Heaven's Light is Our Guide

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

Rajshahi University of Engineering & Technology, Bangladesh

iML-LysPTM: Identification of Multi-Label Lysine PTM Sites with Feature Optimization and Data Imbalance Minimization

Authors

Afrida Rahman
Roll No. 1503009
Department of Computer Science and Engineering
Rajshahi University of Engineering and Technology

Sabit Ahmed
Roll No. 1503056
Department of Computer Science and Engineering
Rajshahi University of Engineering and Technology

Supervised by

Dr. Md. Nazrul Islam Mondal
Professor
Department of Computer Science and Engineering
Rajshahi University of Engineering and Technology

ACKNOWLEDGEMENT

At first we would like to thank Almighty Allah for giving us the opportunity to complete our thesis work successfully.

We would like to express our gratitude to our supervisor Dr. Md. Nazrul Islam Mondal, Professor, Department of Computer Science and Engineering, Rajshahi University of Engineering and Technology, Rajshahi. He has provided us the proper guidelines and directions regarding our thesis work. In addition to that, we appreciate Dr. Md. Al Mehedi Hasan, Professor, Department of Computer Science and Engineering, Rajshahi University of Engineering and Technology, Rajshahi for providing us his continuous support.

We are also grateful to all the respective teachers of Department of Computer Science and Engineering, Rajshahi University of Engineering and Technology, Rajshahi for valuable suggestions and inspirations from time to time.

Finally, we would like to thank our parents, friends and well-wishers for aiding helps and appreciations for our thesis work.

Date: 06/02/2021
RUET, Rajshahi

Afrida Rahman
Sabit Ahmed

Heaven's Light is Our Guide

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

Rajshahi University of Engineering & Technology, Bangladesh

CERTIFICATE

This is to certify that this thesis report entitled ***"iML-LysPTM: Identification of Multi-Label Lysine PTM Sites with Feature Optimization and Data Imbalance Minimization"*** *submitted by* ***Afrida Rahman, Roll:1503009, and Sabit Ahmed, Roll:1503056*** *in partial fulfillment of the requirement for the award of the degree of Bachelor of Science in Computer Science & Engineering of Rajshahi University of Engineering & Technology, Bangladesh is a record of the candidate own work carried out by him under my supervision. This thesis has not been submitted for the award of any other degree*

Supervisor

Dr. Md. Nazrul Islam Mondal
Professor
Department of Computer Science &Engineering
Rajshahi University of Engineering &Technology
Rajshahi-6204

External Examiner

Md. Asifur Rahman
Lecturer
Department of Computer Science & Engineering
Rajshahi University of Engineering & Technology
Rajshahi-6204

ABSTRACT

Identification of post-translational modifications (PTM) is significant in the study of computational proteomics, cell biology, pathogenesis, and drug development due to its role in many bio-molecular mechanisms. Computational methods for predicting multiple PTM at the same lysine residues, often referred to as K-PTM, are still evolving. This paper presents a novel computational tool, abbreviated as iML-LysPTM, for predicting K-PTM, such as acetylation, crotonylation, methylation, succinylation, glutarylation from an uncategorized peptide sample involving single, multiple, or no modification. For informative feature representation, multiple sequence encoding schemes, such as the sequence-coupling, binary encoding, amino acid factor have been used with ANOVA and incremental feature selection. As a core predictor, a cost-sensitive SVM classifier has been adopted which effectively mitigates the effect of class-label imbalance in the dataset. iML-LysPTM predicts multi-label PTM sites with 91.73% accuracy using the top 600 features. It has also achieved 91.89\% aiming and 91.78% coverage rates which are much better than the existing state-of-the-art predictors on the same rigorous validation test. This performance indicates that iML-LysPTM can be used as a supportive tool for further K-PTM study. For the convenience of the experimental scientists, iML-LysPTM has been deployed as a user-friendly web-server at http://103.99.176.239/iML-LysPTM.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction

Introduction

Motivation

Challenges and Contributions

Outlines

Conclusion

1.1 Introduction

Post-translational modifications (PTM) referred to the covalent addition of certain functional groups to a protein after the translation process [1]. These modifications have significant effects on cellular processes and proteomic analysis, such as cellular signal transduction, subcellular localization, protein folding, protein degradation, and are also responsible for various kinds of diseases [2], [3]. Therefore, identifying and understanding PTM sites is critical for the basic research in disease detection, prevention, and various drug developments [4], [5]. Table 1.1 summarizes the features of a few important PTMs.

Table 1.1. Some common and crucial post-translational modifications [6],[7].

PTM type	Function and notes
Phosphorylation pTyr pSer, pThr	Reversible, activation/inactivation of enzyme activity, modulation of molecular interactions, signaling.
Acetylation	Protein stability, protection of N terminus. Regulation of protein–DNA interactions (histones).
Methylation	Regulation of gene expression.
Acylation, fatty acid modification	Cellular localization and targeting signals, membrane, tethering, mediator of protein–protein interactions .
Crotonylation	Enriched at active gene promoters and potential enhancers in mammalian cell genomes .
Glycosylation N-Linked O-linked	Excreted proteins, cell–cell recognition/signaling O-GlcNAc, reversible, regulatory functions .
Ubiquitination	Destruction signal. After tryptic digestion, ubiquitination site is modified with the Gly-Gly dipeptide .
Glutarylation	Impacts metabolic processes and other mitochondrial functions .

1.2 Motivation

There are 20 amino acid residues, such as alanine (A), cysteine (C), lysine (K), arginine (R), etc. Modifications that occur at lysine (K) are named lysine modification or K-PTM. Single or multiple lysine residues may be modified individually or simultaneously where one residue can influence others. In other words, these covalent modifications can aid different K-PTM types, including acetylation, crotonylation, ubiquitination, methylation, butyrylation, succinylation, biotinylation, glutarylation, and ubiquitin-like modifications [6], [7], [8].

Although high throughput experimental procedures to characterize these modification sites are known to achieve higher accuracy, computational methods are getting popularity as an effective alternative because of their laborsaving, time, and cost-efficient characteristics. There are several computational tools for predicting various K-PTMs separately [8], [9], [10], [11], [12]. On the contrary, only two recognized multi-label prediction systems have been developed so far that can take care of the multiplex Lys residues according to the best of the authors' knowledge. One is iPTM-mLys, and the other one is mLysPTMpred, which can predict multiple lysine modification sites as well as their different types [5], [13]. Both predictors have utilized the vectorized sequence-coupled model as a feature extraction method [5], [13], [14], [15]. The former has employed the random forest algorithm while the last one has used the support vector machine for classification [13], [16]. There is one more multi-PTM predictor named as predML-Site which has been constructed by us, awaits the peer-reviewing process. A detailed discussion on these three predictors with corresponding methodologies will be disussed in chapter 2. All the predictors built previously, needed pronounced elevation in terms of the prediction quality. Therefore, a tool with higher efficacy is required to meet the current demand in the study of post-translational modifications.

1.3 Challenges and Contributions

For developing a successful predictor for PTM sites, one of the main challenges is to build a non-redundant benchmark dataset with a collection of experimentally verified protein sequences and elicit features from these sequences, as the appropriate

features can play a crucial role in better prediction performance [13]. The next challenge is to handle the data imbalancement as imbalanced data can lead to a poor prediction outcomes. In multi-label prediction system, performance measurement is another crucial challenge. We proposed a new system iML-LysPTM by addressing these challenges succesfully.

In this study, a new dataset has been constructed which includes five K-PTMs, such as acetylation, crotonylation, methylation, succinylation and glutarylation considering several feature encoding methods to propose a novel multi-label predictor iML-LysPTM, where the vectorized sequence-coupled model [5], [14], [15], encoded binary features, and amino acid factors [11], [17] are aggregated to encode a peptide segment. The next challenge is to reduce the dimensionality as a complex system may require a huge amount of space and computational time. Considering this problem, the analysis of variance (ANOVA) F test statistic along with the incremental feature selection approach was used to eliminate the redundant and trivial features [18], [19], [20]. Furthermore, the support vector machine classifier with the variable cost adjustment process [13] was implemented to handle the imbalance in the dataset [21]. A 5-fold cross-validation [13] scheme was repeated five times for validating the statistical significance of the prediction results, and the average performance of each metric has been reported. For a proper evaluation of our system, we have considered five widely-used multi-label metrics. We will discuss each of the aforementioned steps with the detailed procedures in our upcoming chapters.

1.4 Outlines

Chapter 2 - Post-Translational Modifications

This chapter covers the introductory information about gene expression, DNA, RNA, transcription and translation process, and protein's post-translational modifications.

Chapter 3 - Data Acquisition and Feature Optimization

This chapter introduces the data collection process, peptide sample formulation, and expression of five benchmark datasets with various state-of-art feature extraction

techniques, and the ANOVA F-Test feature selection procedure which was implemented in this study.

Chapter 4 - Methodology

This chapter describes the classification and data imbalance handling techniques, k-fold cross-validation scheme, independent test experiment, and the multi-label evaluation metrics.

Chapter 5 - Experimental Analysis

This chapter covers the performance obtained by the proposed system, performance comparison of different features, performance comparison with the existing predictors, performance on the independent test set, and web-server guide in a detailed manner.

Chapter 6 - Conclusion and Future Scopes

This chapter concludes with the findings and future scopes of our research.

1.5 Conclusion

In this chapter, we have introduced post-translational modifications of proteins and their significance in various biological processes. We have also described the necessity of identifying different individual and simultaneous modifications. Furthermore, we have discussed our contributions, and research outlines to conclude the chapter.

CHAPTER 2

Post-Translational Modification and Related Works

Introduction

Concepts of PTM

Process of Modification

Effects of PTMs

Responsible Residue of PTMs

Related Studies in PTM Prediction

Our Contributions in PTM Prediction

Conclusion

2.1 Introduction

Post-translational modification (PTM) is a crucial step in protein biosynthesis, leading to drastic changes in protein structure when incorporating, folding, or eliminating functional groups. Protein expression and gene expression eventually govern the cell activity for the synthesis of a functional product. There is a significant increase in biological complexity from the nucleus to proteome and the contributions of PTM to this complexity have attracted attention in the post-genomic era [21]. This chapter interprets the core concepts with modification procedure, various types of PTMs along with responsible residues, and finally the consequences due to this mechanism.

2.2 Concepts of PTM

Post-translational modifications (PTMs) significantly improve the functional richness of the proteome by covalently attaching groups of proteins, proteolytic cleavage of perforations, or destruction of whole proteins [22]. These modifications can happen at any step of protein lifespan. After translation, several proteins are altered immediately to mediate proper folding or to guide the resultant protein to different molecular locations. And after the completion of folding, other changes occur and localization to activate or inactivate catalytic activity is achieved. PTMs exist in various side chains or peptide bonds of amino acids which are most commonly regulated by enzymatic action [23].

These modifications can be,

a) **Enzymatic:** The involvement of particular enzymes is needed and enzymes play an important role in the alterations. In this case, some functional groups, co-factors are attached to the enzyme based on different criteria. Such as,

I. **Hydrophobic groups for membrane localization:** To control the membrane positioning, some groups are attached like myristate,

palmitate, isoprenoid, farnesol for myristoylation, palmitoylation, prenylation, farnesylation sites respectively.

II. **Co-factors for enhanced enzymatic activity:** To improve the reactivity in biosynthesis, some co-factors like lipoate, FMN are covalently attached for lipoylation, flavin moiety respectively.

III. **Modifications of translation factors:** After translation, some smaller chemical groups are added based on the corresponding modifications like acetylation, alkylation, amidation, butyrylation, glycosylation,gluterylation, formylation, malonylation, succinylation, phosphoglycerylation, hydroxylation sites, etc.

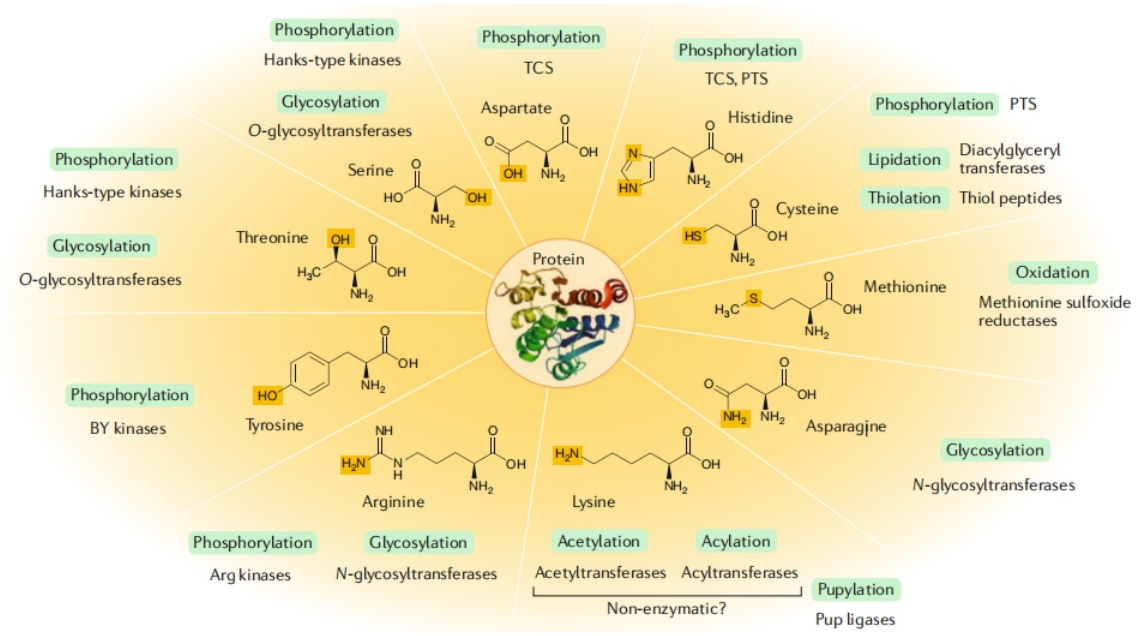


Fig. 2.1. Different types of post-translational modifications [24].

In this study, we have already experimented on multiple PTM sites i.e. acetylation, glutarylation, succinylation, crotonylation, methylation sites. Moreover, prediction of formylation and phosphoglycerylation have also been experimented parallelly on our other studies.

b) **Non-Enzymatic:** It does not need any enzymes to be present. For example, glycation, carbonylation, cambamylation, etc.

Approximately, 5% of the proteome consists of enzymes that carry out more than 200 PTM forms [25]. Enzymes are macromolecular organic molecules that induce the biological reactions of living organisms. Enzymatic shifts are most frequently found in PTMs. These enzymes include kinases, phosphatases, transferases, and ligases that add or remove functional groups, proteins, lipids, or sugars from or onto amino acid chains and proteases that connect peptides to extract certain sequences or regulatory subunits [26]. Post-translational modification of certain forms is a consequence of oxidative stress. The overflow of free radicals indicates oxidative stress. Anti-oxidant assists in the donation of electrons. Oxidative stress represents the equilibrium between the systemic manifestations of the reactive oxygen species and the readily reactive regulation of the biological system or the consequent lack of recovery.

2.3 Process of Modification

Post-translational changes can occur in the amino acid side chains or the C-or N-terminal proteins. The C-terminus (also known as carboxyl-terminus, carboxy-terminus, C-terminal tail, C-terminal end, or COOH-terminus) is the end of the amino acid chain (protein or polypeptide) terminated by the free carboxylic group (-COOH). The N-terminus (also known as the amino-terminus, the NH2-terminus, the N-terminal end, or the amine-terminus) is the beginning of the protein or polypeptide of the free amine group (-NH2) at the end of the polypeptide [27].

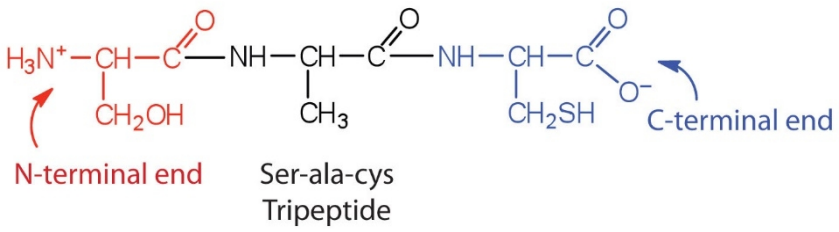


Fig. 2.2. C-termini and N- termini of amino acids [28].

In the figure, the blue portion represents the C-terminus as it contains the carboxyl group (-COOH) and the red portion represents the N-terminus as it contains the amine group (-NH2). The process of introducing the acetyl function group into the chemical compound is acetylation, a common PTM. N-acetylation is an example of an amino group variation. It is an enzymatic mutation catalyzed by a set of enzyme complexes called N-terminal acetyltransferases (NAT). NATs transfer the acetyl group from the acetyl-coenzyme A (AC-CoA) to the amino acid group of the first protein residue. The nascent protein is responsible for the synthesis of the N-terminal allele, and acetylation has yet to be proven to be irreversible [29]. The procedure of how N-terminal acetylation occurs with the presence of catalyst NATs is shown below:

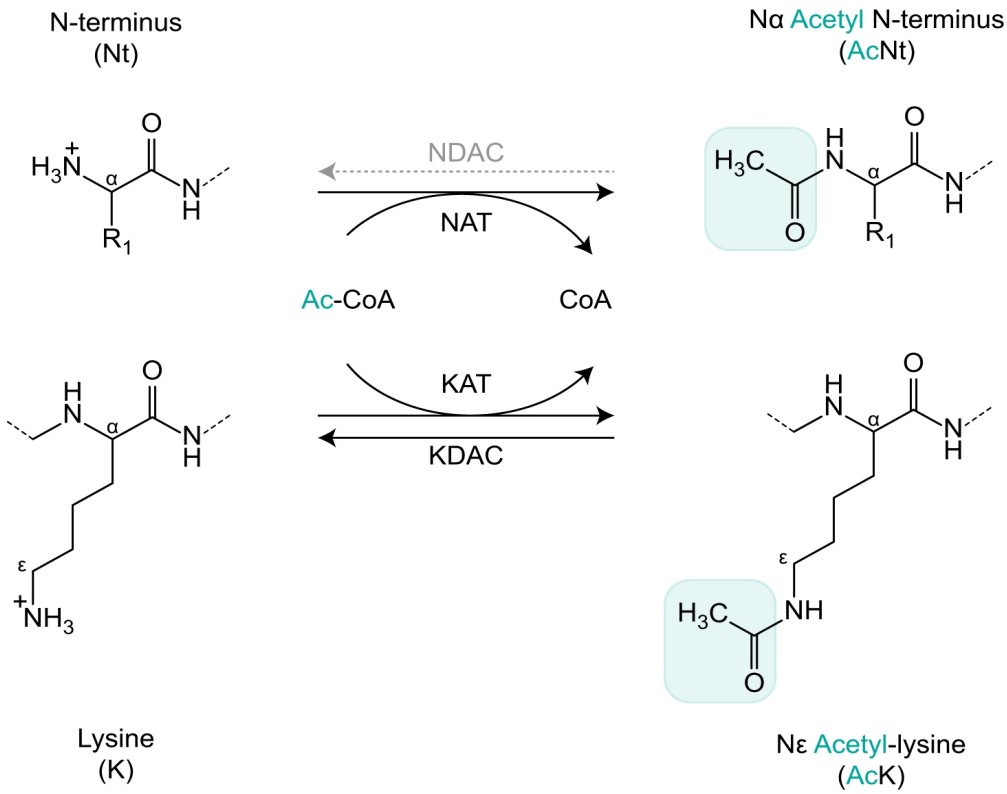


Fig. 2.3. N-terminal acetylation [30].

2.4 Effects of PTMs

Nowadays, PTMs have started to gain increased attention, as new responsive proteomics workflows and computational strategies now enable researchers to obtain large-scale, in-depth, and unbiased information on PTM type and place [31]. These PTMs are responsible for the variation of proteins in function and shape, and the plasticity and motility of the active cells are affected by PTMs [32]. Expansion of the genetic code and manipulation of the physiology of living cells are the product of PTMs [31,33]. Sometimes, post-translational changes can lead to various diseases. Abnormal PTM events occur extensively in a variety of neurodegenerative disorders, such as Alzheimer's Disease (AD) which are associated with neuronal dysfunction and cell death, and also other diseases i.e. diabetes, Parkinson's disease, chronic renal failure, chronic lung disease, sepsis, etc [34]. For more convenience, the associated diseases correlated with our experimented five PTMs are shown below:

Acetylation: In eukaryotes, the role of acetylation focuses primarily on the control of the configuration of the cell chromosome and the activation of nuclear transcription factors. In prokaryotes, protein acetylation is primarily expressed in the following aspects: the immediate effect of enzyme activity, the effect of protein association, the effect of mitochondrial flow [35].

Methylation: The important role of methylation is in gene regulation, many human diseases such as cancer, coronary heart disease, multiple sclerosis, rheumatoid arthritis, and neurodegenerative disorders [34].

Succinylation: some abnormalities and changes in succinylation is correlated with the pathogenesis of many diseases including tumors, cardiac metabolic diseases, liver metabolic diseases, and nervous system diseases [35].

Crotonylation: Previous experiments have shown that histone crotonylation is elevated in tubular kidney cells during acute kidney injury (AKI), but the enzymes responsible for the addition or removal of lysine crotonylation in cellular remain largely unknown [36].

Glutarylation: Glutarylation dysregulation has been related to the etiology of metabolic disorders such as cancer, mycobacterium tuberculosis, diabetes, and brain and liver disorders [37].

The summarized disease association with several PTMs (PDA) is shown in fig. 2.4.

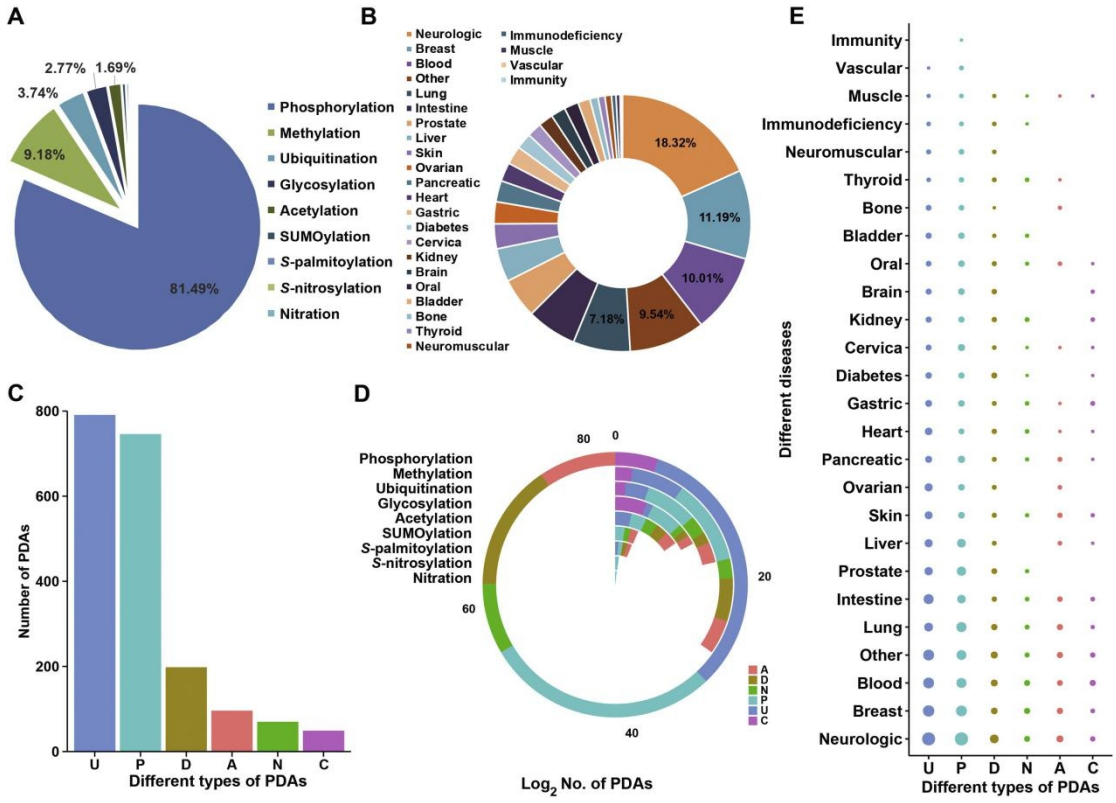


Fig. 2.4 The data is from PTMD [27]

A)The percentage of different PTM supertypes that are associated with diseases.

B)The percentage of different super-types of diseases in PTMD.

C)The number of disease associations within each type of disease-associated PTMs.

D)The number of disease associations within each type of PDAs for each PTM super-type

E)The distribution of the six types of PDAs for each disease super-type.

The dot size represents the number of PDAs for each disease. U and D indicate that the PTM level is upregulated and downregulated in diseases, respectively [38].

2.5 Responsible Residue of PTMs

Each post-translational modification has a set of residues on which that modification works on. A modification can occur both on a single residue or a set of residues. But each modification associated with individual residue represents a different mechanism that corresponds to that specific residue. The following table contains a set of post-translational modifications with the respective residues.

Table 2.1. PTMs with responsible residues [35], [36].

Amino Acid	**Modifications**
Alanine	N-acetylation (N-terminus),Methylation
Arginine	Methylation, N-acetylation
Asparagine	N-linked glycosylation, N-acetylation, Methylation
Aspartic acid	Isomerization to isoaspertic acid, N-acetylation
Cysteine	Oxidation to sulfenic, sulfinic or sulfonic acid, palmitoylation, N-acetylation (N-terminus), S-nitrosylation, Methylation
Glutamic acid	Cyclization to Pyroglutamic acid (N-terminus), deamidation to Glutamic acid, or isopeptide bond formation to a lysine by a transglutaminase
Glutamine	Cyclization to Pyroglutamic acid (N-terminus), gamma-carboxylation
Glycine	N-Myristoylation (N-terminus), N-acetylation (N-terminus), Methylation
Histidine	Phosphorylation
Isoleucine	Amidation
Leucine	Amidation, glycosylation, ubiquitination
Lysine	Acetylation, Ubiquitination, SUMOylation, methylation,

	Hydroxylation, succinylation, malonylation,gluterylation
Methionine	N-acetylation (N-terminus), oxidation to sulfoxide or sulfone
Phenylalanine	None
Proline	Hydroxylation
Serine	Phosphorylation, O-linked glycosylation, N-acetylation (N-terminus)
Threonine	Phosphorylation, O-linked glycosylation, N-acetylation (N-terminus)
Tryptophan	Mono- or di-oxidation, formation of Kynurenine
Tyrosine	Sulfation, phosphorylation, nitration
Valine	N-acetylation (N-terminus)

In this study, we have experimented with five PTM sites. The responsible residues for these five PTM sites are:

Acetylation: Alanine (A), Arginine (R), Aspartate (D), Cysteine(C), Glycine(G), Glutamate (E), Lysine (K), Methionine (M), Proline (P), Serine (S), Threonine (T), Valine(V).

Methylation: Alanine (A), Arginine (R), Aspartate (D), Cysteine(C), Glycine(G), Glutamate (E), Glutamin (Q), Histidine(H), L, Lysine (K), Methionine (M), Proline (P).

Crotonylation: Lysine (K).

Succinylation: Lysine (K).

Glutarylation: Lysine (K).

Among all these residues for these five PTM sites, Lysine (K) residue is common for all, and also a large amount of data are explored compared to the other residues. Moreover, lysine has gained much attention due to its impact on drug development and different diseases mentioned in the previous section. So, we have decided to choose lysine (K) as a common residue and identify it concurrently for all these five PTM sites.

2.6 Related Studies in PTM Site Prediction

There are several computational tools for predicting various K-PTMs separately, such as, iSuc-PseOpt, Formyl_Pred, DeepAcet, Glut_Pred, predCitru, Bigram-PGK, PhoglyPseAAC, etc [8], [9], [10], [11], [12]. These tools are single label predictor which can predict only one type PTM according the functionality. On contrast, to the best of the authors' knowledge, only three multi-label prediction systems i.e. iPTM-mLys, mLysPTMpred and predML-Site have been developed so far that can take care of the multiplex Lys residues [5], [13], [14], [15]. That means, these predictors can predict several modifications simultaneously. The detailed insights about these multi-label PTM predictors are illustrated in Table 2.2 with corresponding methodologies and performances.

Among these prediction systems, the first one named as iPTM-mLys was proposed by W. R. Qiu et al., where the dataset was collected from the Universal Protein Resource (UniProt) [27] for predicting acetylation, crotonylation, methylation, and succinylation modification prediction. iPTM-mLys had utilized a vectorized sequence-coupling model as their feature construction method and applied four random forest classifiers for each of the modification prediction [5]. Later, mLysPTMpred was constructed by M.A.M. Hasan et al. where the datset of iPTM-mLys study was utilized. In this approach, the vectorized sequence-coupled model was used in line with four SVM classifiers for each of the PTM identification. This method obtained a comparatively higher performance as compared to the iPTM-mLys study [13]. After that, predML-Site prediction method was constructed by us for further improvement in K-PTM study. This method utilized the same dataset which was constructed by iPTM-mLys study with several feature extraction techniques such as vectorized sequence-coupling, composition of k-spaced amino acid pairs (CKSAAP), amino acid factor (AAF) and binary encoding (BE). Four separate SVM classifiers were used for classification. The integration of CKSAAP, AAF and BE with the vectorized sequence-coupling technique improved the prediction performance for the predML-Site. In spite of having a higher performance rate, an absolute necessity of further performance elevation with enriched functionality was felt and thus a new system iML-LysPTM has been proposed in this study.

Table 2.2. Summary of the related works.

Paper	Publish Date	Function	Feature	Classifier	Accuracy (%)
iPTM-mLys [5]	2016	4 K-PTMs	Coupling	RF	68.37
mLysPTMpred [13]	2018	4 K-PTMs	Coupling	SVM	83.73
predML-Site	Submitted (08 August, 2020)	4 K-PTMs	Coupling, BE, C KSAAP, AAF	SVM	84.12

Our primary objective is to identify several K-PTMs simultaneously and improve the prediction performance (see Table 2.2) obtained by the previous studies. To achieve more functionality, we aim to include one more type of post-translational modification that previous works lack. Furthermore, our goal is to aid the experimental scientists in further post-translational modification studies.

2.7 Our Contributions in PTM Site Prediction

We have developed two single label PTM predictor such as, Formyl_Pred [9] and predPhogly-Site which can predict proteins formylation and phosphoglycerylation modifications respectively. We have constructed the multi-label predictor predML-Site which can predict four PTMs concurrently. However, the detailed procedures of these prediction systems are not in the scope of this study. We intend to describe about our latest work iML-LysPTM. The contributions behind this prediction tool construction are as follows,

- **Data acquisition:** In this study, a new dataset of five K-PTMs (i.e. acetylation, crotonylation, methylation, succinylation and glutarylation was constructed. The proteins have been collected from the Universal

Protein Resource (UniProt) [27] and later by preprocessing these proteins, the final benchmark dataset has been constructed.

- **Feature optimization:** Three feature encoding techniques such as sequence-coupling, binary encoding (BE), and amino acid factor (AAF) have been considered. Later these encoded features have been combined. After that ANOVA F-Test feature selection technique has been dopted to reduce the dimensionality.

- **Data imbalance management:** Five separate SVM classifiers have been used for five respective K-PTMs. The different error cost (DEC) method has been applied to handle the data imbalancement.

- **Web-server:** A user-oriented web-server has been developed to aid the experimental researchers for future study.

2.8 Conclusion

In this chapter, we have addressed post-translational changes and their consequences. We have briefly explained how the lateral chain or C-or N-terminals of the protein shift and various post-translational modifications are shown with corresponding residues. Additionally, some of the related works in both single label and multi-label PTM site prediction are described in a detailed manner.

CHAPTER 3

Data Acquisition and Feature Optimization

Introduction

Benchmark Dataset

Sample Formulation

Feature Extraction

Conclusion

3.1 Introduction

The primary objectives of any prediction system include data collection and preprocessing as the quality and robustness of a predictor depend on these steps. There are various protein repositories, such as Universal Protein Resource (Uniprot) [27], Protein Lysine Modification Database (PLMD), dbPTM, etc. We have utilized one of these widely recognized protein databases and applied the sliding window method with Chou's scheme [28] to formulate the obtained protein sequences. Later five benchmark datasets have been constructed for each of the K-PTM. After preliminary analysis, it was observed that the vectorized sequence coupling, encoded binary features, and amino acid factor were more appropriate for representing the protein sequences. Additionally, ANOVA F-Test statistic with incremental feature selection helped us to draw the necessary features out of the combined features.

3.2 Benchmark Dataset

In the current study, human protein sequences were utilized for prediction model development and benchmarking. The benchmark dataset was collected from the Universal Protein Resource (UniProt) [27]. Various constraints had been applied to derive the dataset. The number of proteins obtained after applying the query "annotation:(type:mod_res) AND reviewed:yes AND organism:"Homo sapiens (Human) [9606]"" was 9380 (accessed 22 September 2020). As this study was concerned with a multi-label classification problem, 5 different types of K-PTMs (i.e. acetylation, crotonylation, methylation, succinylation, and glutarylation) were considered when constructing the dataset. After applying a preliminary selection process with the specific keywords of each K-PTM, 1842 proteins were obtained. To demonstrate the viability of the proposed predictor iML-LysPTM, we randomly selected 184 proteins (10%) to develop an independent test dataset while the rest was utilized as the training dataset. The training dataset contained 4509 K-PTM sites and 73033 non-K-PTM sites. For reducing the skewness, primarily we applied Cd-hit on the non-K-PTM sites with a 30% identity cut-off. Later common sequence fragments between the K-PTMs and non-K-PTMs were removed and 4490 K-PTM sites with

43785 non-K-PTM sites were obtained. A comprehensive summary of dataset preparation is presented in Fig. 3.1.

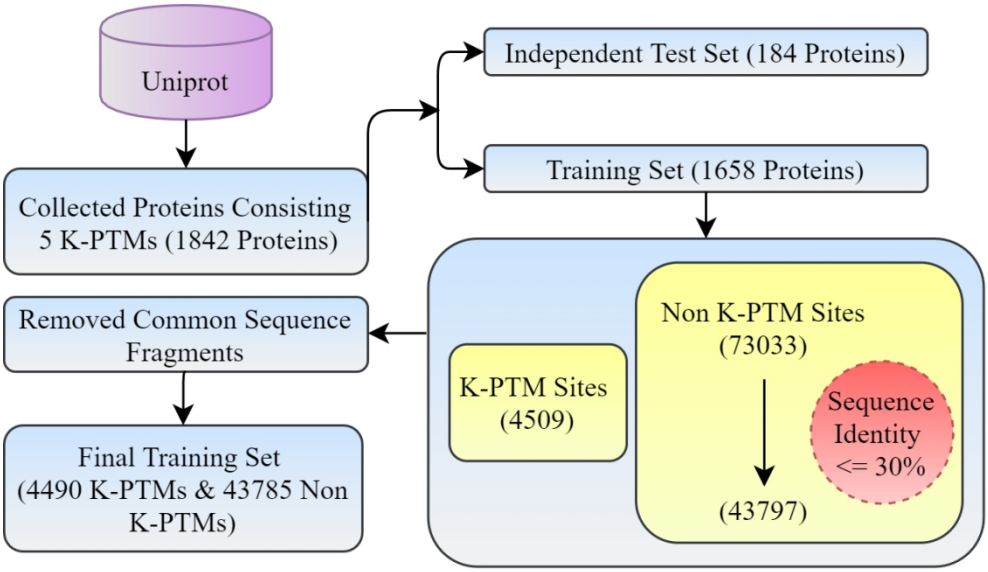


Fig. 3.1. Overview of Dataset Construction.

3.3 Sample Formulation

For formulating PTM sites meticulously and comprehensively, Chou's scheme [28] was adopted. According to this scheme, a peptide segment could generally be expressed by

$$P_{\zeta}(K) = Q_{-\zeta}Q_{-(\zeta-1)} \ldots\ldots Q_{-2}Q_{-1}\, K\, Q_{+1}Q_{+2} \ldots\ldots Q_{+(\zeta-1)}Q_{+\zeta} \qquad (3.1)$$

where the symbol K denoted the responsible residue 'lysine' at the center, the subscript ζ being an integer, $Q_{-\zeta}$ and $Q_{+\zeta}$ denoted the ζ-th leftward and ζ-th rightward amino acid residues from the center, and so forth. Furthermore, a peptide sequence $P_{\zeta}(K)$ could be categorized into two types [13]

$$P_{\zeta}(K) \;\epsilon\; \begin{cases} P_{\zeta}^{+}(K), \; if\ its\ center\ is\ K - PTM\ site \\ P_{\zeta}^{-}(K), \; otherwise \end{cases} \qquad (3.2)$$

where $P_\zeta^+(K)$ contained the positive subset of the peptides and $P_\zeta^-(K)$ contained the negative subset of the peptides with a lysine (K) residue at its center, and the symbol $\in$ indicated the set theory relationship [13]. For equal-sized K-PTM site formation, (2ζ+1)-tuple peptide window with K at its center was employed. During segmentation, the lacking amino acid at both the right and left end was filled with the nearest residue [5], [13]. After the peptide fragments went through some screening, such as the elimination of sequences in case of redundancy, a total of five benchmark datasets were constructed with the given form

$$S_\zeta(K-type) = S_\zeta^+(K-type) \cup S_\zeta^-(K-type) \quad (3.3)$$

where the positive subset $S_\zeta^+(K-type)$ could contain only the true K-type peptide samples, while the negative subset $S_\zeta^-(K-type)$ could contain only the false K-type samples with K at the center [5], [13]. It is to be noted that only the specific type of K-PTM, i.e. 'acetylation', or 'crotonylation', or 'methylation' or 'succinylation' or 'glutarylation' must be used separately and consistently as the 'K-type' described in Eq. (3.3). After going through many preliminary tests, the window size was chosen as (2ζ+1)= 27, where ζ=13. Therefore, Eq. (3.1) had been reduced to

$$P(K) = Q_{-13}\, Q_{-12} \ldots\ldots Q_{-2}\, Q_{-1}\, K\, Q_{+1}\, Q_{+2} \ldots\ldots Q_{+12}\, Q_{+13} \quad (3.4)$$

The overviews of the benchmark dataset and the independent test set obtained by this process are provided in Table 3.1, Table 3.2, and Table 3.3.

Table 3.1. Five Benchmark Datasets Overview.

Attribute	Ace	Cro	Met	Succ	Glut
Positive	3659	201	294	1112	189
Negative	44616	48074	47981	47163	48086

Ace: Acetylation, Cro: Crotonylation, Met: Methylation, Succ: Succinylation, Glut: Glutarylation

Table 3.2. Five Independent Test Datasets Overview.

Attribute	Ace	Cro	Met	Succ	Glut
Positive	502	12	35	151	61
Negative	8582	9072	9049	8933	9023

Ace: Acetylation, Cro: Crotonylation, Met: Methylation, Succ: Succinylation, Glut: Glutarylation

Table 3.3. Frequency of K-PTMs in Training Set and Independent Test Set.

Attribute	1 K-Type	2 K-Types	3 K-Types	4 K-Types	5 K-Types	Non-K-Types
Training Set	3641	849	81	29	6	43785
Independent Test Set	446	141	30	3	0	8497

3.4 Feature Optimization

With the evolution of the biological sequences, several encoding methods have been developed for extracting pertinent features hidden in the sequences. After exploratory analysis, three feature encoding techniques, i.e. the vectorized sequence coupling [14], encoded binary features and amino acid factor [11], [17] were chosen to be more appropriate for representing the protein sequences of the multiple lysine modification sites than any other encoding methods. To draw out the features required in the combined features, ANOVA F-Test statistic with the incremental feature

selection was also used. A summary of the optimized feature construction procedure is depicted in Fig. 3.2.

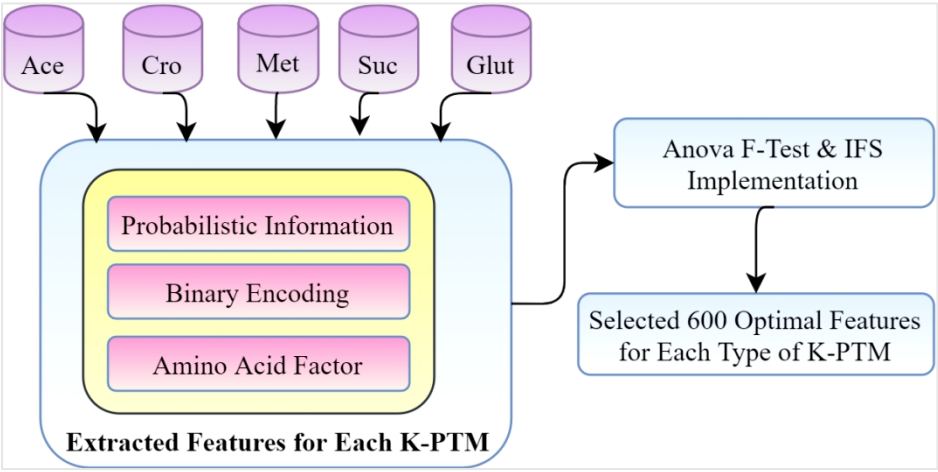


Fig. 3.2. An overview of feature construction steps.

3.4.1 Sequence-Coupling

The composition of pseudo amino acid or PseAAC [29], [30], [31], [32], [33] was designed to preserve the sequence pattern information, which is a much harder task for any existing machine learning algorithm [34]. In this study, incorporating sequence coupling information into Chou's general PseAAC was adopted for extracting features from peptide sequences [5], [13], [14], [15], [30]. Based on this conception, the segmented proteins of Eq. (4.1) could be defined as,

$$P(K) = P^{+}(K) - P^{-}(K) \tag{3.5}$$

where,

$$P^{+}(K) = \left[\; P^{C+}_{-13} \; P^{C+}_{-12} \; \ldots\ldots \; P^{+}_{-1} \; P^{+}_{+1} \ldots\ldots \; P^{C+}_{+12} \; P^{C+}_{+13} \; \right]^{T} \tag{3.6}$$

$$P^{-}(K) = \left[\; P^{C-}_{-13} \; P^{C-}_{-12} \; \ldots\ldots \; P^{-}_{-1} \; P^{-}_{+1} \ldots\ldots \; P^{C-}_{+12} \; P^{C-}_{+13} \; \right]^{T} \tag{3.7}$$

where P_{-13}^{C+} in Eq. (4.2) denoted the conditional probability [5], [13] of amino acid Q_{-13} at the leftmost position given that its adjacent right member was Q_{-12} and so forth [35].

In contrast, only P_{-1}^{+} and P_{+1}^{+} were of non-contingent probability as *K* was the adjoining member of both amino acids at position Q_{-1} and Q_{+1}. All the conditional probability values were extracted from the positive training dataset [5], [8], [12], [13]. Additionally, all the probability values in Eq. (4.3) were identical to those of Eq. (4.2) other than that they could be derived from the negative training dataset. Thus, after omitting *K* from the center, *(27-1)=26* dimension feature vectors were obtained.

3.4.2 Binary Encoding

Binary encoding [11] could represent the amino acid position and composition by using *20* binary bits for one amino acid [11]. But one additional bit was conjoined to handle the complexity of sliding windows. For *21* amino acids structured as *'ACDEFGHIKLMNPQRSTVWYZ'*, each residue inside a sequence fragment was formed by a *21*-dimension binary vector [11]. For instance, residue *'A'*, *'G'* and *'Z'* were encoded as *'100000000000000000000'*, *'000000100000000000000'* and *'000000000000000000001'* respectively. According to this concept, each resultant peptide segment was expressed as *27×21 = 567*-dimensional feature vectors.

3.4.3 Amino Acid Factor

Five multidimensional attributes [17], [36], which included polarity, secondary structure, molecular volume, electrostatic charge, and codon diversity [23], were constructed from AAIndex by using multivariate statistical analysis [11]. These five transformed properties could be introduced as amino acid factors (AAF) [23]. Since the AAF could reduce the dimensionality of the feature space of physicochemical properties efficiently, it had been utilized in many biological studies [11], [23], [37]. The dimensionality of feature vectors was calculated as follows,

$$D = peptide\ sequence\ length \times number\ of\ factors \quad (3.8)$$

With a peptide sequence of length *27* and previously described five amino acid factors, *27×5 = 135* dimension features were derived by using this formula.

3.4.4 Feature Ensembling

Initially, the three aforestated feature encoding techniques (i.e. sequence coupling, BE, and AAF) were implemented separately to encode the training peptides. However, for extracting more PTM-contextual information from the protein sequences, encoded features were ensembled serially, scaled through standardization, and obtained *26 + (27×21) + (27×5) = 728* features.

3.4.5 Analysis of Variance

Since the dimension of the encoded features was higher, irrelevant, and redundant features needed to be removed to avoid learning complexity. For this reason, the analysis of variance (ANOVA) F test statistic technique [18], [38] was adopted in this study. It tested the null hypothesis (i.e. all the means of different groups were equal) against the alternative hypothesis (i.e. all the means differed from each other). The one-way ANOVA could be defined as

$$F = \frac{(n-k)\sum n_i\,(\overline{Y}_{i.}-\overline{Y}_{..})^2}{(k-1)\sum (n_i-1)s_i^2} \tag{3.9}$$

where

$$n = \sum_{i=1}^{k} n_i \ , \quad \overline{Y}_i = Y_i/n_i \ , \quad \overline{Y}... = Y.../n$$

and

$$s_i^2 = \sum_{j=1}^{n_i} \left(Y_{ij} - \overline{Y}_{i.}\right)^2/(n_i - 1)$$

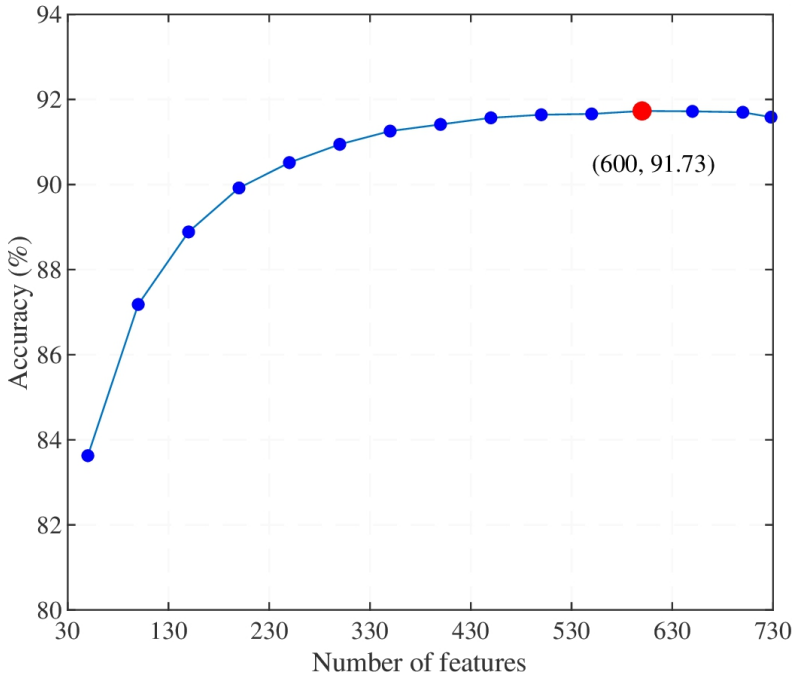
It should be mentioned that the dot in $Y_{i.}$ indicates an aggregation over the j index [38], where

$$Y_{i.} = \sum_{j=1}^{n_i} Y_{ij} \quad \text{and} \quad Y... = \sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}$$

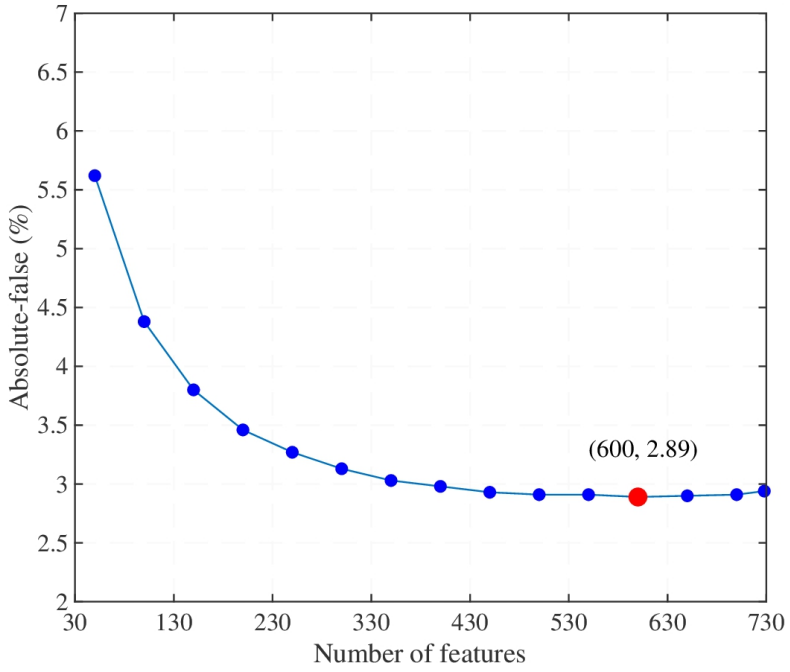
The calculated *F* values were used to rank the features. for higher *F* values, the discriminative capability of a predictor is better.

3.4.6 Incremental Feature Selection

The incremental feature selection (IFS) [19] algorithm was applied for selecting the optimal number of features [11], [53]. For each feature subset of top *m (m = 50, 100, 150, ..., 728)*, one SVM classifier with LibSVM's default parameter [40], [54] , [55],[56],[57],[58] was trained for each K-PTM type and measured its accuracy and absolute-false rate by adopting 5-fold cross-validation. The accuracy and absolute-false rate are two very important performance measurement metrics that will be discussed thoroughly in chapter 4. As depicted in Fig. 3.3a and 3.3b, the highest accuracy of 91.73% with the lowest absolute-false rate of 3.47% was achieved with 600 leading features. Finally, the proposed predictor iML-LysPTM was constructed by utilizing the top 600 features.



(a)



(b)

Fig. 3.3. The IFS curves: (a) Feature range 50 to 728 (Features Vs Accuracy); (b) Feature range 50 to 728 (Features Vs Absolute-false).

3.4.7 Optimal Features Analysis

The feature distribution for different K-PTM types was shown in Fig. 3.4. For the acetylation feature set, out of 600 optimal features, 26 belonged to the sequence-coupling, 456 belonged to the BE, and 118 belonged to the AAF. Therefore, the ratios of selected dimensions of these three types of features are 100% (26/26), 80.42% (456/567), and 87.41% (118/135), respectively.

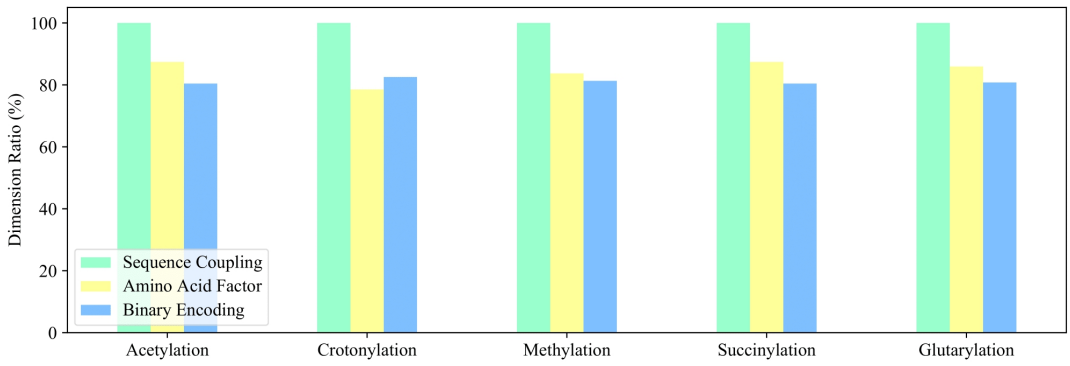


Fig. 3.4. Feature distribution in the optimal feature sets.

The crotonylation feature set was made of 26 sequence coupling features, 468 BE features, and 106 AAF features. The ratios of selected dimensions of three types of features are 100% (26/26), 82.54% (468/567), and 78.52% (106/135), respectively. Besides, the methylation feature set consisted of 26 sequence-coupling features, 461 BE features, and 113 AAF features, and selected dimensions ratios for each type of feature are 100% (26/26), 81.31% (461/567), and 83.70% (113/135), respectively. For the succinylation dataset, 26, 456, and 118 features belonged to the sequence-coupling, BE, and AAF respectively. The dimension ratios for sequence-coupling, BE, and AAF are 100% (26/26), 80.42% (456/567), and 87.41% (118/135), respectively. And for glutarylation set, 26, 458, and 116 features belonged to the sequence-coupling, BE and AAF respectively and the dimension ratio for sequence-coupling is 100% (26/26), for BE it is 80.78% (458/567), and for AAF it is 85.93% (116/135).

As reflected in Fig. 3.4, the sequence-coupled features had a stronger influence while the BE and AAF features had a moderate impact on the identification of acetylation, crotonylation, methylation, succinylation, and glutarylation sites. The selected feature dimensions for BE and AAF varied over different types of K-PTM site prediction. It may be concluded that the proposed model augmented the sequence-coupling effect with the essential features of BE and AAF and intensified the prediction performance.

3.5 Conclusion

This chapter consists of data preprocessing, peptide segmentation techniques, redundancy reduction criteria, and a thorough discussion on different feature encoding schemes along with the basic understanding of the feature selection criteria according to the benchmark datasets used in this study.

CHAPTER 4
Methodology

Introduction

Support Vector Machine

Cross-Validation

Independent Test

Evaluation Metrics

Model Development

Conclusion

4.1 Introduction

The support vector machine (SVM) [39], [40], [41], [59], [60] one of the dominant statistical learning algorithms was adopted as a core prediction algorithm. It seeks the optimum hyperplane with the highest margin between the two groups. In this chapter, we tend to understand the basic SVM prediction engine initially. To evaluate the statistical significance of a novel predictor's anticipated performance, three validation schemes, such as k-fold cross-validation, jackknife test, and independent test are widely used [13]. We have utilized the validation schemes which are suitable for our study. Later, we have measured the prediction performance of our predictor iML-LysPTM with five widely used multi-label metrics.

4.2 Support Vector Machine

The support vector machine solves the problem of constraint optimization as described below

$$maximize_{\alpha} \ \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \ \sum_{i=1}^{n} \ \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j k\,(x_i x_j) \qquad (4.1)$$

Subject to: $\sum_{i=1}^{n} y_i \alpha_i = 0$, $0 \leq \alpha \leq C$, for all *i=1,2,3,..,n*.

After involving the kernel function [13], the discriminant function of SVM took the following form

$$f(x) = \sum_{i}^{n} \alpha_i y_i k(x, x_i) + b \qquad (4.2)$$

In this paper, the radial basis function kernel [13], [42] was applied to construct SVM classifier and given by, $k(x_i, x_j) = \exp\left(-\gamma \left\| x_i - x_j \right\|^2\right)$, where $\gamma > 0$ [19], [43]. Using Eq. (4.2), five separate SVM classifiers have been utilized for each of the K-PTM types. The datasets have been preprocessed for each K-PTM and fed to each classifier as input. The hyperparameters for each classifier have been selected as, *C=1* and $\gamma = 1$ / number of features. As the benchmark dataset was highly imbalanced,

different error costs (DEC) [13] method had been used to tackle the class imbalance problem [21], [44],[61],[62],[63]. According to this approach, the SVM soft margin objective function [13],[64],[65] was adjusted to allocate two costs for misclassification [11], [12], [19],[66],[67],[68],[69] such as C^+ for the positive class instances and C^- for the negative class instances

$$C^+ = C \times W^+ \quad , \qquad C^- = C \times W^- \tag{4.3}$$

In Eq. (5.3), W^+ is the weight for the positive instances and W^- is the weight for the negative instances and defined by

$$W^+ = \frac{M}{2 \times M_1} \quad , \qquad W^- = \frac{M}{2 \times M_2} \tag{4.4}$$

where M is the total number of elements, M_1 is the number of elements for the positive class, and M_2 is the number of elements for the negative class. There are five sets of positive and negative weights for five classifiers, as the distribution of positive and negative classes differ from each other. The detailed descriptions of the datasets used for each classifier have been given in Table 3.1, and 3.3.

Finally, the output from the five separate classifiers have been merged to obtain the final multi-label prediction output. Later, the performances have been measure by using this multi-lable output.

4.3 Cross-Validation

Among various statistical techniques for validating the performance of prediction models, K-fold cross-validation [13] is the widely used method, particularly for time complexity. Additionally, to establish a fair comparison with the state-of-the-art K-PTM prediction methods, a 5-fold cross-validation test [13],[45],[70],[71],[72], [73],[74] has been executed five times (i.e. 25 iterations in total), because of having insufficient details about the exact 5-way splits [13]. In each 5-fold cross-validation, the training dataset was randomly divided into five approximately equal-sized disjoint sets [13], [19], [75],[76],[77]. The average results of all metrics were reported with their respective standard deviations for the evaluation of the novel predictor.

4.4 Independent Test

To demonstrate the viability of the proposed predictor iML-LysPTM for new proteins, an independent test set has been constructed with 10% of the total proteins obtained from the Uniprot database. These protein sequences are utterly unknown to the predictor as they are not present in the benchmark datasets used for prediction model development. The summary of the independent test set was discussed in chapter 3.

4.5 Evaluation Metrics

In our benchmark dataset, peptide samples were 48275 in total, of which 3659 were labeled with 'acetylation', 201 with 'crotonylation', 294 with 'methylation', 1112 with 'succinylation', 189 with 'glutarylation', and 43785 with 'non-K-PTM'. Since a sample can contain more than one labels, metrics for multi-label systems [5], [13] were utilized instead of ordinary metrics for single-label systems [9], [10], [11], [12], [19], [47], [78], [79], [80], [81], [82], [83], [84]. For evaluation, we have estimated Aiming, Coverage, Accuracy, Absolute-True and Absolute-False [22], [85], [86] (defined in Eq. (4.5)).

$$\begin{cases} Aiming = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\|Y_i \cap Y_i^{'}\|}{\|Y_i^{'}\|}\right) \\ Coverage = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\|Y_i \cap Y_i^{'}\|}{\|Y_i\|}\right) \\ Accuracy = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\|Y_i \cap Y_i^{'}\|}{\|Y_i \cup Y_i^{'}\|}\right) \\ Absolute - True = \frac{1}{N}\sum_{i=1}^{N}\left(\Delta\left\|Y_i, Y_i^{'}\right\|\right) \\ Absolute - False = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\|Y_i \cup Y_i^{'}\| - \|Y_i \cap Y_i^{'}\|}{L}\right) \end{cases} \quad (4.5)$$

where N and L were the total number of the samples and the total number of labels in the system respectively [5], [13], $\cup$ and $\cap$ denoted the 'union' and

'intersection' in the set theory, || || meant the operator acting on the set to calculate the number of its elements [13], Y_i and $Y_i^{'}$ denoted the subset that contained all the labels experiment-observed and all the labels predicted [5], [13] for the i^{th} sample respectively, and

$$(\Delta Y_i, Y_i^{'}) = \begin{cases} 1, & \text{if all labels in } Y_i \text{ and } Y_i^{'} \text{ are identical} \\ 0, & \text{otherwise} \end{cases}$$

The metrics defined above have been applied effectively in several multi-label-based systems [5], [13].

4.6 Model Development

In this study, five separate SVM classifiers [13] were used to predict the acetylation, crotonylation, methylation, succinylation, and glutarylation sites separately. Each of the classifiers performed binary classification on the benchmark dataset described in Table 3.1. For all five K-PTM types, necessary features were extracted by integrating multiple encoding methods and selected 600 optimal features with ANOVA F-test to train the models, as shown in Fig. 4.1.

The radial basis function (RBF) kernel [22], [55] was used for each SVM classifier. As there was a lack of details about the exact 5-way splits of the dataset [22], five complete runs of 5-fold cross-validation were executed [5], [13], [56],[58],[59],[61],[63],[64]. The misclassification cost *C* was calculated according to Eq. (5.3) and (5.4) for handling the data imbalance issue. In this study, LibSVM's default parameters (i.e. *C=1* and $\gamma = 1$ / number of features) were selected to train the model. Eventually, after training the five binary SVM classifiers with the appropriate hyperparameters, multi-label predictor iML-LysPTM was constructed by combining the outputs from these classifiers [22], as depicted in Fig. 4.1. Five times repetition of the 5-fold cross-validation [22] produced five sets of values of all metrics defined in Eq. (4.5). The average results of each multi-label metric were taken to evaluate the final model.

Later, an independent test set evaluation was performed for further verification. Finally, our proposed predictor was deployed as a web-server for aiding future researchers in terms of the PTM study. It should be mentioned that Matlab 2019a and python 3.7.3 were utilized to implement the system.

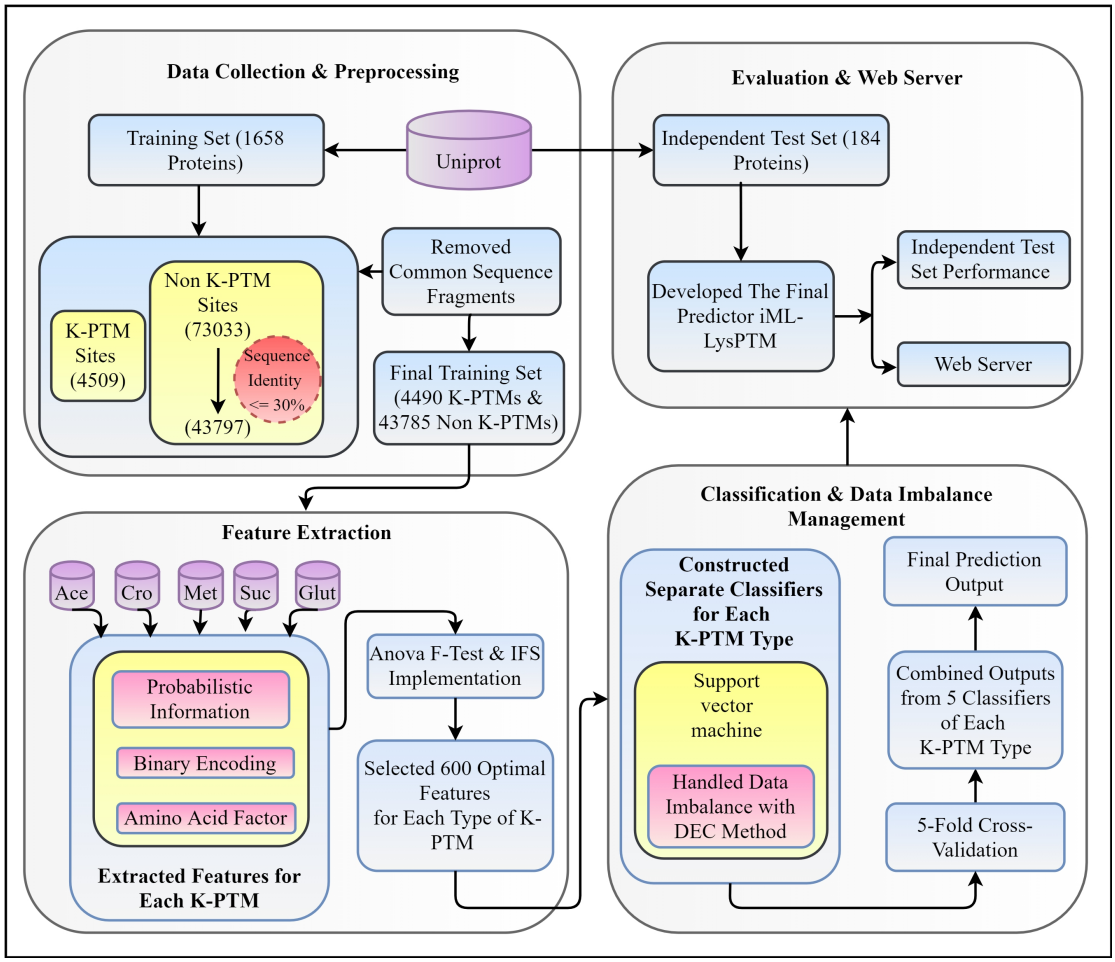


Fig. 4.1. The system flowchart of iML-LysPTM.

4.7 Conclusion

In this chapter, we have gone through the classification technique applied in this research. Then we have discussed how to handle the data imbalance problem with variable cost adjustment. Moreover, we have discussed different validation criteria, such as cross-validation, independent test, and introduced to the evaluation metrics for the multi-label system. Finally, we have developed the final model iML-LysPTM after rigorous experimentation and deployed it as a web-server.

CHAPTER 5

Experimental Analysis

Introduction

Performance of iML-LysPTM

Performance of Different Features

Comparison with Existing Predictors

Performance of Independent Test

Web-Server

Conclusion

5.1 Introduction

This chapter contains the performance measurements of our proposed predictor iML-LysPTM. We have compared the performance obtained by our predictor with existing predictors. We have also shown the corresponding outcomes achieved by different feature extraction techniques which have been discussed in chapter 3. After some comparative discussion, we have shown the best outcome obtained by our experiments, and based on the observations, we have constructed the final predictor iML-LysPTM. Based on the recent collections of publications [5], [8], [13], [19], [22], [23], [24], [25], to develop an efficient predictor with regards to computational biology, one must go through Chou's five-step [12], [13], [26] guidelines: i) generating an acceptable benchmark dataset for training and testing the system, ii) formulating the sequences with the proper mathematical representations, iii) developing a prediction engine or introducing to a powerful prediction algorithm, iv) conducting cross-validation tests properly to evaluate the predictive accuracy, and v) providing an accessible and user-oriented web-server. We have performed each of these steps carefully and reported the detailed procedures in the previous chapters.

5.2 Performance of iML-LysPTM

The performance of the iML-LysPTM predictor derived from the aforementioned multi-label metrics was given in both Table 5.1 and Table 5.2 with corresponding standard deviations. The values of the five metrics were the average result of five times complete run of 5-fold cross-validation on the benchmark dataset.

In Eq. (4.5), for the first four metrics, the higher the rate is, the better the performance will be, and for the last one, it is entirely the opposite [13], [57]. The rate of the most crucial metric 'Accuracy' for our proposed predictor iML-LysPTM is 91.73%. Besides, it has achieved a 91.89% 'Aiming' or 'Precision' rate [57], [58] which represents the average ratio of the predicted labels that hit the target of the original labels [5]. The average ratio of the original labels that were covered by the hits of prediction referred to 'Coverage' [57], [58] is 91.78%. In addition to that, the

experimentally obtained rate of the most stringent and harsh metric 'Absolute-True' [57] is 91.52% which is significant for any multi-label prediction system. Very few multi-label predictors in computational biology could reach over 50% for the absolute-true rate. More specifically, only two predictors, one of which was constructed by us, could be able to reach over 80% absolute-true rate [13]. Furthermore, the rate of 'Absolute-False' or 'Hamming-Loss' [13], [57], [58] denoting the average ratio of completely wrong hits over the total prediction events is 2.89%

5.3 Performance on Different Feature Encoding Schemes

The performance obtained by iML-LysPTM was further compared with multiple baseline K-PTM prediction methods, developed using different feature extraction methods, such as the incorporation of sequence coupling information into general PseAAC [23], binary encoding, and amino acid factors [11], [14], [17], [37], [59] to estimate iML-LysPTM's K-PTM related information extraction capability. The performances of the specified feature encoding schemes evaluated by 5-fold cross-validation are reported in Table 5.1 with the corresponding standard deviations.

Table 5.1. Cross-Validation Performance on Various Feature Encoding Techniques.

Feature Type	**Feature Count**	**Aiming (%)**	**Coverage (%)**	**Accuracy (%)**	**Absolute-True (%)**	**Absolute-False(%)**
AAF	135	87.89	87.81	87.77	87.6	4.24
BE	567	91.57	91.57	91.53	91.44	3.01
Sequence-Coupling	26	82.08	82.17	81.87	81.34	6.26
Combined	728	91.74	91.64	91.58	91.37	2.94
Optimal Features$_{(iML\text{-}LysPTM)}$	**600**	**91.89**	**91.78**	**91.73**	**91.52**	**2.89**

It may be observed that sequence-coupling, which was one of the most crucial encoding strategies, attained a considerably lower accuracy rate of 81.87%, a coverage rate of 82.17%, and an aiming rate of 82.08% with a higher absolute-false rate of 6.26%. However, slightly better results were picked up by The amino acid factor (AAF) schemes. It reached 87.77% accuracy with an 87.89% aiming rate and 87.81% coverage rate. The absolute-false rate was reduced to 4.24% with an absolute-true rate of 87.6%. The binary encoding (BE) technique obtained a higher accuracy rate of 91.53%, aiming rate of 91.57%, absolute-true rate of 91.44%. It has also achieved a lower absolute-false rate of 3.01% than that of the previous schemes.

Integrating all the feature extraction methods was considered a successful approach for developing a multi-label predictor. Therefore, the sequence-coupling was combined with amino acid factor, binary encoding. Consequently, the performance of the integrated features improved and for 728 dimension features, accuracy was increased to 91.58% with the reduced absolute-false rate of 2.94%. Later, 600 optimal features were selected from the high dimension features by conducting ANOVA F-test. By using the libSVM's default parameter value of C and gamma, accuracy and aiming rate were reached 91.73% and 91.89% respectively [54]. The most uncompromising metric absolute-true rate was 91.52%, which is a rare example in bioinformatics. Table 5.1 points out that iML-LysPTM achieved a discernible performance among all the feature encoding techniques described earlier.

5.4 Performance Comparison with Existing Predictors

Two multi-label prediction systems can predict multiple K-PTM sites simultaneously according to the best of the authors' knowledge. Both the predictors were constructed for identifying four types of K-PTMs i.e. acetylation, crotonylation, methylation, and succinylation. One of the multi-label predictors named as iPTM-mLys utilized the random forest (RF) classifier with the sequence-coupling encoding technique. The other predictor termed as mLysPTMpred applied cost-sensitive SVM classifiers with the same encoding method used in the iPTM-mLys predictor. The latter achieved comparatively higher prediction

performance than iPTM-mLys i.e. 83.73% accuracy, 84.82% aiming, 86.56% coverage, and 6.66% absolute-false rate (see Table 5.2). It also surpassed the milestone of reaching over 80% absolute-true rate.

However, we have constructed a novel multi-PTM site predictor iML-LysPTM which can predict 5 K-PTM sites concurrently. In comparison with the recently developed multi-label predictor mLysPTMpred, it can be observed that the rate of the most crucial metric 'Accuracy' for the proposed predictor iML-LysPTM had been increased from 83.73% to 91.73%. Apart from that, the acquired absolute-true rate reached over 90%, which is another significant achievement in a multi-label PTM prediction system. It should also be mentioned that the absolute-false rate was 2.89%, which is 3.55% lower than mLysPTMpred [13].

Table 5.2. Performance Comparison with Existing Predictors on the Respective Datasets.

Predictor	**Functionality**	**Aiming (%)**	**Coverage (%)**	**Accuracy (%)**	**Absolute-True (%)**	**Absolute-False (%)**
iPTM-mLys [5]	4 K-PTMs	69.78	74.54	68.37	60.92	13.40
mLysPTMpred [13]	4 K-PTMs	84.82	86.56	83.73	79.73	6.66
predML-Site	4 K-PTMs	85.25	86.67	84.12	80.39	6.44
iML-LysPTM	**5 K-PTMs**	**91.89**	**91.78**	**91.73**	**91.52**	**2.89**

Furthermore, as for iML-LysPTM, the rate of aiming [57], [58], and the rate of coverage were 91.89% and 91.78% which outperformed both the predictors iPTM-mLys and mLysPTMpred. Therefore, the experimental results reported in Table 5.2 indicated that the novel multi-label predictor iML-LysPTM achieved better performance even after inclusion of one more type of PTM site prediction capability

than both the iPTM-mLys and mLysPTMpred in terms of ‘Aiming’, 'Coverage', ‘Accuracy’, ‘Absolute-True’ and ‘Absolute-False’ [5], [13], [22]. Additionally, it could be observed that iML-LysPTM outperformed our previous predictor “predML-Site” in terms of all the multi-label metrics.

5.5 Performance of Independent Test

In this work, an independent test has been conducted for further evaluation of our proposed model iML-LysPTM. The test set contained 184 proteins completely unknown to the predictor iML-LysPTM (see Table 3.2 and 3.3 for details). The proposed predictor predicted 448 of one K-PTM type, 41 of two K-PTM types, 11 of three K-PTM types, 6 of four K-PTM types, 1 of five K-PTM types, and 8595 of non-K-PTM types.

According to Eq. (6.1), the obtained results were: aiming rate = 89.03%, coverage rate = 88.96%, accuracy = 88.93%, absolute-true rate = 88.82% and absolute-false rate = 4.04%, almost identical to the cross-validation performance delineated in Table 5.2. However, the existing multi-PTM predictors do not account for glutarylation sites to be predicted, we could not report the independent test results for these predictors. The superior performance obtained from both the cross-validation and independent test demonstrates the validity of our proposed model and it could be a high throughput tool for multi-label PTM identification.

5.6 Web-Server

For intensifying user accessibility without the concern of experimental implementations, an easy-to-use web-server for iML-LysPTM has been developed. It can be accessed at http:/103.99.176.239/iML-LysPTM, where proper guidelines for submitting query protein sequences are provided. Users are allowed to submit query sequences either in the input box or in a batch file. For better understanding, a few protein sequences taken from the independent test dataset are included as an example. In addition to that, the benchmark dataset and the training features used for

constructing iML-LysPTM will be provided upon user request. A snapshot of the iML-LysPTM web-server is illustrated in Fig. 5.1.

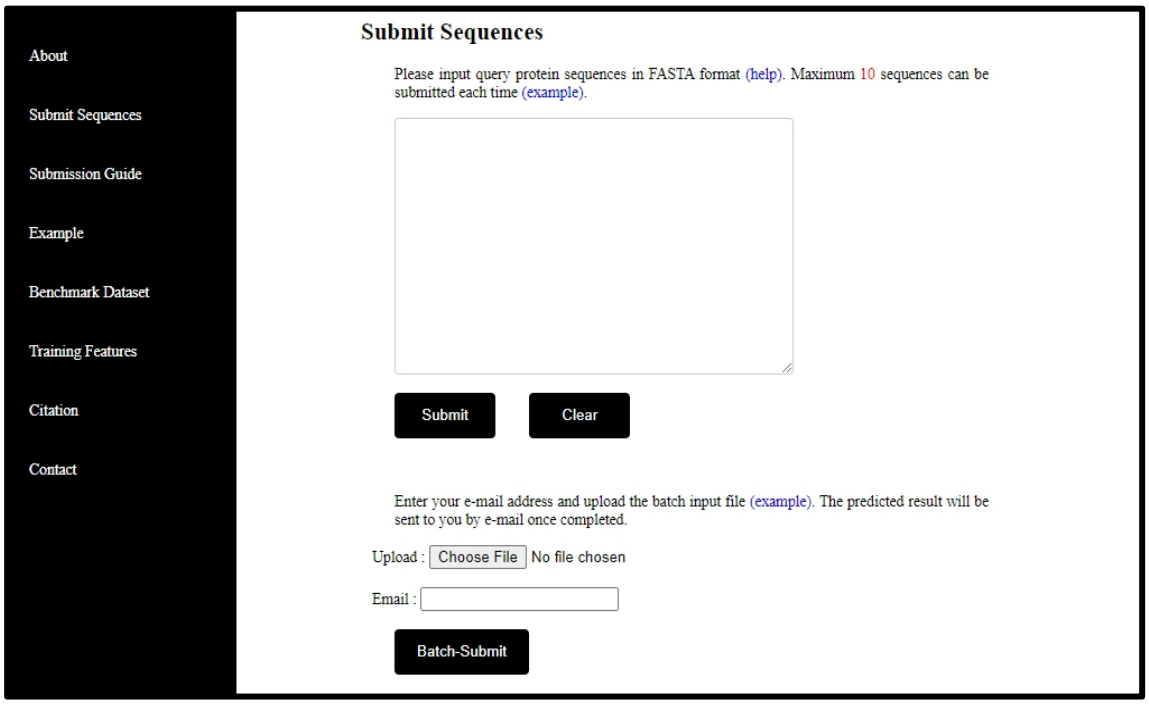


Fig. 5.1. Web Interface of the iML-LysPTM predictor.

5.6.1 Submission Guide

Users can submit their protein sequences in FASTA format. We currently accept a maximum of 10 sequences for the browser. If users have more than 10 sequences, they can also submit a batch file. The file must contain FASTA sequences. The FASTA format for the current predictor can be described as follows:

- ✓ Each query protein must begin with a greater-than (">") symbol.
- ✓ The identifier and description of the sequence might be placed after the ">" symbol (Optional).
- ✓ The sequence begins in a different line and ends when a ">" appears, which indicates the start of another query protein.

5.7 Conclusion

This chapter has covered the overall implementation procedure and corresponding performances obtained through rigorous experiments. Finally, it concluded with the construction of a robust and viable PTM predictor iML-LysPTM. Furthermore, we have brought up a user-oriented web-server for a multi-label PTM prediction system.

CHAPTER 6

Future Scopes and Conclusion

Introduction

Limitations and Future Scopes

Conclusion

6.1 Introduction

This study presents what is post-translational modifications, their impact on drug design, disease detection, and prevention with many other structural and functional diversities, why a computational approach for K-PTM study is required, and what it takes to develop a simultaneous K-PTM prediction system. We aim to improve our prediction performance compared to the existing multi-label predictors while including more functionality. We have built one after thorough discussion and analysis and deployed it with a web interface.

6.2 Limitations and Future Scopes

In this study, we have trained our SVM model with libSVM's default parameter for saving computational time and complexity. As we have considered several K-PTM sites at a time, there are a lot of scopes to optimize our model. In our future study, we will optimize the hyperparameters with the grid-search technique. Further, we would apply different classifiers to observe the prediction outcomes. iML-LysPTM was designed for five K-PTM types. Other PTM types with new protein sequences can be added to extend their capability in the future. Besides, a similar methodology of the proposed predictor can be used in the study of other PTMs such as C-PTM, R-PTM, and S-PTM that correspond to multi-label PTM sites at Cys, Arg, and Ser residues respectively.

6.3 Conclusion

For constructing a PTM predictor with higher efficacy, one of the main challenges is to construct a benchmark dataset and draw meaningful features out of it. Furthermore, for concurrent prediction of K-PTM sites, one has to consider the multi-label metrics for evaluating the output. We have built five benchmark datasets for each type of PTM site from the Univer Protein Resource. After that, we have utilized three different feature extraction techniques with the ANOVA F-Test statistic

feature selection technique. We have observed that the best result was picked by the top 600 features. We have also seen that among the three feature encoding methods, the probabilistic information has proved to be the most impactful compared to the binary encoding and amino acid factors feature sets. Additionally, we have minimized the imbalance between the positive and negative samples with different error costs (DEC) methods augmented with the support vector machine classifiers. After some rigorous analysis and performance comparison, we have developed our final predictor iML-LysPTM with five support vector machine prediction engines which outperformed all the existing state-of-art predictors. Finally, we have deployed our system as a user-oriented web-server to aid the researchers for further K-PTM study.

REFERENCES

[1] N. Saraswathy and P. Ramalingam, Concepts and techniques ingenomics and proteomics. Elsevier, 2011.

[2] G. McDowell and A. Philpott, “New insights into the role of ubiq-uitylation of proteins,” inInternational review of cell and molecularbiology. Elsevier, 2016, vol. 325, pp. 35–88.

[3] J. D. Weissman, A. Raval, and D. S. Singer, “Assay of an intrinsi cacetyltransferase activity of the transcriptional coactivator CIITA,”inMethods in enzymology. Elsevier, 2003, vol. 370, pp. 378–386.

[4] K.-C. Chou, “Impacts of bioinformatics to medicinal chemistry,”Medicinal chemistry, vol. 11, no. 3, pp. 218–234, 2015.

[5] W.-R. Qiu, B.-Q. Sun, X. Xiao, Z.-C. Xu, and K.-C. Chou, “iPTM-mLys: identifying multiple lysine PTM sites and their differenttypes,”Bioinformatics, vol. 32, no. 20, pp. 3116–3123, 2016.

[6] R. N. Freiman and R. Tjian, “Regulating the regulators: lysinemodifications make their mark,”Cell, vol. 112, no. 1, pp. 11–17,2003.

[7] Y. Xu and K.-C. Chou, “Recent progress in predicting posttrans-lational modification sites in proteins,”Current topics in medicinalchemistry, vol. 16, no. 6, pp. 591–603, 2016.

[8] J. Jia, Z. Liu, X. Xiao, B. Liu, and K.-C. Chou, “iSuc-PseOpt:identifying lysine succinylation sites in proteins by incorporatingsequence-coupling effects into pseudo components and optimiz-ing imbalanced training dataset,”Analytical biochemistry, vol. 497,pp. 48–56, 2016.

[9] A. Rahman, S. Ahmed, J. Rahman, and M. A. M. Hasan, “Predic-tion of formylation sites by incorporating sequence coupling intogeneral pseaac,” in2020 IEEE Region 10 Symposium (TENSYMP).IEEE, 2020, pp. 921–924.

[10] M. Wu, Y. Yang, H. Wang, and Y. Xu, “A deep learning methodto more accurately recall known lysine acetylation sites,”BMCbioinformatics, vol. 20, no. 1, p. 49, 2019.

[11] Z. Ju and J. - J. He, "Prediction of lysine glutarylation sites bymaximum relevance minimum redundancy feature selection,"Analytical biochemistry, vol. 550, pp. 1–7, 2018.

[12] M. A. Hasan, M. K. Ben Islam, J. Rahman, and S. Ahmad, "Cit-rullination Site Prediction by Incorporating Sequence CoupledEffects into PseAAC and Resolving Data Imbalance Issue,"CurrentBioinformatics, vol. 15, no. 3, pp. 235–245, 2020.

[13] M. A. M. Hasan and S. Ahmad, "mLysPTMpred: Multiple LysinePTM Site Prediction Using Combination of SVM with ResolvingData Imbalance Issue,"Natural Science, vol. 10, no. 9, pp. 370–384,2018.

[14] K.-C. Chou, "A vectorized sequence-coupling model for predict-ing HIV protease cleavage sites in proteins."Journal of BiologicalChemistry, vol. 268, no. 23, pp. 16 938–16 948, 1993.

[15] —— , "Prediction of human immunodeficiency virus proteasecleavage sites in proteins,"Analytical biochemistry, vol. 233, no. 1,pp. 1–14, 1996.

[16] W.-Z. Lin, J.-A. Fang, X. Xiao, and K.-C. Chou, "iDNA-Prot:identification of DNA binding proteins using random forest withgrey model,"PloS one, vol. 6, no. 9, 2011.

[17] J. Zhe and S. - Y. Wang, "Prediction of 2-hydroxyisobutyrylationsites by integrating multiple sequence features with ensemblesupport vector machine,"Computational Biology and Chemistry, p.107280, 2020.

[18] D. Chen, Z. Liu, X. Ma, and D. Hua, "Selecting genes by teststatistics,"BioMed Research International, vol. 2005, no. 2, pp. 132–138, 2005.

[19] Z. Ju and S.-Y. Wang, "Prediction of citrullination sites by incor-porating k-spaced amino acid pairs into Chou's general pseudoamino acid composition,"Gene, vol. 664, pp. 78–83, 2018.

[20] ——, "iLys-Khib: Identify lysine 2-Hydroxyisobutyrylation sitesusing mRMR feature selection and fuzzy SVM algorithm,"Chemo-metrics and Intelligent Laboratory Systems, vol. 191, pp. 96–102, 2019.

[21] K. Veropoulos, C. Campbell, N. Cristianiniet al., “Controllingthe sensitivity of support vector machines,” inProceedings of theinternational joint conference on AI, vol. 55, 1999, p. 60.

[22] M. A.M. Hasan, S. Ahmad, and M. K. I. Molla, “iMulti-HumPhos:a multi-label classifier for identifying human phosphorylatedproteins using multiple kernel learning based support vectormachines,”Molecular BioSystems, vol. 13, no. 8, pp. 1608–1618,2017.

[23] Z. Ju and J.-J. He, “Prediction of lysine propionylation sites usingbiased SVM and incorporating four different sequence featuresinto Chou’s PseAAC,”Journal of Molecular Graphics and Modelling,vol. 76, pp. 356–363, 2017.

[24] M. A. M. Hasan, J. Li, S. Ahmad, and M. K. I. Molla, “predCar-site: Carbonylation sites prediction in proteins using supportvector machine with resolving data imbalanced issue,”Analyticalbiochemistry, vol. 525, pp. 107–113, 2017.

[25] W. Bao, B. Yang, D.-S. Huang, D. Wang, Q. Liu, Y.-H. Chen, andR. Bao, “IMKPse: Identification of protein malonylation sites bythe key features into general PseAAC,”IEEE Access, vol. 7, pp.54 073–54 083, 2019.

[26] W.-R. Qiu, X. Xiao, W.-Z. Lin, and K.-C. Chou, “iMethyl-PseAAC:identification of protein methylation sites via a pseudo amino acidcomposition approach,”BioMed research international, vol. 2014,2014.

[27] U. Consortium, “UniProt: a worldwide hub of protein knowl-edge,”Nucleic acids research, vol. 47, no. D1, pp. D506–D515, 2019.

[28] K .- C. Chou, “Prediction of signal peptides using scaled window,”peptides, vol. 22, no. 12, pp. 1973–1979, 2001.

[29] ——, “Using amphiphilic pseudo amino acid composition to pre-dict enzyme subfamily classes,”Bioinformatics, vol. 21, no. 1, pp.10–19, 2005.

[30] ——, “Some remarks on protein attribute prediction and pseudoamino acid composition,”Journal of theoretical biology, vol. 273,no. 1, pp. 236–247, 2011.

[31] J.-L. Min, X. Xiao, and K.-C. Chou, “iEzy-Drug: A web server foridentifying the interaction between enzymes and drugs in cellularnetworking,”BioMed research international, vol. 2013, 2013.

[32] P. Du, X. Wang, C. Xu, and Y. Gao, “PseAAC-Builder: A cross-platform stand-alone program for generating various specialChou’s pseudo-amino acid compositions,”Analytical biochemistry,vol. 425, no. 2, pp. 117–119, 2012.

[33] Y. Xu, X. Wen, L.-S. Wen, L.-Y. Wu, N.-Y. Deng, and K.-C. Chou,“iNitro-Tyr: Prediction of nitrotyrosine sites in proteins with gen-eral pseudo amino acid composition,”PloS one, vol. 9, no. 8, p.e105018, 2014.

[34] Z. Zhang, M. Tan, Z. Xie, L. Dai, Y. Chen, and Y. Zhao, “Identifica-tion of lysine succinylation as a new post-translational modifica-tion,”Nature chemical biology, vol. 7, no. 1, p. 58, 2011.

[35] J. Jia, L. Zhang, Z. Liu, X. Xiao, and K.-C. Chou, “pSumo-CD: pre-dicting sumoylation sites in proteins with covariance discriminantalgorithm by incorporating sequence-coupled effects into generalPseAAC,”Bioinformatics, vol. 32, no. 20, pp. 3133–3141, 2016.

[36] W. R. Atchley, J. Zhao, A. D. Fernandes, and T. Dr uke, “Solvingthe protein sequence metric problem,”Proceedings of the NationalAcademy of Sciences, vol. 102, no. 18, pp. 6395–6400, 2005.

[37] A. Torkamani and N. J. Schork, “Accurate prediction of deleteriousprotein kinase polymorphisms,”Bioinformatics, vol. 23, no. 21, pp.2918–2925, 2007.

[38] M. H. Kutner, C. J. Nachtsheim, J. Neter, W. Liet al.,Applied linearstatistical models. McGraw-Hill Irwin New York, 2005, vol. 5.

[39] V. Vapnik, The nature of statistical learning theory. Springer science& business media, 2013.

[40] Z.Ju and S.-Y. Wang, “Prediction of lysine formylation sites usingthe composition of k-spaced amino acid pairs via Chou’s 5-stepsrule and general pseudo components,”Genomics, vol. 112, no. 1,pp. 859–866, 2020.

[41] C. Cortes and V. Vapnik, “Support - vector networks, ”Machinelearning, vol. 20, no. 3, pp. 273–297, 1995.

[42] X. Ruan, D. Zhou, R. Nie, and Y. Guo, “Predictions of ApoptosisProteins by Integrating Different Features Based on

ImprovingPseudo-Position-Specific Scoring Matrix,"BioMed Research Inter-national, vol. 2020, 2020.

[43] Y. Ma, Z. Yu, G. Han, J. Li, and V. Anh, "Identification of pre-microRNAs by characterizing their sequence order evolution in-formation and secondary structure graphs,"BMC bioinformatics,vol. 19, no. 19, p. 521, 2018.

[44] R. Batuwita and V. Palade, "Efficient resampling methods fortraining support vector machines with imbalanced datasets," inThe 2010 International Joint Conference on Neural Networks (IJCNN).IEEE, 2010, pp. 1–8.

[45] A. Dehzangi, Y. L opez, S. P. Lal, G. Taherzadeh, A. Sattar, T. Tsun-oda, and A. Sharma, "Improving succinylation prediction accu-racy by incorporating the secondary structure via helix, strand andcoil, and evolutionary information from profile bigrams,"PloS one,vol. 13, no. 2, p. e0191900, 2018.

[46] W. Qiu, C. Xu, X. Xiao, and D. Xu, "Computational Predictionof Ubiquitination Proteins Using Evolutionary Profiles and Func-tional Domain Annotation,"Current Genomics, vol. 20, no. 5, pp.389–399, 2019.

[47] C. Jia, M. Zhang, C. Fan, F. Li, and J. Song, "Formator: predictinglysine formylation sites based on the most distant undersam-pling and safe-level synthetic minority oversampling,"IEEE/ACMTransactions on Computational Biology and Bioinformatics, 2019.

[48] Q. Ning, Z. Ma, and X. Zhao, "dForml (KNN) - PseAAC : Detectingformylation sites from protein sequences using K-nearest neighboralgorithm via Chou's 5-step rule and pseudo components,"Journalof theoretical biology, vol. 470, pp. 43–49, 2019.

[49] Y. Xu, Y.-X. Ding, J. Ding, L.-Y. Wu, and N.-Y. Deng, "Phogly–PseAAC: prediction of lysine phosphoglycerylation in proteinsincorporating with position-specific propensity,"Journal of Theo-retical Biology, vol. 379, pp. 10–15, 2015.

[50] L . -M. Liu, Y. Xu, and K.-C. Chou, "iPGK-PseAAC: identify lysinephosphoglycerylation sites in proteins by incorporating four dif-ferent

tiers of amino acid pairwise coupling information into thegeneral PseAAC,"Medicinal Chemistry, vol. 13, no. 6, pp. 552–559,2017.

[51] A.Chandra, A.Sharma, A. Dehzangi, D. Shigemizu, and T. Tsun-oda, "Bigram-PGK: phosphoglycerylation prediction using thetechnique of bigram probabilities of position specific scoring ma-trix,"BMC molecular and cell biology, vol. 20, no. 2, pp. 1–9, 2019.

[52] J. Yu, S. Shi, F. Zhang, G. Chen, and M. Cao, "PredGly: predictinglysine glycation sites for Homo sapiens based on XGboost featureoptimization,"Bioinformatics, vol. 35, no. 16, pp. 2749–2756, 2019.

[53] V. B. Semwal, J. Singha, P. K. Sharma, A. Chauhan, and B. Behera,"An optimized feature selection technique based on incrementalfeature analysis for bio-metric gait data classification,"Multimediatools and applications, vol. 76, no. 22, pp. 24 457–24 475, 2017.

[54] C . -C. Chang and C .-J.Lin, "LIBSVM: A library for support vectormachines,"ACM Transactions on Intelligent Systems and Technology,vol. 2, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/~ cjlin/libsvm.

[55] M. Jiang and J.-Z. Cao, "Positive-Unlabeled learning for pupyla-tion sites prediction,"BioMed research international, vol. 2016, 2016.

[56] M. A. M. Hasan, S. Ahmad, and M. K. I. Molla, "Protein subcel-lular localization prediction using multiple kernel learning basedsupport vector machine,"Molecular BioSystems, vol. 13, no. 4, pp.785–795, 2017.

[57] K. - C. Chou, "Some remarks on predicting multi - label attributesin molecular biosystems,"Molecular Biosystems, vol. 9, no. 6, pp.1092–1100, 2013.

[58] X. Cheng, S.-G. Zhao, X. Xiao, and K.-C. Chou, "iATC-mISF:a multi-label classifier for predicting the classes of anatomicaltherapeutic chemicals,"Bioinformatics, vol. 33, no. 3, pp. 341–346,2017.

[59] Z. Ju and S. - Y. Wang , "Computational Identification of LysineGlutarylation Sites Using Positive-Unlabeled Learning,"CurrentGenomics, vol. 21, no. 3, pp. 204–211, 2020.

[60] B. - J. Chen, T. C. Lam, L. - Q. Liu, and C.-H. To, “Post-translational modifications and their applications in eye research,”Molecular medicine reports,vol. 15, no. 6, pp. 3923–3935, 2017.

[61] T. Scientific, “Overview of post-translational modifications,” 2018.

[62] V. Uversky, “Post translational modification,” inBrenner’s Encyclopedia of Genetics (Second Edition), second edition ed., S. Maloy and K. Hughes,Eds. San Diego: Academic Press, 2013, pp. 425 – 430.

[63] G. Duan and D. Walther, “The roles of post-translational modifications in the context of protein interaction networks,”PLoS Comput Biol, vol. 11, no. 2,p. e1004049, 2015.

[64] B. Macek, K. Forchhammer, J. Hardouin, E. Weber-Ban, C. Grangeasse, and I. Mijakovic, “Protein post-translational modifications in bacteria,”Naturereviews Microbiology, vol. 17, no. 11, pp. 651–664, 2019.

[65] D. G. Christensen, X. Xie, N. Basisty, J. Byrnes, S. McSweeney, B. Schilling, and A. J. Wolfe, “Post-translational protein acetylation: an elegantmechanism for bacteria to dynamically regulate metabolic functions,”Frontiers in microbiology, vol. 10, p. 1604, 2019.

[66] Y. Xu, J. Ding, L.-Y. Wu, and K.-C. Chou, “isno-pseaac: predict cysteine s-nitrosylation sites in proteins by incorporating position specific amino acidpropensity into pseudo amino acid composition,”PloS one, vol. 8, no. 2, p. e55844, 2013.

[67] C. T. Walsh, S. Garneau - Tsodikova, and G. J. Gatto Jr, “Protein posttranslational modifications: the chemistry of proteome diversifications,”AngewandteChemie International Edition, vol. 44, no. 45, pp. 7342–7372, 2005.

[68] E. S. Witze, W. M. Old, K. A. Resing, and N. G. Ahn, “Mapping protein post-translational modifications with mass spectrometry,”Nature methods,vol. 4, no. 10, pp. 798–806, 2007.

[69] H. Xu, Y. Wang, S. Lin, W. Deng, D. Peng, Q. Cui, and Y. Xue, “Ptmd: a database of human disease-associated post-translational modifications,”Genomics, proteomics & bioinformatics, vol. 16, no. 4, pp. 244–251, 2018.

[70] A. Meinhart and P. Cramer, "Recognition of RNA polymerase II carboxy-terminal domain by 3′-RNA-processing factors", Nature, vol. 430, no. 6996, pp. 223-226, 2004.

[71] Peptides , chem . libretexts .org [Online] , Available-https://chem.libretexts.org/Courses/Eastern_Mennonite_University/EMU%3A_Chemistry_for_the_Life_Sciences_(Cessna)/18, Accessed - 20th August, 2019.

[72] K. Starheim, K. Gevaert and T. Arnesen, "Protein N - terminal acetyltransferases: when the start matters", Trends in Biochemical Sciences, vol. 37, no. 4, pp. 152-161, 2012.

[73] Q. Ning, M. Yu, J. Ji, Z. Ma, and X. Zhao, "Analysis and prediction of human acetylation using a cascade classifier based on support vector machine,"BMC bioinformatics, vol. 20, no. 1, pp. 1–15, 2019.

[74] W. Zheng, Q. Wuyun, M. Cheng, G. Hu, and Y. Zhang, "two-level protein methylation prediction using structure model-based features,"Scientific reports,vol. 10, no. 1, pp. 1–15, 2020.

[75] L. Zhang, M. Liu, X. Qin, and G. Liu, "Succinylation site prediction based on protein sequences using the ifs-lightgbm (bo) model,"Computationaland Mathematical Methods in Medicine, vol. 2020, 2020.

[76] Z. Ju and J. -J. He, "Prediction of lysine crotonylation sites by incorporating the composition of k-spaced amino acid pairs into chou's general pseaac,"Journal of Molecular Graphics and Modelling, vol. 77, pp. 200–204, 2017.

[77] M. Arafat, M. Ahmad, S. Shovan, A. Dehzangi, S. R. Dipta, M. Hasan, A. Mehedi, G. Taherzadeh, S. Shatabda, A. Sharmaet al., "Accurately predictingglutarylation sites using sequential bi-peptide-based evolutionary features,"Genes, vol. 11, no. 9, p. 1023, 2020.

[78] M. Mann and O. N. Jensen, "Proteomic analysis of post-translational modifications,"Nature biotechnology, vol. 21, no. 3, pp. 255–261, 2003.

[79] O. N. Jensen, "Modification-specific proteomics: systematic strategies for analysing post-translationally modified proteins,"Trends in Biotechnology,vol. 18, pp. 36–42, 2000.

[80] K.-Y. Huang, H.-J. Kao, J. B.-K. Hsu, S.-L. Weng, and T.-Y. Lee, “Characterization and identification of lysine glutarylation based on intrinsicinterdependence between positions in the substrate sites,”BMC bioinformatics, vol. 19, no. 13, pp. 13–25, 2019.

[82] M. D. Hirschey and Y. Zhao, “Metabolic regulation by lysine malonylation, succinylation, and glutarylation,”Molecular & Cellular Proteomics, vol. 14,no. 9, pp. 2308–2315, 2015.

[83] M. Tan, C. Peng , K. A. Anderson, P. Chhoy, Z. Xie, L. Dai, J. Park, Y. Chen, H. Huang, Y. Zhanget al., “Lysine glutarylation is a protein posttranslationalmodification regulated by sirt5,”Cell metabolism, vol. 19, no. 4, pp. 605–617, 2014.

[84] M. Tan, H. Luo, S. Lee, F. Jin, J. S. Yang, E. Montellier, T. Buchou, Z. Cheng, S. Rousseaux, N. Rajagopalet al., “Identification of 67 histone marksand histone lysine crotonylation as a new type of histone modification,”Cell, vol. 146, no. 6, pp. 1016–1028, 2011.

[85] K. - Y. Huang, T.-Y. Lee, H.-J. Kao, C.-T. Ma, C.-C. Lee, T.-H. Lin, W.-C. Chang, and H.-D. Huang, “dbptm in 2019: exploring disease association andcross-talk of post-translational modifications,”Nucleic acids research, vol. 47, no. D1, pp. D298–D308, 2019.

[86] K .- Y. Huang, M.-G. Su, H.-J. Kao, Y.-C. Hsieh, J.-H. Jhong, K.-H. Cheng, H.-D. Huang, and T.-Y. Lee, “dbptm 2016: 10-year anniversary of a resourcefor post-translational modification of proteins,”Nucleic acids research, vol. 44, no. D1, pp. D435–D446, 2016.