ACCURATELY PREDICTING PROTEIN ACETYLATION SITE USING
CONVOLUTIONAL NEURAL NETWORK

By

KANGAN HU

A thesis submitted to the

Graduate School-Camden

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of Master of Science

Graduate Program in Computer Science

Written under the direction of

Iman Dehzangi

And approved by

_________________________
Iman Dehzangi

_________________________
Desmond Lun

_________________________
Sunil Shende

Camden, New Jersey

January 2022
THESIS ABSTRACT

Accurately Predicting Protein Acetylation Site Using Convolutional Neural Network

by KANGAN HU

Thesis Director:
Iman Dehzangi

Post Translational Modification (PTM) is an important regulatory mechanism and plays an important role in normal and pathological cells. A wide range of PTM types including ubiquitination, methylation, crotonylation, and acetylation have been identified so far. Among these PTMs, acetylation is one of the most important ones as it is associated with specific human diseases like hypertension, arrhythmia, heart failure, and angiogenesis. Experimental methods to detect acetylation in proteins include radioactive chemical method, chromatin immunoprecipitation, and mass spectrometry. However, all these methods are time-consuming and costly. Therefore, finding fast and effective computational approaches to effectively detect acetylation sites is attracting tremendous attention.

In this study, we propose a new machine learning approach to enhance the lysine acetylation prediction performance. First, we used two different traditional machine learning classifiers (RF and SVM) and use both structural and evolutionary features to investigate the effectiveness of the traditional machine learning classifiers on the prediction of lysine acetylation sites. Our results for this step demonstrate that SVM has better performance than RF in almost all aspects in both human and mouse samples. We
then investigate related attributes that are important in the prediction task such as the training and testing ratio and the impact of our employed features. As a result, we identify that the 5:1 training and testing ratio demonstrates the best performance compared to other ratios. We also show that the combination of evolutionary and structural features demonstrates better results than using each one, separately.

Second, we improved our model by employing Convolutional Neural Networks (CNN) to our extracted features for lysine acetylation sites prediction. By comparing the results to RF and SVM, we demonstrate that CNN achieves average 0.05 improvement over RF and average 0.04 improvement over SVM in MCC. We then compared our CNN model to two state-of-the-art models proposed in recent years. Results demonstrate that our model achieves 0.04 and 0.09 improvements in term of MCC for these two models, respectively.

In conclusion, we develop an effective and accurate computational predictor that enables us to identify lysine acetylation sites better than previously proposed methods found in the literature.
# Table of Contents

Abstract .............................................................................................................................................. ii

Table of Contents ................................................................................................................................. iv

List of Illustrations ............................................................................................................................... vi

Chapter 1 Introduction ......................................................................................................................... 1

1.1 PTM .............................................................................................................................................. 1

1.2 Machine Learning ......................................................................................................................... 3

Chapter 2 Literature review ............................................................................................................... 8

2.1 Overview ....................................................................................................................................... 8

Chapter 3 Dataset/Features .............................................................................................................. 13

3.1 Dataset ........................................................................................................................................ 13

3.2 Feature extraction ....................................................................................................................... 15

3.2.1 Structural features .................................................................................................................. 15

3.2.2 Evolutionary features ............................................................................................................. 17

Chapter 4 Traditional machine learning .......................................................................................... 19

4.1 Introduction ................................................................................................................................. 19

4.2 Results ........................................................................................................................................ 20

4.2.1 Performance Metrics ............................................................................................................. 20

4.2.2 Comparison of RF and SVM ................................................................................................. 21
4.2.3 Comparison of different ratio of training and testing data ........................................... 22
4.2.4 Comparison of different feature extractions ................................................................. 22
4.3 Conclusion ..................................................................................................................... 23

Chapter 5 Deep learning ...................................................................................................... 24
5.1 Introduction .................................................................................................................... 24
5.2 Results .......................................................................................................................... 25
5.2.1 Comparison of CNN and traditional machine learning classifier ......................... 25
5.2.2 Comparison of our CNN model with state-of-the-art models ............................... 26
5.3 Conclusion .................................................................................................................... 27

Chapter 6 Conclusion .......................................................................................................... 29
6.1 Contributions ................................................................................................................ 29
6.2 Future Works ............................................................................................................... 30

Reference ............................................................................................................................ 32
LIST OF ILLUSTRATIONS

Figure 1.1. Some of the most common post-translational modifications (Retrieved from https://www2.mrc-lmb.cam.ac.uk/ccsb/research/post-translational-modifications) ........2

Figure 1.2. Arthur Samuel demonstrates how machine learning can be used to play checkers in 1962 (Retrieved from https://www.ibm.com/ibm/history/ibm100/us/en/icons/ibm700series/impacts) ..........4

Figure 1.3. Supervised machine learning .................................................................4

Figure 1.4. Clustering (unsupervised learning)............................................................5

Figure 3.1. Secondary structure of a protein (Retrieved from https://www.mdpi.com/2073-4425/11/12/1431/html) .........................................................16

Figure 5.1. The general architecture of our proposed CNN method .........................25
Chapter 1 Introduction

1.1 PTM

Protein post-translational modification (PTM) is an important regulatory mechanism and plays an important role in normal and pathological cell physiology [1]. PTMs are defined as enzymatic modification of proteins after translational process [2]. PTM occurs in different amino acid side chains or peptide chains, which is mainly mediated by enzyme activity. In fact, it is estimated that 5% of the proteome contains enzymes that undergo more than 200 post-translational modifications [3].

Moreover, post translational modification can occur at any stage of the protein lifespan. For instance, many proteins are modified shortly after translation to guide neonatal proteins to different cellular compartments or to ensure proper folding or stability of the protein [4]. Other changes occur after folding to activate or inactivate catalytic activity and affect the biological activity of proteins [5]. In addition to a single modification, proteins are usually modified by the combination of post-translational cleavage and the addition of functional groups through the stepwise mechanism of protein maturation or activation [6].

Studies have shown that most proteins in cells are controlled by reversible post-translational modification. This process helps cells to respond quickly to the changes of external environment and signal stimulation [5-7]. Therefore, the analysis and understanding of proteins and their post-translational modifications is particularly important to understand the functioning of the proteins and specially for the studies of heart disease, cancer, neurodegenerative diseases and diabetes. With the advance of
above studies, the need for identification of PTMs is becoming dramatically important [8]. Traditional PTM detection methods include radioactive chemical method [3], chromatin immunoprecipitation [4], and mass spectrometric detection [5]. All above methods are experimental approaches, which are time-consuming and costly [7]. Hence, it is urgent to develop computational methods to identify PTM sites.

A wide range of protein post-translational modification types are identified which include phosphorylation, prenylation, ubiquitination, methylation, crotonylation, acetylation [8]. Among these PTMs, the regulation of lysine acetylation is one of the most important ones [10]. Acetylation is the reaction of acetyl functional groups (acetoxy, molecular formula: CH₃CO) into organic compounds (Figure 1.1), that is, acetylation replaces hydrogen atoms [9].

The lysine acetylation widely exists in mammals including humans and mouse [11]. Acetylation regulates many gene expressions, such as localization, stability, and synthesis [12]. More importantly, it is associated with specific human diseases like hypertension, vascular diseases, arrhythmia, heart failure, and angiogenesis [13]. In
addition, the abnormal function of lysine acetyltransferases (KATs) and lysine deacetylases (KDACs) affects cell division [12]. Therefore, the study of acetylation regulation will improve the understanding of cell and proteins and help to reveal the potential biological process of many disease [13].

1.2 Machine Learning

In 2021, machine learning is not an unfamiliar concept. It could be defined as a branch of artificial intelligence (AI) and computer science. The focus of machine learning is to use data and algorithms to simulate human learning and gradually improve its accuracy. The idea of machine learning was introduced by Arthur Samuel (Figure 1.2), an American pioneer in the field of computer gaming and artificial intelligence, in 1959. He stated that “it gives computers the ability to learn without being explicitly programmed” [14]. It has been since a continuous pursuit for every researcher in the field of computer science to let the computer do the learning automatically and solve the encountered problems.

Traditionally, machine learning approaches could be classified into three major categories namely, supervised learning, unsupervised learning, and reinforcement learning [15]. However, in recent years, with generation of huge data as well as the improvement of learning algorithms, there are more and more semi-supervised learning algorithms entering the field as the fourth category.
Supervised learning is a machine learning algorithm that learns patterns from example data with label and map inputs to target responses [16]. It infers function from a set of training data that composed of numeric or numeric values or string labels. Which means in supervised learning, a new input object has an expected output values generated by the functions (Figure 1.3). This method is similar to human learning under the supervision of teachers, where the teacher gives students some good examples, and then the students draw the general rules from these examples.

Unsupervised learning is to develop the learning algorithm based on the structure of input data. It is an algorithm that learns from relevant answers, that is no labels are
provided for training data [17]. The algorithm should self-discover data into types, such as new features that can represent a class or a new set of unrelated values. Clustering is a common example of unsupervised learning (Figure 1.4), which algorithm automatically dividing training samples into different categories based on their characteristics [17]. It is similar to the method used by humans to determine whether certain objects or events belong to the same category.

Reinforcement learning is a machine learning technique, which allows algorithms to learn through reward and punishment in an interactive environment [18]. It involves some concepts of how software works in the environment to obtain the maximum cumulative return (Figure 1.5).
Semi-supervised learning is a machine learning system with incomplete training signals. The training data combines a small amount of labeled data with a large amount of unlabeled data. Therefore, semi-supervised learning falls between supervised learning and unsupervised learning.

Since machine learning can easily identify patterns and no human intervention is needed, it is a more efficient and fast method to identify acetylation sites compared to traditional experimental methods. Therefore, new machine learning methods have been proposed to predict acetylation sites [8, 19].

In this study, we first use traditional machine learning methods based on Random Forest (RF) and Support Vector Machines (SVM) to predict acetylation sites. Then we apply a new deep learning model based on convolutional neural networks (CNN) to predict acetylation sites. The CNN method shows more promising results than the traditional machine learning methods. The CNN method demonstrates accuracy of 72.3% with sensitivity of 24.2%, specificity of 84.3% and Matthew’s correlation coefficient (MCC) of 0.73 on the human samples. It also predicts the acetylation sites for mouse with 71.1%, 23.2%, 86.7%, and 0.68 in terms of accuracy, sensitivity, specificity, and Matthew’s correlation coefficient (MCC), respectively. The CNN method beats the random forests method in terms of accuracy, sensitivity, specificity, and MCC on both human and mouse samples as well. It also outperforms SVM method in terms of accuracy, specificity, and MCC on both human and mouse samples, too.

We then compare our methods with two established methods based on deep neural networks (DNNAce [8] and XGBoost PPISP-XGBoost [19]) to predict lysine acetylation sites. Our model beats the DNNAce in terms of accuracy, sensitivity, and MCC on both
human and mouse samples. It also outperforms PPISP-XGBoost on sensitivity, specificity, and MCC on both human and mouse samples.

The rest of this thesis is organized as follows. In Chapter two, a comprehensive review of the literature and the state of art models is conducted. In this chapter, the development path of acetylation sites prediction methods besides the merits and shortcomings of these methods are presented and reviewed.

In Chapter three, the procedure of how our dataset used and how training and testing sets are built is presented, comprehensively. We also introduce the mixture techniques to extract features from both structural and evolutionary information.

In Chapter four, we present our proposed lysine acetylation prediction method based on traditional machine learning methods (Random Forest and SVM). We present and compare their performance with both structural and evolutionary feature extraction methods.

In Chapter five, we introduce a novel acetylation site prediction method based on CNN and compare its performance with traditional machine learning methods. We also compare our method with two other deep learning acetylation prediction methods proposed in recent two years and show how our method is able to enhance the acetylation prediction.

In Chapter six, we present the conclusions and the potential research direction in the future.
Chapter 2 Literature review

2.1 Overview

As mentioned in chapter one, the traditional experimental methods to identify lysine acetylation sites are time consuming and costly. In recent years, several studies based on machine learning for acetylation detection have been proposed. Among these studies, some focused on feature extraction based on the properties of lysine such as physicochemical properties, sequence information, structural and evolutionary features, and functional annotation. The others mostly focused on the classification methods, such as support vector machine (SVM), Random Forest (RF), and Artificial Neural Network (ANN).

Support vector machine (SVM) is a supervised machine learning method that maps input samples to higher dimension space and uses an optimal hyperplane for classification and regression [22]. In 2010, Gnad et al. [26] first introduced SVM for lysine acetylation prediction and achieved 78% accuracy using 1750 samples. However, no further performance evaluations like sensitivity and specificity were reported. Shi et al. [27] first combined protein sequence information such as secondary structure and amino acid properties and used SVM to predict lysine acetylation sites. The accuracy reached 83% using 1639 protein samples.

More recently, Lee et al. [29] first introduced solvent accessibility and the physicochemical properties of proteins to predict lysine acetylation. With the combination of using SVM as classifier and amino acid sequences, accessible surface area and physicochemical properties of proteins as features, they achieved 5% to 14% higher accuracy compared to the model using only amino acid sequences. In a different
study, Wuyun et al. [28] first focused on human samples and developed an online tool based on SVM called KA-predictor to predict acetylation sites. They reached 68.9% accuracy in Homo sapiens (human) samples and 60.4% in Mus musculus (mouse) samples. Xu et al. [31] first introduced an ensemble of SVM classifiers to predict acetylation sites. Their accuracy reached 87.9% using a dataset of 830 acetylation sites and outperformed the same model with only single SVM classifier. Later on, Wang et al. [32] used SVM with multiple kernels to boost predictive performance, and it reached better accuracy than the existing prediction methods. In recent years, a series of studies have carried out on how to extract more informative features from protein sequences.

Besides traditional structural information such as accessible surface area (ASA), more evolutionary information has been used [23]. Combining structural information with evolutionary information, more comprehensive and effective features can be extracted. Position-specific scoring matrix (PSSM) is a common patterns representation in biological sequences to extract evolutionary information of proteins [37]. It generates substitution score among amino acids with respect to their positions along protein sequence. Murakami et al. [34] first introduced PSSM in prediction of acetylation site in 2010. They reached a recall of 41.6% on dataset consisting of 186 protein samples with MCC of 0.15. Although the results seem not very satisfactory in today’s standard, this research gave us a breakthrough on how to extract more evolutionary features to tackle this problem.

In a different study, Tien et al. [21] first introduces RSA in the prediction of acetylation in 2013. The relative solvent accessibility (RSA) of the residue in protein that measures the residue’s exposure in the 3D structure. Then Dhole et al. [35] combined the
PSSM and RSA in the prediction and reported 66.2% in accuracy and 64.3% in specificity in 2014. Later on, Xie et al. [36] first combined PSSM and PSFM as the evolutionary information to predict the acetylation sites. Position specific frequency matrix (PSFM) is similar to PSSM and is created by counting the occurrences of each nucleotide at each position. They reported 81.1% in terms of AUC.

Besides these studies that used SVM to improve prediction performance, other studies put their focus on how to choose different machine learning classifiers to outperform existing models. Random Forest (RF) is a classifier that fits a number of decision tree classifiers and uses average value to control overfitting and improve the accuracy [33]. Hou et al. [28] combined RF classifier with feature backbone flexibility to predict acetylation sites. This RF predictor performs better than existing predictors on the independent heteromeric dataset.

Convolutional neural network (CNN) is another classifier that was used for this task. CNN uses the weighted architecture of the convolution kernels and provide translation equivariant responses. It is widely used in analyzing visual images [36]. Yu et al. [21] first used CNN as the classifier as well as PSSM as the evolution information of amino acid residues to predict the acetylation sites in 2020. They utilized the advanced deep learning technique and reported 85% in accuracy on prokaryote acetylation sites.

Light Gradient Boosting Machine (LightGBM) is another classifier that is also used for this task. LightGBM is a distributed gradient boosting framework based on decision tree algorithms. It is mainly used for classification and ranking [24]. Zhang et al. [18] used the LightGBM classifier to predict acetylation and reported a promising result of 93%
in terms of accuracy on E. coli species. Their work showed that ensemble classifiers could be a powerful tool for acetylation cite prediction task.

Similarly, Extreme gradient boosting (XGBoost) is another classifier that is recently used to predict acetylation sites. XGBoost is an implementation of gradient boosted decision trees that designed for better speed and performance [25]. Wang et al. [18] introduced a tool based on XGBoost classifiers for this task. They used evolutionary protein information like PSSM and AAC and structural information like solvent accessible surface area for prediction task. They reported 85% in accuracy on 10-fold cross validation for their employed dataset.

It is worth noting that despite all efforts that have achieved many extinguished results and set good examples for future studies in recent years, there are still several deficiencies in prediction of lysine acetylation sites. First, many studies mainly focused on how to choose the classifiers for acetylation sites prediction and therefore underestimated the importance of features that represent the sequence information. For these studies, despite using complex classifiers such as recurrent neural network (RNN) methods and long short-term memory (LSTM) [33], their results were not satisfactory as they just used features based on a single aspect of physicochemical properties.

Second, several other studies mainly focused on the traditional machine learning models such as SVM and RF, focusing on enriching sequence information. Their results are also remained limited. Finally, some studies just use prokaryote datasets and no mammal species data were used. In this study, we propose a prediction model that addresses these three shortcomings. First, we apply deep learning methods (CNN) as well as traditional machine learning classifiers (RF, SVM) for lysine acetylation site
prediction. Second, we extract structural features such as secondary structure, accessible surface area and backbone torsion angles. We also extract evolutionary features from PSSM for our prediction model. Finally, our dataset is focused on *H. sapiens* (human) and *M. musculus* (mouse) to conduct our experiments.
Chapter 3 Dataset/Features

3.1 Dataset

To build our dataset for lysine acetylation site prediction, we first extracted data from integrative resource of protein lysine modification database (PLMD) [49]. In PLMD, there are various kind of PTMs including ubiquitination, acetylation, sumoylation, and methylation. Also, there are samples from different species including H. sapiens, M. musculus, E. coli and many more.

This acetylation dataset consists of 2241 H. sapiens (human) and 1766 M. musculus (mouse) proteins. To reduce the bias, we first removed proteins with over 40% sequence similarity threshold using CD-HIT to discard redundancy from our dataset. The remaining dataset consist of 1741 human proteins with 4561 acetylation sites and 72,145 non-acetylation sites, and 1267 mouse proteins with 3454 and 43,943 acetylation and non-acetylation sites, respectively.

Due to the enormous difference of the number of acetylation sites and non-acetylation sites, the prediction could be bias because of this imbalance. Although there are more non-acetylation sites exist in real biological world, we should balance the data to avoid biased towards non-acetylation sites as it is the main purpose of this task [37]. There exist many methods to balance the positive and negative data, such as SMOTe [38], ADASYN [39] and K-Nearest Neighbor (KNN) [40]. Among these methods, either the minor class could be oversampled in SMOTe and ADASYN, or the major class could be down sampled. On the one hand, it is nice to have more samples to training and testing in machine learning.
On the other hand, oversampling will introduce some artificial data to the datasets and bias the prediction. After considering the pros and cons of both methods, we used KNN to down sample the dataset so that every data is real (no artificial sample is produced). To balance the positive and negative acetylation sites, the Euclidean distance was calculated for each pair of lysine residues. Then, those negative lysine sites were removed from dataset that had one positive points in its k-nearest neighbors. Some studies [20, 29] have shown that 1:1 ratio of positive and negative sites could have best performance, and therefore our aim is to balance the dataset to this ratio.

After these procedures, our human dataset consists of 2490 positive and 2493 negative sites, and our mouse dataset consists of 1455 positive and 1450 negative sites.

To investigate the performance of our model, an independent test dataset was separated from the original dataset. The ratio is critical as larger training dataset gives more data for training while it may cause overfitting and larger testing dataset reduces the training dataset and may cause the insufficient data for training. Therefore, different training: testing ratio range from 5:1, 3:1 and 1:1 was investigated in this study. The detail of our employed datasets is presented in Table 3.1. The number of acetylation sites used for different training: testing ratio is presented in Table 1. Later in Chapter 4, we will show that among these ratios, using 5:1 we obtain the best performance.

Table 3.1 Numbers of sites for human and mouse

<table>
<thead>
<tr>
<th>Training</th>
<th></th>
<th></th>
<th></th>
<th>Testing</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>Mouse</td>
<td>Human</td>
<td>Mouse</td>
<td>Human</td>
<td>Mouse</td>
<td>Human</td>
<td>Mouse</td>
</tr>
<tr>
<td>Positive</td>
<td>Negative</td>
<td>Positive</td>
<td>Negative</td>
<td>Positive</td>
<td>Negative</td>
<td>Positive</td>
<td>Negative</td>
</tr>
<tr>
<td>5:1</td>
<td>2075</td>
<td>2077</td>
<td>1212</td>
<td>1208</td>
<td>415</td>
<td>415</td>
<td>242</td>
</tr>
<tr>
<td>3:1</td>
<td>1867</td>
<td>1870</td>
<td>1091</td>
<td>1088</td>
<td>622</td>
<td>623</td>
<td>364</td>
</tr>
<tr>
<td>1:1</td>
<td>1245</td>
<td>1246</td>
<td>728</td>
<td>725</td>
<td>1245</td>
<td>1247</td>
<td>727</td>
</tr>
</tbody>
</table>
3.2 Feature extraction

In this study, we use structural and evolutionary features to predict lysine acetylation sites. The following sections present how we extract these features and the mixture technique used for prediction task.

3.2.1 Structural features

The structural features used in this study are the Accessible Surface Area (ASA), the predicted secondary structure (coil, strand and helix) and the backbone torsion angles (θ, τ, ψ, and φ). Previous studies [20–23] show that these features contain important structural discriminatory information for the PTM classification purpose. In this study, we used SPIDER2 as the toolbox to extract the predicted value of above parameters. SPIDER2 is a machine learning toolbox to predict local structure of protein based on deep learning and it outperforms other similar tools [38-40]. A matrix consisted of all amino acid’s predicted values is produced as the results of SPIDER2.

Accessible surface area (ASA) can be defined as the accessible area to a solvent of an amino acid [41]. Since the predicted value of ASA differs in various protein configuration and proteins interacted with each other differently, we could get essential information about the structure of protein. The obtained ASA value is computed for every protein sequence.

Secondary structure defines local structures of proteins and delivers significant information [42]. The three local components, namely helix (ph), strand (pe), and coil (pc) build the 3D structure of proteins (Figure 3.1). These three local structures help to determine which amino acid are more likely to interact with each other. We executed
SPIDER2 to produce a $L \times 3$ matrix for secondary structure prediction where $L$ represents the length of protein sequence.

Local backbone angles give torsion angles between neighboring amino acids and also show the local structure of protein [43]. The four angles are backbone torsion angle $\psi$ and $\phi$ represent the angle along the protein backbone and dihedral angles $\theta$ and $\tau$ represents the rotation angle [44]. Executing SPIDER2 on protein sequence will give the $L \times 4$ matrix for torsion angle prediction where $L$ represents the length of protein sequence.

![Figure 3.1. Secondary structure of a protein](image)

To combine all these structural information parameters, a structural matrix ($L \times 8$) is constructed. Where $L$ is the length of protein and 8 represents all of eight parameters. Recent studies have shown 7-residue window size around an acetylation site for feature extraction obtain the best results [42,44]. Therefore, in this study, 3 upstream, 3 downstream amino acids and lysine residue ($K$) in the middle are used to obtain structural matrix (7-residue in total). The peptide S could be represented as:

\[ S = \{A_{-3}, A_{-2}, A_{-1}, K, A_1, A_2, A_3\} \]
3.2.2 Evolutionary features

Evolutionary features are extracted using Position-Specific Scoring Matrix (PSSM) [45]. PSSM is built using PSI-BLAST (Position-Specific Iterated BLAST). PSI-BLAST is a widely used tool for detection of related but evolutionarily distant sequence of protein [50]. The achieved PSSM is a $L \times 20$ matrix where $L$ represents the length of protein sequence and 20 represents different amino acids. To generate features for a given acetylation site, we investigated different window sizes. Among them using the 31 windows size we obtained the best for acetylation site prediction which is similar to those reported in [42-44]. Therefore, 15 upstream and 15 downstream (and a central lysine) amino acids are used to extract features from PSSM (31 residues in total). The peptide $S$ could be represented as:

$$S = \{A_{-15}, A_{-14}, \ldots, A_{-1}, K, A_1, \ldots, A_{14}, A_{15}\}$$

For both sides of the protein tail, where there are less than 3 (structural) or 15 (evolutionary) neighboring amino acids, we use left mirroring and right mirroring to fill out the missing amino acids [48].

Bigram is a sequence of two adjacent elements from a string of tokens and is widely used as a frequency vector. Research has shown that using bigram transformation from PSSM has promising results in protein analysis problems than using PSSM matrix [45, 46, 51]. The original dimension of PSSM is $31 \times 20$, then we convert PSSM matrix ($M$) to a frequency vector of bigrams and generate a $20 \times 20$ matrix ($B$). It was shown
in [45, 46] that using bigram we can extract effective features from PSSM [45, 46]. The bigram could be calculated as:

$$B_{p,q} = \sum_{k=1}^{30} m_{k,p} m_{k+1,q}$$

where $1 \leq p \leq 20$ and $1 \leq q \leq 20$

Since the structural matrix’s dimension is $7 \times 8$, and the converted evolutionary matrix is $20 \times 20$, we generate a 456-dimensional vector ($56 + 400$) and this vector captures both structural and evolutionary information and will be used for further acetylation site prediction.
Chapter 4 Traditional machine learning

4.1 Introduction

In this Chapter, we aim at exploring the prediction result of acetylation sites based on the traditional machine learning methods such as random forest (RF) and support vector machine (SVM). We then compare these two methods and summarize their advantages and disadvantages, separately. In this Chapter, we also explore what is the best ratio for training and testing and how feature extraction affects our model results.

With more than six decade’s developments since it was coined in 1959, machine learning has advanced, dramatically. As in this study, we aim at using supervised learning, as the machine learning algorithm learns patterns from example data with labels [16]. There are many methods of supervised learning, such as Naïve Bayes, decision tree, random forest, KNN, and SVM [18].

The concept of Random Forest (RF) was first proposed in 1995 and was developed in 2006 as a tree-based ensemble method [55]. It is defined as forest because $K$ decision trees that are sensitive to rotation of the feature axes are chosen and the feature set is randomly split into $K$. These $K$ decision trees transform features separately and the algorithm aggregates all the outcomes and produce the final result. The concept behind RF is to keep individual accuracy of each decision tree and diversity of the group at the same time. Therefore, RF is typically more accurate than single decision tree method [56].

Support Vector Machine (SVM) is a supervised learning model that mainly used for classification and regression analysis [56]. Given a set of training dataset, SVM maximize width among two hyperplanes representing linear boundaries of two different
sets. Comparing to other common classifiers, prediction model based on SVM has lower prediction error when there exists large number of features [57]. As mentioned in previous Chapter, we have 456 features available for this study. SVM was proved very sufficient and widely used in different research areas in bioinformatics and many state-of-the-art predictors for acetylation sites have used it [22-28]. There have been studies based on SVM in chemogenomic, genomics, protein function prediction, protease functional site recognition and gene expression data classification [52-54].

4.2 Results

4.2.1 Performance Metrics

In this study, four metrics have been selected to compare the performance of prediction with different classifiers. These four metrics are accuracy (Acc), sensitivity (Sn), specificity (Sp), and Matthew’s correlation coefficient (MCC). Accuracy presents the total number of correctly classified acetylation and non-acetylation sites. Sensitivity presents the number of correctly identified acetylation sites. Specificity presents the number of correctly non-identified acetylation sites. MCC indicates the quality of the predictor. Acc, Sn and Sp ranges from 0 to 1 and the higher the score, the better the predictor is. MCC ranges from -1 to 1 and -1 indicates a totally negative correlation and +1 means a highly positive relationship. These metrics are calculated as follows:

\[
Acc = \frac{TP + TN}{TP + TN + FP + FN}
\]

\[
Sn = \frac{TP}{TP + FN}
\]
where TP (True Positive) is the number of correctly classified acetylation sites. TN (True Negative) represents the number of correctly classified non-acetylation sites. FP (False Positive) is the number of misclassified non-acetylation sites as acetylation sites. FN (False Negative) represents the number of misclassified acetylation sites as non-acetylation sites.

4.2.2 Comparison of RF and SVM

The results of using SVM and RF classifiers for Human and Mouse datasets are presented in Table 4.1. As shown in this table, SVM classifier is performing better than RF in all metrics in human samples and almost all metrics except accuracy in mouse samples. SVM method achieves up to 5%, 25%, 10% and 4% improvement over RF method for Acc, Sn, Sp, and MCC for human samples and achieved up to 14%, 3% and 2% improvement over RF method for Sn, Sp, and MCC for mouse samples, respectively. And RF shows less than 1% improvement over SVM classifier in term of accuracy.

<table>
<thead>
<tr>
<th>Method</th>
<th>Acc</th>
<th>Sn</th>
<th>Sp</th>
<th>MCC</th>
<th>Acc</th>
<th>Sn</th>
<th>Sp</th>
<th>MCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>RF</td>
<td>68.7%</td>
<td>22.3%</td>
<td>81.1%</td>
<td>0.69</td>
<td>70.7%</td>
<td>21.1%</td>
<td>83.1%</td>
<td>0.62</td>
</tr>
<tr>
<td>SVM</td>
<td>71.9%</td>
<td>27.8%</td>
<td>81.9%</td>
<td>0.71</td>
<td>70.1%</td>
<td>24.1%</td>
<td>84.6%</td>
<td>0.63</td>
</tr>
</tbody>
</table>
4.2.3 Comparison of different ratio of training and testing data

In Table 4.2, we present the results of SVM classifier with different training and testing ratio. As shown in this table, using the 5:1 ratio, the best performance for SVM is achieved in almost all metrics except specificity in both human and mouse samples (80% training and 20% testing samples). This result is consistent with our prior hypothesis that it is likely that our model would train better and obtain better results with more training data.

<table>
<thead>
<tr>
<th>Method</th>
<th>Human Acc</th>
<th>Human Sn</th>
<th>Mouse Acc</th>
<th>Mouse Sn</th>
</tr>
</thead>
<tbody>
<tr>
<td>5:1</td>
<td>71.9%</td>
<td>27.8%</td>
<td>5:1</td>
<td>71.9%</td>
</tr>
<tr>
<td>3:1</td>
<td>71.7%</td>
<td>24.2%</td>
<td>3:1</td>
<td>71.7%</td>
</tr>
<tr>
<td>1:1</td>
<td>71.2%</td>
<td>21.3%</td>
<td>1:1</td>
<td>71.2%</td>
</tr>
</tbody>
</table>

4.2.4 Comparison of different feature extractions

Here we investigate the impact of our employed features. We use SVM with features extracted from SPIDER 2.0, PSMM, and then compare it to the results when using both feature groups, together. The results of this comparison are presented in Table 4.3. As shown in this table, combination of evolutionary and structural information performing better than using evolutionary information or structural information separately in terms of sensitivity for both human and mouse samples. Note that sensitivity is the most important metric as high sensitivity means the predictor could effectively detect Acetylation sites. Hence, having higher Sensitivity is the desired outcome.
Table 4.3 Performance of different feature extractions based on SVM classifier

<table>
<thead>
<tr>
<th>Method</th>
<th>Human</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>Mouse</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Acc</td>
<td>Sn</td>
<td>Sp</td>
<td>MCC</td>
<td>Acc</td>
<td>Sn</td>
<td>Sp</td>
<td>MCC</td>
</tr>
<tr>
<td>PSSM+SPIDER</td>
<td>71.9%</td>
<td>27.8%</td>
<td>81.9%</td>
<td>0.71</td>
<td>70.1%</td>
<td>24.1%</td>
<td>84.6%</td>
<td>0.63</td>
</tr>
<tr>
<td>PSSM</td>
<td>77.0%</td>
<td>9.8%</td>
<td>98.7%</td>
<td>0.35</td>
<td>74.0%</td>
<td>8.3%</td>
<td>93.2%</td>
<td>0.31</td>
</tr>
<tr>
<td>SPIDER</td>
<td>68.0%</td>
<td>9.9%</td>
<td>94.5%</td>
<td>0.34</td>
<td>76.0%</td>
<td>7.8%</td>
<td>94.5%</td>
<td>0.32</td>
</tr>
</tbody>
</table>

4.3 Conclusion

In this Chapter, we explored the performance of lysine acetylation sites predictors based on SVM and RF. We also explored how the ratio of training and testing dataset affects the performance of predictor. Finally, we investigated the impact of our employed features.

Our reported results showed that by using SVM classifier, we outperformed the acetylation site prediction compared to RF classifiers in all aspects in human samples and in almost all aspects in mouse samples. Also, by comparing different training and testing data, the 5:1 training and testing ratio (80% training and 20% testing samples) represents best compared to other ratios and will be used for the rest of this study. Finally, we showed that the combination of evolutionary and structural features demonstrates better results than using each one, separately.

Despite achieving promising results, our performances are by far lower than the highest reported results for lysine acetylation site prediction in the literature [8, 19]. The state-of-the-art predictor has shifted the focus to deep learning methods. In next Chapter, we aim to exploring the lysine acetylation site prediction based on deep learning classifier.
Chapter 5 Deep learning

5.1 Introduction

In this Chapter, we aim at exploring the prediction result of acetylation sites based on the deep learning methods such as Convolutional Neural Network (CNN). We then compare this method with traditional machine learning model presented in last chapter (RF, SVM) and summarize their advantages and disadvantages, separately. In this Chapter, we also compare our model with state-of-the-art model proposed in recent year.

With the multi-layer architectures of neural network first proposed in 1991 by Kurt Hornik [60], deep learning has experienced its rapid development and was widely spread. Deep learning is widely used in different field including computer vision, natural language processing, and bioinformatics [61]. Many deep learning methods have been introduced under this high demand situation, such as deep neural networks, deep reinforcement learning, recurrent neural networks and convolutional neural networks [62].

The concept of convolutional neural networks was inspired by biological processes [60]. The basic structure of the neural network consists of one input layer, several hidden layers and one output layers. The units in each layer, except for input layer, are connected to units in the previous layer. It is very similar to connectivity pattern between the neurons in visual cortex [58]. Compared to other deep learning methods, CNNs take a different approach towards regularization. Smaller and simpler patterns in the hidden layers are used because of the advantage of hierarchical pattern. Therefore, CNN has lower scale and complexity compared to other classification algorithms [59].
The following parameters are crucial to the CNN model: the number of hidden layers, the number of kernel size and filters in hidden layers, the activation function used and the cross-entropy loss function optimizer.

In this study, we set 4 hidden layers as it balanced the training time and results. We set filters to 32 and kernel size to $3 \times 3$ as experimental results showed it is the optimal choice [18]. We set the activation function to Relu and optimizer to Adam [59].

In this section, we present our results and compare them with previous studies found in the literature. The general architecture of our proposed CNN model is presented in Figure 5.1.

![Figure 5.1. The general architecture of our proposed CNN method](image)

5.2 Results

5.2.1 Comparison of CNN and traditional machine learning classifier

The results of using CNN compared to traditional machine learning classifiers (RF, SVM) for Human and Mouse datasets are presented in Table 5.1. As shown in this table, CNN classifier is performing better than RF in all metrics in both human and mouse
samples. The CNN method achieves up to 5%, 9%, 4%, and 0.04 improvement over RF method for Acc, Sn, Sp, and MCC, respectively, for human samples and achieved up to 1%, 10%, 5%, and 0.06 improvement over RF method for Acc, Sn, Sp, and MCC, respectively, for mouse samples, respectively. CNN classifier also performs better than SVM in all metrics excepts sensitivity in both human and mouse samples. The CNN method achieves up to 1%, 3%, and 0.02 improvement over SVM method for Acc, Sp, and MCC for human samples and achieves up to 2%, 3%, and 0.05 improvement over SVM method for Acc, Sp, and MCC for mouse samples, respectively. SVM shows less than 3% improvement over CNN classifier in term of sensitivity.

<table>
<thead>
<tr>
<th>Method</th>
<th>Human</th>
<th>Mouse</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Acc</td>
<td>Sn</td>
</tr>
<tr>
<td>RF</td>
<td>68.7%</td>
<td>22.3%</td>
</tr>
<tr>
<td>SVM</td>
<td>71.9%</td>
<td>27.8%</td>
</tr>
<tr>
<td>CNN</td>
<td>72.3%</td>
<td>24.2%</td>
</tr>
</tbody>
</table>

5.2.2 Comparison of our CNN model with state-of-the-art models

In Table 5.2, we present the results of CNN classifier compared to two state-of-the-art models [8,19] named DNNAce and PPISP-XGBoost separately. As shown in this table, CNN classifier is performing better than DNNAce and PPISP-XGBoost in term of sensitivity and MCC in both human and mouse samples. The CNN method achieves up to 20% and 0.12 improvement over DNNAce for Sn and MCC for human samples and achieves up to 17% and 0.06 improvement over DNNAce method for Sn and MCC for mouse samples, respectively. On the other hand, DNNAce shows less than 3% and 6% improvement over CNN classifier in term of accuracy and specificity. The CNN method achieves up to 3% and 0.03 improvement over PPISP-XGBoost for Sn and MCC for
human samples and achieves up to 8% and 0.04 improvement over PPISP-XGBoost method for Sn and MCC for mouse samples, respectively. On the other hand, PPISP-XGBoost shows less than 3% and 2% improvement over CNN classifier in term of accuracy and specificity in both human and mouse samples.

<table>
<thead>
<tr>
<th>Method</th>
<th>Human</th>
<th>Mouse</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Acc</td>
<td>Sn</td>
</tr>
<tr>
<td>CNN</td>
<td>72.3%</td>
<td>24.2%</td>
</tr>
<tr>
<td>DNNAce</td>
<td>74.5%</td>
<td>19.4%</td>
</tr>
<tr>
<td>PPISP-XGBoost</td>
<td>74.1%</td>
<td>23.5%</td>
</tr>
</tbody>
</table>

Our results demonstrate the ability of CNN and the impact of our proposed features to enhance lysine acetylation site prediction compared to previous studies found in the literature.

5.3 Conclusion

In this Chapter, we used CNN to predict lysine acetylation sites. We also compared the performance of our model to the state-of-the-art models proposed in recent year. Our reported results showed the using CNN, we outperformed the acetylation site prediction compared to RF classifiers in all aspects in both human and mouse samples. The CNN model also demonstrates better performance than SVM classifier in almost all aspects in both human and mouse samples. Also, our proposed CNN model compared to two different state-of-the-art models demonstrates better performance in terms of sensitivity and MCC. As it was discussed in Chapter 4, high sensitivity means the predictor can effectively detect Acetylation sites. Hence, having higher sensitivity is the desired
outcome we want. We propose a new acetylation site prediction model based on CNN that could detect acetylation site 19% more effectively than DNNAce and 5% more effectively than PPISP-XGBoost.

Despite achieving promising results, the performance of CNN still has some deficiency such as 5% lower accuracy and specificity. It is because 9 different datasets are used in DNNAce and 5 different datasets are used in PPISP-XGBoost. Therefore, their model is trained based on more data. Our future studies aim at exploring how the number and size of datasets influence the performance of acetylation site prediction.
Chapter 6 Conclusion

This Chapter summarizes the progression and contribution of this thesis. In Section 6.1 we present the contributions to the research of lysine acetylation site prediction. The contribution presented covers the aims and objectives set before the thesis, problems encountered during the experiments and achievements gained after the research. We also review our research results and outline some future directions in Section 6.2.

6.1 Contributions

In this study, we tried to enhance the lysine acetylation prediction performance using a new machine learning approach. To achieve this goal, we looked through the literature published in recent years to investigate their lysine acetylation prediction models and identified their shortcoming as follows. First, despite all the efforts have been made so far, many studies mainly focused on how to choose the classifiers for acetylation sites prediction and therefore underestimated the importance of features that represent the sequence information. In most cases, despite using complex classifiers such as recurrent neural network (RNN) methods and long short-term memory (LSTM) [33], their results were not satisfactory as they just used features based on a single aspect of physicochemical properties.

Therefore, some other studies have changed their focus to using more features extracted from structural and evolutionary information [34, 35]. At the same time, their studies mainly focused on the traditional machine learning models such as SVM and RF. Their results are unable to provide effective lysine acetylation prediction.

In this study, we built a more effective lysine acetylation sites predictor in the following steps.
First, we used two different traditional machine learning classifiers (RF and SVM) and use both structural and evolutionary feature extractions to investigate the effectiveness of the traditional machine learning classifiers on the prediction of lysine acetylation sites. Using these two classifiers showed (in Chapter 4) that SVM has better performance than RF in almost all aspects in both human and mouse samples. We then discovered related attributes on the prediction such as the training and testing ratio and the impact of our employed features. We identified that the 5:1 training and testing ratio demonstrate the best performance compared to other ratios and is used for the rest of this study (80% training and 20% testing samples). We also showed that the combination of evolutionary and structural features demonstrates better results than using each one, separately.

Second, in Chapter 5, we explored the novel deep learning techniques (CNN) and apply it on the lysine acetylation sites prediction. By comparing the results to two traditional machine learning mentioned in Chapter 4, we found out that the CNN could have 2% and 3% improvement in accuracy in predicting both on human and mouse samples. We then compared our CNN model to two state-of-the-art models proposed in recent years and achieved 20% and 5% improvements in terms of sensitivity for human and mouse dataset, respectively. However, we still have not been able to enhance the prediction accuracy better than these two models.

6.2 Future Works

There exists a number of possible future directions of lysine acetylation site prediction based on the research conducted in this thesis. Our future works can be summarized into two parts as follows:
- **Investigating the relationship between accuracy and sensitivity:** As it was shown in Chapter 5, we have successfully explored the prediction model based on the deep learning and we have achieved better sensitivity in the predicting lysine acetylation sites. However, we were not able to perform the state-of-the-art models in term of accuracy. Sensitivity is the most important metric as high sensitivity means the predictor can effectively detect Acetylation sites. Our future work will be to find the relationship of the accuracy and sensitivity so that our future model could outperform these models in all metrics.

- **Extending our experiments for more datasets and developing our online lysine acetylation sites predictor:** After exploring a wider range of literatures as well as their models, we find out some models with good performance has been trained on several different datasets instead of only one. In future works, we aim at extending our model training to more datasets and try to improve the performance. We will also develop our online lysine acetylation sites predictor and make it publicly available.
Reference


[40] A. Easin, A. Dehzangi, Accurately predicting glutarylation sites using sequential bi-peptide-based evolutionary features, Genes, (2020)9, 1-16


