…	0	0	Negative
…	…	…	…	…	…	…
17400	0.04762	0	…	0	0	Positive
17401	0.04762	0.09524	…	0	0	Positive
17402	0.04762	0.09524	…	0	0	Positive

Furthermore, protein data that has been extracted with each descriptor is divided into training data and test data with the 10-Fold Cross Validation rule where the data is divided into 10 parts. 1 part as test data and the other 9 parts as training data. The shared data is illustrated in Figure 1.

Iteration	Subset data									
	1	2	3	4	5	6	7	8	9	10
	1740	1740	1741	1740	1741	1740	1740	1740	1740	1740
1	■									
2		■								
3			■							
4				■						
5					■					
6						■				
7							■			
8								■		
9									■	
10										■

Figure 1 Separating dataset using k-fold cross validation

Furthermore, to create a learning model, the training data is classified using the Support Vector Machine Gaussian kernel method and tested using confusion matrix calculations. The performance tested were accuracy, specificity, sensitivity, and MCC. The test results are shown in Figure 2.

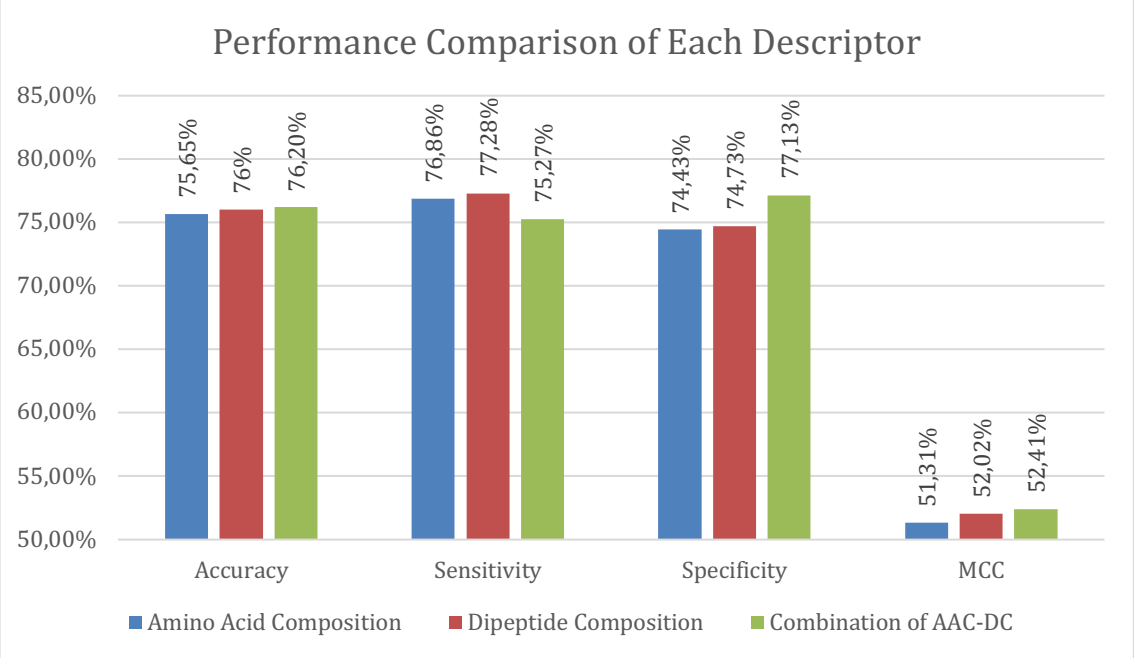


Figure 2 Performance chart

3.2 Discussion

From the three experiments conducted on three kinds of protein descriptors, the results obtained in the form of accuracy, sensitivity, specificity, and MCC for each descriptor. Overall, the best accuracy was obtained from the AAC-DC combination descriptor, namely 76.20%. Similar to accuracy, the AAC-DC combination also got the best performance on the specificity and MCC tests where the AAC-DC combination got 77.13% for the specificity and 52.41% for MCC. The three performance test scores on this AAC-DC combination descriptor are higher than the other two descriptors. Meanwhile, the best sensitivity performance was obtained from the DC descriptor. This means that the combination descriptor AAC-DC is the best at predicting negative and positive data correctly which can be seen from the accuracy value, and the best at predicting negative data correctly which can be seen from its specificity performance. On the other hand, in predicting positive data correctly, the AAC-DC combination gave the lowest value and DC obtained the highest value measured from its sensitivity value.

Overall, the three descriptors used in this study can be used to identify PTM in protein, namely acetylation of lysine protein. However, the best performance is obtained from the use of combination descriptors AAC and DC.

4. CONCLUSION

Based on the results of the research and discussion that has been described, it can be concluded that the three descriptors namely Amino Acid Composition (AAC), Dipeptide Composition (DC), and the combination of AAC-DC with the implementation of the Support Vector Machine classification method can be used in identifying PTM protein in this study. is the acetylation of lysine protein. The performance obtained from each descriptor is in the value range of 70-80% with the best accuracy value obtained from the AAC-DC combination descriptor, which is

76.20%. For future research, it is suggested to use protein descriptors or other classification methods to find out how the performance results from these other descriptors.

REFERENCES

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

[2] Hou, T. Guangyong, Z., Pingyu, Z., Jia, J., Jing, L., Lu, X., Chaochun, W., & Yixue, L., 2014. LAceP: Lysine acetylation site prediction using logistic regression classifiers. PLoS ONE, 9(2).

[3] Huang, K. Y., Su, M., Kao, H., Hsieh, Y., Jhong, J., Cheng, K., Huang, H., & Lee, T., 2016. dbPTM 2016: 10-year anniversary of a resource for posttranslational modification of proteins. Nucleic Acids Research, 44.

[4] Lumbanraja, F. R., Mudyaningsih, W., Hermanto, B., & 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.

[5] Lumbanraja, F. R., Silalahi, E. D. P., Kurniawan, D., & 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

[6] Wuyun, Q., Zheng, W., Zhang, Y., Ruan, J., & Hu, G., 2016. Improved SpeciesSpecific Lysine Acetylation Site Prediction Based on a Large Variety of Features Set. PLoS ONE, 11(5), pp. 1–21.