A NEW STRUCTURAL FEATURE FOR LYSINE POST-TRANSLATION MODIFICATION PREDICTION USING MACHINE LEARNING

A Dissertation
Submitted to the Graduate Faculty
of the
North Dakota State University
of Agriculture and Applied Science

By

Yuan Liu

In Partial Fulfillment of the Requirements
for the Degree of
DOCTOR OF PHILOSOPHY

Major Department:
Computer Science

June 2021

Fargo, North Dakota

North Dakota State University

Graduate School

Title

A NEW STRUCTURAL FEATURE FOR LYSINE POST-TRANSLATION MODIFICATION PREDICTION USING MACHINE LEARNING

By

Yuan Liu

The Supervisory Committee certifies that this ***disquisition*** complies with North Dakota State University's regulations and meets the accepted standards for the degree of

DOCTOR OF PHILOSOPHY

SUPERVISORY COMMITTEE:

Dr. Changhui Yan

Chair

Dr. Anne Denton

Dr. Lu Liu

Dr. Megan Orr

Approved:

June 28, 2021

Date

Dr. Simone Ludwig

Department Chair

ABSTRACT

Lysine post-translational modification (PTM) plays a vital role in modulating multiple biological processes and functions. Lab-based lysine PTM identification is laborious and time-consuming, which impede large-scale screening. Many computational tools have been proposed to facilitate PTM identification in silico using sequence-based protein features. Protein structure is another crucial aspect of protein that should not be neglected. To our best knowledge, there is no structural feature dedicated to PTM identification. We proposed a novel spatial feature that captures rich structure information in a succinct form. The dimension of this feature is much lower than that of other sequence and structural features that were used in previous studies. When the proposed feature was used to predict lysine malonylation sites, it achieved performance comparable to other state-of-the-art methods that had much higher dimension. The low dimensionality of the proposed feature would be very helpful for building interpretable predictors for various applications involving protein structures. We further attempted to develop a reliable benchmark dataset and evaluate performance of multiple sequence- and structure-based features in prediction. The result indicated that our proposed spatial structure achieved competent performance and that other structural features can also make contribution to PTM prediction. Even though utilizing protein structure in lysine PTM prediction is still in the early stage, we can expect structure-based features to play a more crucial role in PTM site prediction.

ACKNOWLEDGEMENTS

Throughout the years of pursuing my Ph.D. degree, I have received a great deal of support and assistance from many people.

I would first like to thank my advisor, Professor Dr. Changhui Yan, whose expertise was invaluable in formulating the research questions and methodology. Dr. Yan's insightful feedback pushed me to sharpen my thinking and brought my work to a higher level. I would like to acknowledge my committee members, Dr. Anne Denton, Dr. Lu Liu, and Dr. Megan Orr, who all provided valuable guidance and support throughout my studies. In addition, I would like to thank Dr. Xuehui Li, my M.S. degree advisor in Department of Plant Sciences, who encouraged and supported my interdisciplinary careers.

I gratefully acknowledge the College of Graduate and Interdisciplinary Studies, whose award of the Doctoral Dissertation Fellowship allowed me to focus on the final stages of this work.

I would like to thank my parents, my wife, and my daughter for their patience and encouragement.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LITERATURE REVIEW

Living cell regulates cellular function and physiology mostly by its inventory of all proteins, i.e. proteome [1]. Proteome in eukaryotic cells consists of hundreds of thousand proteins across species, exceeding the coding capacity of its corresponding genome [1]. The mRNA splicing and protein post translation modification (PTM) are two major mechanisms to expand protein complexity over coding capacity [2]. PTMs are enzymatic, covalent chemical modifications after transcription, which change the protein's physical or chemical properties, activity, localization, stability [3], [4].

Currently, over three hundred types of PTMs have been identified [1]. The detection of PTM sites were mostly achieved through mass spectrometry-based techniques [1]. Briefly, modified proteins show lower or higher molecular mass compared to normal proteins. The observed modified protein is digested into peptides using enzymes. Each peptide is examined, and its molecular mass is compared to its expected molecular mass. Then, the peptide with unexpected mass is sequenced by tandem mass spectrometry [1]. Due to instability of some PTM types during mass spectrometry, other biotechnology methods, including Affinity Enrichment, Affinity Tagging, and Mass Tagging, are also used to identify PTMs [1].

Lysine malonylation (Kmal) is a recently identified PTM type [2]. The PTM type was validated using multiple biotechnology methods, Western blot, tandem MS, and high-performance liquid chromatograph. Du et al. [5] showed Kmal was associated with type 2 diabetes, and elevated Kmal sites were observed in liver tissues of mice models. The Kmal sites were involved in an enzyme of the glycolysis pathway. Ma et al. [6] validated the presence of Kmal sites in Cyanobacteria, and further bioinformatics analysis indicated Kmal sites participated metabolic enzyme activity of phosphoglycerate kinase, which played a role in

photosynthesis. In addition, number of Kmal sites were identified in common wheat, and these sites were involved in diverse pathways, including carbon metabolism, the Calvin cycle, and biosynthesis of amino acids [7].

The detailed cellular regulation mechanism of Kmal is based on identification of Kmals [8]. However, the laborious and low efficiency of PTM site identification via experimental method limits the scale of Kmal sites searching. With the advance of artificial intelligence and proteomics, many models utilizing characteristics of proteins have been proposed to predict PTM sites *in silico*. Even though models differ in many aspects, such as feature extraction and training algorithms, they all can be generalized into 'Chou's 5 step rule': benchmark dataset construction, protein sample representation, prediction algorithm, method validation, and model releasing [9].

Data Collection

With the increase of identified Kmal sites by experiments, databases, such as dbPTM [10], PLMD [11], and PTM-SD [12], have been established and collected a number of modified sites. The collected protein sequences are highly redundant, which means many sequences share sequence similarities. Directly using these datasets with extreme redundancy will introduce bias into prediction models [13]. For example, if highly similar sequences exist in both training and validation dataset, the model may just 'memorize' these sequences instead of generalizing characteristics. In addition, the sequences and structures submitted into the biological databases increase at an exponential speed, and removing redundancy can reduce computational complexity [13].

Feature Construction

Sequence-based feature

Sequence-based features are features directly derived from the whole protein sequence, or a segment of the protein sequence. The following are some sequence-based features that have been used to encode the original protein sequences.

The 1-gram and 2-gram are two features based on the idea of k-gram for proteins [14]. The k in k-gram defines the length of unique block when calculating frequency. For example, 1-gram extracts frequency of each single residue in the protein sequence, and 2-gram calculates the frequency of each possible pair of residues in the sequence. Given a total of 21 residue types (including 20 natural types and 1 non-natural type), the 1-gram and 2-gram features are 21- and 441-dimensional vectors, respectively.

The numerical representation converts original sequences into numerical sequences, and each residue type is represented by a unique number [15]. Therefore, a protein sequence with *n* residues is represented as a *n*-dimensional vector.

The binary encoding, also known as one-hot encoding, converts each residue into a 21 dimensional orthogonal binary vector [15]. Each orthogonal binary vector consists of one '1' and twenty '0'. For example, alanine (A) is alphabetically the first residue and is encoded as '10000000000000000000'. Then, a sequence of n residues is represented as a (n-1)*21 dimensional vector. The difference between numerical representation and binary encoding is how a single residue is represented, as a single value or a vector. Because residue type is nominal data, consecutive numbering does not always make sense. However, binary encoding can suffer from a high dimensional problem when the length of protein sequence increases.

The Two Sample Logo incorporates both positive and negative datasets and calculates frequency difference between positive and negative samples at the sequence position [16]. For example, at position *i*, a twenty-dimensional vector is produced to represent frequency differences of each residue type at this position. The method is a good visualization tool to present major compositions in the dataset.

Word embedding, techniques of natural language processing, has also been applied in sequence representation. For instance, Word2vec [17] is a successful word embedding application that combines both continuous bag-of-word and skip-gram models. It maps input words into dimensional spaces. The protein sequences can also be considered as a collection of words and be mapped into spaces. Inspired by that, Asgari and Mofrad proposed ProtVec, a dedicated word embedding tool for protein sequences [15].

Physicochemical property-based features

AAindex is a database collecting various physicochemical and biochemical properties of amino acids [18]. Selected properties of residues are concatenated into a single vector for each sample. Various properties were used in protein function site prediction, including but not limited to physicochemical index, hydrophobicity, polarity, polarizability, hydration potential, accessibility reduction ratio, net charge, molecular weight, PK-N, PK-C, melting point, optical rotation, entropy of formation, heat capacity, and absolute entropy. The method represents a sequence of N residues as a vector of (N)*(number of properties selected) dimension.

The 20 standard amino acids can be grouped into groups based on their own characteristics. For example, based on the property of their R group, amino acids were classified into six categories: hydrophobic – aliphatic (Ala, Lle, Leu, Met, and Val), hydrophobic – aromatic (Phe, Trp, and Tyr), polar neutral (Asn, Cys, Gln, Ser, and Thr), electrically charged –

acidic (Asp and Glu), electrically charged – basic (Arg, His, and Lys), and unique amino acids (Gly and Pro). EBGW is an encoding method using such information [19]. The method creates several dummy vectors to record the presence of the property at each site and slices each dummy vector into a couple of sectors. Then, mean value is calculated for each sector and output.

Evolutionary-derived features

Position-Specific Scoring Matrix (PSSM) is a pattern matrix derived from multiple sequence alignment [20]. The query protein sequence is searched against a sequence database. Higher weight value is assigned to a specific site if it is more conserved compared to other sites. Based on the PSSM, many derivatives were proposed by row transformation, column transformation, or mixture of transformations [21].

The KNN encoder, originating from natural language processing [22], produces local similarities/dissimilarities of a sample to comparison dataset (containing both positive and negative samples) [23]. The similarity between two protein sequences is the summation of site substitution values. The substitution values can be obtained from a substitution matrix, such as BLOSUM62 matrix [24]. The normalized similarity score is used to perform k-nearest neighbor analysis, and the percentage of positive neighbor is the final score of the sample.

Feature Normalization

Features extracted from methods discussed above have various ranges. For example, 1- and 2-gram generates features ranging from 0 to 1, whereas features from numerical representation range from 0 to 20. Distance-based algorithms suffer from features with various variation, but tree-based algorithms are not sensitive to scaling [25]. Popular normalization methods include linear scaling to unit range, linear scaling to unit variance, transformation to a

uniform random variable, rank normalization, and normalization after fitting distributions, and empirical experiments revealed performance improvement after data normalization [26].

Feature Selection

A comprehensive classifier may utilize features extracted from multiple methods. But these features may have thousands of dimensions, which leads to a problem called 'curse of dimensionality' [27].

For continuous variables, correlation analysis can detect linear dependencies between variable and target. Redundant variables showing extremely high correlation with others may be discarded. However, correlation analysis is limited to linear relationships and cannot handle non-linear relationships [27]. In addition, correlation-based feature selection requires a huge amount of computation because the number of pairs increases exponentially when the number of features increases. To make the work feasible, a genetic algorithm was implemented to perform a stochastic general search to find an optimal subset [28].

The single-variable classifier method is another brute-force way for feature selection. Each single feature is used to train a classifier to obtain the predictive power of the feature. The drawback of the single-variable classifier is its intensive computation and the difficulty to distinguish top ranking variables [27].

Gain ratio, a concept arising from information theory, describes how much information can be obtained from an attribute with respect to the class through evaluating entropy [29]. High entropy indicates that the feature is uniformly distributed over various classes, and low entropy means that the feature forms cluster/s and tends to give high predicting power.

In addition to selecting features for further analysis, dimension reduction is an alternative way to address the high dimensionality problem [27]. Matrix factorization, e.g. singular value

decomposition, can extract a set of principal components, thereby maximizing variance with lower dimensions. Another way of reconstructing features to reduce dimension is clustering. Variables can be clustered into several groups, and the centroid of the group can be used to represent the whole group [27].

Machine Learning Algorithms

A number of machine learning algorithms have been proposed and applied in bioinformatics and computational biology, such as Bayesian classifiers, logistic regression, discriminant analysis, classification trees, nearest neighbor, neural networks, support vector machine (SVM) [30]. Of these algorithms, SVM, random forest (RF), and neural networks are widely used in PTM prediction [8].

SVM aims to find a hyperplane in an N-dimensional space that can distinctly separate data points and maximize the margin [31]. The success of SVM mostly results from selecting the particular hyperplane with maximum margin, which maximizes the classifier's ability to predict the correct classification on future samples [32]. Because a real-world dataset is always not perfect and may contain errors, the soft margin of SVM handles such errors under predefined tolerance. It allows some misclassifications without moving the margin of the separating hyperplane. Another concept introduced into SVM is kernel function, which solves nonlinear relationship. The kernel function projects an original dataset from a low-dimensional space into a higher one, where different classes become linearly separable.

RF is a tree-based algorithm incorporating an ensemble learning technique [33]. RF aggregates a large number of decision trees and outputs the majority voting or the average, and the method reduces the variance compared to a single decision tree. The original RF randomly draws a number of samples from the original dataset and constructs a decision tree. The key

hyperparameter is the number of trees in bagging. In most empirical experiments, the performance of the model reaches a plateau when a few hundred of trees are constructed.

Another recently popularized tool is deep learning. Basic deep learning structures include an input layer, several nonlinear layers, and output layers. The current deep learning architectures can be categorized into four groups, deep neural networks (DNN), convolutional neural network (CNN), recurrent neural network (RNN), and emergent architectures [34]. DNN usually consists of an input layer, multiple hidden layers, and an output layer, the numerical values move along the architectures. Nodes in a middle layer receive weighted summation from nodes of a previous layer, and they are activated by a non-linear function, named the activation function, including sigmoid, hyperbolic tangent, rectified linear unit, etc. CNN is extended based on DNN by adding convolution and pooling layers. Convolution layers utilize filters (small weight matrices) and perform a convolution operation to catch patterns across input data. Pooling layers split received tensors into small regions and take maximum or average as output values. Usually, the convolution layer and pool layer are combined together several times, and connect to a couple of fully-connected layers to increase non-linear properties [34]. RNN shows good performance on sequential information, where input data are not independent. The architecture consists of extra hidden units where cyclic connection exists. Because the cyclic connection, gradient vanishing and gradient explosion hinder long context input. To relieve such problem, long short-term memory (LSTM) [35] and gradient recurrent unit (GRU) [36] were proposed to serve as memory cells determining ‘memorize’ or ‘forget’.

Current Progress on Kmal Prediction

Available tools for Kmal prediction

To our knowledge, several methods were proposed for Kmal prediction. Mal-Lys is the first proposed method using k-gram encoding and AAindex property feature. The method performed feature selection using a correlation-based method and made final prediction using SVM with radial basis function [37]. Wang et al. [23] also used SVM and built a new classifier, MaloPred. MaloPred incorporated 1-gram, binary encoding, EBGW, KNN and PSSM to construct original feature set, and then information gain was employed for feature selection. Zhang et al. [15] built three species-specific ensemble models, named kmal-sp. The original feature set of kmal-sp consists of 1-gram, 2-gram, quasi-sequence order, numerical encoding, binary encoding, Logo, EBGW, AAindex, KNN, PSSM, and S-FPSSM. The original feature set was normalized and selected by information gain. The final ensemble models were built based on RF, SVM, gradient boosting decision tree, K-nearest neighbor, and logistic regression. LEMP, a LSTM-based ensemble malonylation prediction, used enhanced amino acid content and employed word embedding, RNN, and RF [38]. MUscADEL is the acronym for Multiple Scalable Accurate Deep Learner for lysine PTMs, which also employed word embedding and RNN, similar to part of LEMP [8]. Sun et al. [39] proposed CNN-based method K_net, which used enhanced amino acid composition and EBGW features. Kmalo a newly proposed ensemble classifier, combined five classifiers [40]. Each classifier made use of one feature, including binary encoding, AAindex, PSSM, amino acid composition, or pseudo-amino acid composition, and RF, SVM, or CNN were implemented depending on the preliminary test performance.

Performance of available tools

The performance of these available tools was evaluated by some measures. Popular measures include accuracy, specificity, sensitivity, precision, F-score, Matthew's correlation coefficient (MCC), and the receiver operating characteristic carvers, area under curve (AUC). These metrics describe the model performance in terms of true positive, true negative, false positive, and false negative. All available tools achieved AUC values over 0.8 (Table 1). Currently, there is no global comparison among all tools but some pair-wise comparisons (Table 1 and 2). All proposed models were built based on a specific dataset, and some models did not achieve good performance when evaluated with other datasets. For example, Mal-Lys and MaloPred showed much lower AUC values on LEMP and MUscADEL's datasets, even higher than just random guessing. The accuracy of kmal-sp dropped from 0.833 to 0.597 when changing dataset into Kmalo's (Table 2). LEMP achieved commensurable performance to Kmalo in terms of accuracy because of sharing the same dataset in training. The predictive performance degrading over datasets indicates current models still require further improvements.

Table 1. Performance comparison by AUC within each study in *Homo sapiens*.

	Mal-Lys	MaloPred	kmal-sp	LEMP	MUscADEL	K_net	Kmalo
Mal-Lys	0.814[1]	NA[2]	NA	0.561[3]	0.529	NA	NA
MaloPred		0.871	0.874	0.656	0.756	NA	NA
kmal-sp			0.923	NA	NA	NA	NA
LEMP				0.827	NA	NA	NA
MUscADEL					0.834	NA	NA
K_net						0.800	NA
Kmalo							0.943

[1]Diagonal values are performance based on its own testing dataset.
[2]NA means no comparison or values were not listed in the context.
[3]The AUC values were derived by the dataset from the tool (column name).

Table 2. Performance comparison by accuracy within each study in *Homo sapiens*.

	Mal-Lys	MaloPred	kmal-sp	LEMP	MUscADEL	K_net	Kmalo
Mal-Lys	NA[1]	NA[2]	NA	NA[3]	NA	NA	NA
MaloPred		0.737	0.802	NA	NA	NA	NA
kmal-sp			0.833	NA	NA	NA	0.597
LEMP				NA	NA	NA	0.862
MUscADEL					0.807	NA	NA
K_net						NA	NA
Kmalo							0.866

[1]Diagonal values are performance based on its own testing dataset.
[2]NA means no comparison or values were not listed in the context.
[3]The AUC values were derived by the dataset from the tool (column name).

Accessing Structural Similarity

In addition to features directly or indirectly derived from sequence information, the protein structure is an alternative resource of information for functional prediction. The three-dimensional structure of proteins at near atomic level resolution implies their potential function and evolutionary evidence [41]. Functions of proteins can be predicted from structurally similar proteins because protein structure is even more conserved compared to protein sequence [42]. Distantly related proteins might show dramatic differences on sequences but function in similar ways, which challenges the sequence-based method and favors structurally-motivated approach. Multiple structural alignment is a widely-used tool for structure prediction, motif detection, analysis of evolutionary, and even classification [43]. So far, multiple structure alignment methods can be categorized into two groups, 'horizontal-first' and 'vertical-first' [44]. The 'horizontal-first' methods utilize pair-wise alignments and merge pairs into multiple alignment results. The 'vertical-first' methods begin with identifying similar fragment blocks among queried proteins and extend these blocks into multiple alignments. To compare protein structures, the first question to answer is how to represent an amino acid/residue. Generally, there are four types of representation, backbone atom (C-alpha), distance map-based method (C-

map), secondary structure, and amino acid type or structural alphabet [37]. Because the identified structure blocks are three-dimensional objects with internal motion, aligners differed at treating these blocks as rigid, flexible, or elastic object [45]. Another important topic in structure alignment is scoring function, and aligners make use of scoring function and maximize it for optimal solutions. Because of the different choices in protein structure representation, there are three groups of scoring function, three-, two-, and one-dimensional [44]. Three-dimensional scoring function measures the positional deviations of equivalent atoms of the whole or substructures. Two-dimensional scoring function describes the similarities of residue-residue interactions, such as contact maps, graphs, and distance matrices. One dimensional scoring function profiles amino acid type and backbone conformational state [45].

So far, protein multiple structural alignment is still an open challenge. The 'horizontal-first' approaches merge pairwise alignment result progressively and also accumulate errors step by step [44]. The 'vertical-first' approaches identify similar fragment blocks first among proteins, but the number of similar fragment blocks grow exponentially with respect to the number of proteins [44]. Thus, the 'vertical-first' approaches require intensive computation. In addition to the two aligner groups, the consensus method provides an alternative way for multiple alignment. Ilinkin et al. [46] proposed an algorithm that a consensus (pseudo) structure generating from a subset of queries is used for iteratively searching similar structures and updating consensus structures in a larger dataset. Limitations of this algorithm are the bias underneath the generation of initial consensus and error accumulation of pairwise comparison in the iterations [44].

References

[1] O. Nørregaard Jensen, "Modification-specific proteomics: characterization of post-translational modifications by mass spectrometry," *Current Opinion in Chemical Biology*, vol. 8, no. 1, pp. 33–41, Feb. 2004, doi: 10.1016/j.cbpa.2003.12.009.

[2] C. Peng *et al.*, "The First Identification of Lysine Malonylation Substrates and Its Regulatory Enzyme," *Molecular & Cellular Proteomics*, vol. 10, no. 12, Dec. 2011, doi: 10.1074/mcp.M111.012658.

[3] V. N. Uversky, "Posttranslational Modification," in *Brenner's Encyclopedia of Genetics (Second Edition)*, S. Maloy and K. Hughes, Eds. San Diego: Academic Press, 2013, pp. 425–430. doi: 10.1016/B978-0-12-374984-0.01203-1.

[4] A. R. Farley and A. J. Link, "Chapter 40 Identification and Quantification of Protein Posttranslational Modifications," in *Methods in Enzymology*, vol. 463, R. R. Burgess and M. P. Deutscher, Eds. Academic Press, 2009, pp. 725–763. doi: 10.1016/S0076-6879(09)63040-8.

[5] Y. Du *et al.*, "Lysine Malonylation Is Elevated in Type 2 Diabetic Mouse Models and Enriched in Metabolic Associated Proteins," *Molecular & Cellular Proteomics*, vol. 14, no. 1, pp. 227–236, Jan. 2015, doi: 10.1074/mcp.M114.041947.

[6] Y. Ma, M. Yang, X. Lin, X. Liu, H. Huang, and F. Ge, "Malonylome Analysis Reveals the Involvement of Lysine Malonylation in Metabolism and Photosynthesis in Cyanobacteria," *J. Proteome Res.*, vol. 16, no. 5, pp. 2030–2043, May 2017, doi: 10.1021/acs.jproteome.7b00017.

[7] J. Liu *et al.*, "Systematic analysis of the lysine malonylome in common wheat," *BMC Genomics*, vol. 19, no. 1, p. 209, Mar. 2018, doi: 10.1186/s12864-018-4535-y.

[8] Z. Chen *et al.*, "Large-scale comparative assessment of computational predictors for lysine post-translational modification sites," *Brief Bioinform*, vol. 20, no. 6, pp. 2267–2290, Nov. 2019, doi: 10.1093/bib/bby089.

[9] K.-C. Chou, "Artificial Intelligence (AI) Tools Constructed via the 5-Steps Rule for Predicting Post-Translational Modifications," *Trends in Artificial Intelligence*, vol. 3, no. 1, Art. no. 3, Aug. 2019, doi: Artificial Intelligence (AI) Tools Constructed via the 5-Steps Rule for Predicting Post-Translational Modifications.

[10] K.-Y. Huang *et al.*, "dbPTM in 2019: exploring disease association and cross-talk of post-translational modifications," *Nucleic Acids Res*, vol. 47, no. D1, pp. D298–D308, Jan. 2019, doi: 10.1093/nar/gky1074.

[11] H. Xu, J. Zhou, S. Lin, W. Deng, Y. Zhang, and Y. Xue, "PLMD: An updated data resource of protein lysine modifications," *Journal of Genetics and Genomics*, vol. 44, no. 5, pp. 243–250, May 2017, doi: 10.1016/j.jgg.2017.03.007.

[12] P. Craveur, J. Rebehmed, and A. G. de Brevern, "PTM-SD: a database of structurally resolved and annotated posttranslational modifications in proteins," *Database (Oxford)*, vol. 2014, May 2014, doi: 10.1093/database/bau041.

[13] K. Sikic and O. Carugo, "Protein sequence redundancy reduction: comparison of various method," *Bioinformation*, vol. 5, no. 6, pp. 234–239, Nov. 2010.

[14] H. Liu and L. Wong, "Data mining tools for biological sequences," *J. Bioinform. Comput. Biol.*, vol. 01, no. 01, pp. 139–167, Apr. 2003, doi: 10.1142/S0219720003000216.

[15] Y. Zhang *et al.*, "Computational analysis and prediction of lysine malonylation sites by exploiting informative features in an integrative machine-learning framework," *Briefings in Bioinformatics*, Aug. 2018, doi: 10.1093/bib/bby079.

[16] V. Vacic, L. M. Iakoucheva, and P. Radivojac, "Two Sample Logo: a graphical representation of the differences between two sets of sequence alignments," *Bioinformatics*, vol. 22, no. 12, pp. 1536–1537, Jun. 2006, doi: 10.1093/bioinformatics/btl151.

[17] X. Rong, "word2vec Parameter Learning Explained," *arXiv:1411.2738 [cs]*, Nov. 2014, Accessed: Apr. 10, 2019. [Online]. Available: http://arxiv.org/abs/1411.2738

[18] S. Kawashima, H. Ogata, and M. Kanehisa, "AAindex: Amino Acid Index Database.," *Nucleic Acids Res*, vol. 27, no. 1, pp. 368–369, Jan. 1999.

[19] Z.-H. Zhang, Z.-H. Wang, Z.-R. Zhang, and Y.-X. Wang, "A novel method for apoptosis protein subcellular localization prediction combining encoding based on grouped weight and support vector machine," *FEBS Letters*, vol. 580, no. 26, pp. 6169–6174, 2006, doi: 10.1016/j.febslet.2006.10.017.

[20] M. M. Gromiha, "Chapter 2 - Protein Sequence Analysis," in *Protein Bioinformatics*, M. M. Gromiha, Ed. Singapore: Academic Press, 2010, pp. 29–62. doi: 10.1016/B978-8-1312-2297-3.50002-3.

[21] J. Wang *et al.*, "POSSUM: a bioinformatics toolkit for generating numerical sequence feature descriptors based on PSSM profiles," *Bioinformatics*, vol. 33, no. 17, pp. 2756–2758, Sep. 2017, doi: 10.1093/bioinformatics/btx302.

[22] S. Tan, "An effective refinement strategy for KNN text classifier," *Expert Systems with Applications*, vol. 30, no. 2, pp. 290–298, Feb. 2006, doi: 10.1016/j.eswa.2005.07.019.

[23] L.-N. Wang, S.-P. Shi, H.-D. Xu, P.-P. Wen, and J.-D. Qiu, "Computational prediction of species-specific malonylation sites via enhanced characteristic strategy," *Bioinformatics*, p. btw755, Dec. 2016, doi: 10.1093/bioinformatics/btw755.

[24] S. R. Eddy, "Where did the BLOSUM62 alignment score matrix come from?," *Nature Biotechnology*, vol. 22, no. 8, Art. no. 8, Aug. 2004, doi: 10.1038/nbt0804-1035.

[25] Z. Zhang, "Understand Data Normalization in Machine Learning," *Medium*, Aug. 11, 2019. https://towardsdatascience.com/understand-data-normalization-in-machine-learning-8ff3062101f0 (accessed Sep. 22, 2020).

[26] S. Aksoy and R. M. Haralick, "Feature normalization and likelihood-based similarity measures for image retrieval," *Pattern Recognition Letters*, vol. 22, no. 5, pp. 563–582, Apr. 2001, doi: 10.1016/S0167-8655(00)00112-4.

[27] I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," *Journal of Machine Learning Research*, vol. 3, no. Mar, pp. 1157–1182, 2003.

[28] A. G. Karegowda, A. S. Manjunath, and M. A. Jayaram, "Comparative study of attribute selection using gain ratio and correlation based feature selection," *International Journal of Information Technology and Knowledge Management*, vol. 2, no. 2, pp. 271–277, 2010.

[29] H. Dağ, K. E. Sayin, I. Yenidoğan, S. Albayrak, and C. Acar, "Comparison of feature selection algorithms for medical data," in *2012 International Symposium on Innovations in Intelligent Systems and Applications*, Jul. 2012, pp. 1–5. doi: 10.1109/INISTA.2012.6247011.

[30] P. Larrañaga *et al.*, "Machine learning in bioinformatics," *Brief Bioinform*, vol. 7, no. 1, pp. 86–112, Mar. 2006, doi: 10.1093/bib/bbk007.

[31] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and Computing*, vol. 14, no. 3, pp. 199–222, Aug. 2004, doi: 10.1023/B:STCO.0000035301.49549.88.

[32] W. S. Noble, "What is a support vector machine?," *Nature Biotechnology*, vol. 24, no. 12, Art. no. 12, Dec. 2006, doi: 10.1038/nbt1206-1565.

[33] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[34] S. Min, B. Lee, and S. Yoon, "Deep learning in bioinformatics," *Briefings in Bioinformatics*, p. bbw068, Jul. 2016, doi: 10.1093/bib/bbw068.

[35] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: continual prediction with LSTM," pp. 850–855, Jan. 1999, doi: 10.1049/cp:19991218.

[36] K. Cho *et al.*, "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation," *arXiv:1406.1078 [cs, stat]*, Sep. 2014, Accessed: Sep. 22, 2020. [Online]. Available: http://arxiv.org/abs/1406.1078

[37] Y. Xu, Y.-X. Ding, J. Ding, L.-Y. Wu, and Y. Xue, "Mal-Lys: prediction of lysine malonylation sites in proteins integrated sequence-based features with mRMR feature selection," *Scientific Reports*, vol. 6, no. 1, Art. no. 1, Dec. 2016, doi: 10.1038/srep38318.

[38] Z. Chen, N. He, Y. Huang, W. T. Qin, X. Liu, and L. Li, "Integration of A Deep Learning Classifier with A Random Forest Approach for Predicting Malonylation Sites," *Genomics, Proteomics & Bioinformatics*, vol. 16, no. 6, pp. 451–459, Dec. 2018, doi: 10.1016/j.gpb.2018.08.004.

[39] J. Sun, Y. Cao, D. Wang, W. Bao, and Y. Chen, "K_net: Lysine Malonylation Sites Identification With Neural Network," *IEEE Access*, vol. 8, pp. 47304–47311, 2020, doi: 10.1109/ACCESS.2019.2961941.

[40] C.-R. Chung *et al.*, "Incorporating hybrid models into lysine malonylation sites prediction on mammalian and plant proteins," *Scientific Reports*, vol. 10, no. 1, Art. no. 1, Jun. 2020, doi: 10.1038/s41598-020-67384-w.

[41] P. Brown, W. Pullan, Y. Yang, and Y. Zhou, "Fast and accurate non-sequential protein structure alignment using a new asymmetric linear sum assignment heuristic," *Bioinformatics*, vol. 32, no. 3, pp. 370–377, Feb. 2016, doi: 10.1093/bioinformatics/btv580.

[42] C. Chothia and A. M. Lesk, "The relation between the divergence of sequence and structure in proteins," *EMBO J.*, vol. 5, no. 4, pp. 823–826, Apr. 1986.

[43] S. Wang, J. Peng, and J. Xu, "Alignment of distantly related protein structures: algorithm, bound and implications to homology modeling," *Bioinformatics*, vol. 27, no. 18, pp. 2537–2545, Sep. 2011, doi: 10.1093/bioinformatics/btr432.

[44] J. Ma and S. Wang, "Algorithms, Applications, and Challenges of Protein Structure Alignment," in *Advances in Protein Chemistry and Structural Biology*, vol. 94, Elsevier, 2014, pp. 121–175. doi: 10.1016/B978-0-12-800168-4.00005-6.

[45] H. Hasegawa and L. Holm, "Advances and pitfalls of protein structural alignment," *Current Opinion in Structural Biology*, vol. 19, no. 3, pp. 341–348, Jun. 2009, doi: 10.1016/j.sbi.2009.04.003.

[46] I. Ilinkin, J. Ye, and R. Janardan, "Multiple structure alignment and consensus identification for proteins," *BMC Bioinformatics*, vol. 11, no. 1, p. 71, Feb. 2010, doi: 10.1186/1471-2105-11-71.

PAPER 1: A NOVEL SPATIAL FEATURE FOR PREDICTING LYSINE MALONYLATION SITES USING MACHINE LEARNING

Abstract

Lysine malonylation is a recently identified post-translation modification type with known effects on type 2 diabetes. Several machine learning algorithms have been used to predict lysine malonylation sites using various protein features. We proposed a novel spatial feature that captures rich structure information in a succinct form. The dimension of this feature is much lower than that of other sequences and structural features that were used in previous studies. When the proposed feature was used to predict lysine malonylation sites, it achieved performance comparable to other state-of-the-art methods that had much higher dimension. The low dimensionality of the proposed feature would be very helpful for building interpretable predictors for various applications involving protein structures.

Introduction

Post-translation modifications (PTMs) are chemical alterations on protein structure, typically catalyzed by subtrate-specific enzymes [1]. Through PTM, one gene can produce diverse, complex, and heterogeneous gene products. Currently, over three hundred types of PTMs have been identified in protein structures. Among them, lysine malonylation (Kmal) is a recently identified type [2]. Kmal was shown to be present in both eukaryotic and prokaryotic cells. Du et al. [3] showed that Kmal was elevated in type 2 diabetic mouse models. In addition, Kmal plays a role in metabolism and photosynthesis in Cyanobacteria [4].

Due to the importance of Kmal, identifying new Kmal sites is crucial to understanding biological processes and advancing disease treatment. Due to the high cost and low efficiency of experimental validation procedures, *in silico* prediction is needed to guide the design of

experimental procedures. To date, several machine-learning-based prediction algorithms, MaloPred [5], Mal-Lys [6], and kmal-sp [7], have been proposed to predict Kmal sites [8]. MaloPred extracted features from the sequence-based features, physicochemical properties, and evolutionary-derived information. Then, feature selection and model learning were performed using support vector machine (SVM). Mal-Lys utilized sequence order information, position-specific amino acid propensity, and physicochemical properties to conduct prediction with SVM. Zhang et al. [7] constructed a comprehensive feature set, subset feature set, and merged ensemble models into the final model, kmal-sp.

As reviewed in Chen et al. [8], types of features that have been used in lysine PTM site prediction include sequence-derived features, predicted structural features, physicochemical properties, position-specific scoring matrices (PSSMs), and peptide similarity features. In addition to these features, the 3-dimensional (3D) structure of protein in the Protein Data Bank [9], is a rich source of spatial features that could be used for PTM site prediction. However, one main challenge is how to encode the 3D structural features into 1-dimensional (1D) vectors that can be efficiently processed by machine learning methods. Additionally, this encoding must be invariant to the rotation and transition of the coordinate system. To our knowledge, there is no model that utilizes spatial information in PTM site prediction. Herein, we proposed a spatial feature that captured the spatial environment of a point of interest in the protein structure. We used unsupervised clustering to select features for prediction model training. We demonstrated the efficiency of the proposed feature in the prediction of Kmal sites by comparing it with other state-of-the-art features.

Material and Methods

Data collection and pre-processing

The original dataset was collected in [5] with 9,760 experimentally validated malonylation sites from 3,433 Uniprot entries. These Uniprot entries were mapped to PDB chains via SIFTS [10], and a total of 6,254 unique and available pdb chains were derived. To reduce dataset redundancy, sequences with over 70% sequence identity were discarded using CD-HIT online server [11]. The resulting dataset comprises 692 PDB chains and 1,036 validated Kmal sites. Another 1,036 non-Kmal lysine sites were randomly selected from the chains to be used as negative samples.

Representation schemes for different features

We proposed a spatial feature and compared its performance with a few common features that have been widely used in previous studies.

The proposed spatial feature

The orientation of a residue was represented by a vector, which will be referred to as side chain vector, originating from its alpha carbon and ending at the mass center of its R group. A special case was Glycine whose R group has only a single hydrogen that doesn't have coordinates in the PDB structure. In that case, a pseudo-R group consisting of the carbon in the alpha-carboxyl group, N, and O in Glycine were created. For an amino acid of interest, each of its neighboring residues was represented using a triplet (t, d, θ), where t was the type of the adjacent residue, d was the distance between the mass center of the R group of the neighboring residue and that of the residue of interest, and θ was the dihedral angle between the side chain vectors of the two residues. Among them, d and θ described the neighboring residue's spatial proximity and orientation relative to the residue of interest. For each residue of interest, we

considered N spatially nearest neighbors, where N was a parameter to be explored. Therefore, the spatial environment of a residue of interest was described as a bag of triplets (t_i, d_i,θ_i), where *i* ranged from 1 to N.

We used the following procedure to select M recurring triplet features from the training set of positive examples and used them as features to encode examples. The guiding principle for this procedure was that features that are important for Kmal function will occur repeatedly in the positive data set, i.e., they are conservative in the positive data set. Let *P* be the number of positive examples in the training set, and (t_i^j, d_i^j,θ_i^j) be the triplet describing the i^{th} neighbor of the j^{th} positive example. For every triplet (t_i^j, d_i^j,θ_i^j) in the positive set, its conservation score was calculated by

$$C_i^j = \sum_{k=1,..,P \ \& \ k \neq j} \min_{q=1,..,N} \{ \sqrt[2]{(d_i^j - d_q^k)^2 + (\theta_i^j - \theta_q^k)^2} \ if \ t_i^j = t_q^k, \quad 100 \ if \ t_i^j \neq t_q^k \}$$

Basically, for every (t_i^j, d_i^j,θ_i^j) this formula iterated overall positive example other than *i* *(i.e. k≠ i)* and for every k, it found the minimum Euclidean distance between (t_i^j, d_i^j,θ_i^j) and all (t_q^k, d_q^k,θ_q^k) in the bag of triplets associated with k. Then, C_i^j was the sum of such minimum distances over all positive examples in the training set. If (t_i^j, d_i^j,θ_i^j) *and* (t_i^j, d_i^j,θ_i^j) don't have the same type *(i.e.,* $t_i^j \neq t_q^k$*)*, the distance between the two triplets was arbitrarily set to a large value, 100. So, if a triplet occurred in the feature bags of all positive examples, its conservation score would be 0. If a triplet occurs only in one positive example, its conservation score would be *(P-1)*100*. A lower conservation score indicated that the triplet was more conservative and therefore was more important for Kmal site prediction. We then sorted all triplets in the order of ascending conservation score.

Some of the triplets were similar to each other with minor variations in d and θ. Therefore, we clustered all triplets that had the same type, t, using the k-mean cluster, with k=3. So, each type of triplets was clustered into 3 bins. Triplets in the same bin were considered equivalent. Thus, each triplet was associated with one bin and each bin corresponding to a group of residues that had the same spatial proximity and orientation relative to the residue of interest. Therefore, we could treat each bin as a spatial feature for function prediction.

Using the order of the triplets, we picked the M most conservative bins as features to encode examples. Each example was encoded as a vector of M values of 0 or 1, indicating the absence or presence of the corresponding spatial features in the bag of triplets associated with the example.

PSSM

Position-specific scoring matrix (PSSM) represents the evolutionary information of each amino acid site. The PSSMs for the dataset were constructed by running PSI-BLAST [12] against the uniref50 database with three iterations and e-value at 0.0001.

FEATURE

Halperin et al. [13] proposed the FEATURE model that included a large number of physicochemical properties from several spherical shells centering at a point of interest on the protein structure. The FEATURE combines distance and other traditional features, such as solvent accessibility, hydrophobicity, etc.

Residue identity

One hot encoding was used to represent the identity of amino acids. The 20 types of amino acids were represented using a 20-dimensional binary vector.

Side chain property

Based on the property of their R group, amino acids were classified into six categories: hydrophobic – aliphatic (Ala, Lle, Leu, Met, and Val), hydrophobic – aromatic (Phe, Trp, and Tyr), polar neutral (Asn, Cys, Gln, Ser, and Thr), electrically charged – acidic (Asp and Glu), electrically charged – basic (Arg, His, and Lys), and unique amino acids (Gly and Pro). A 6-dimensional vector was used to represent the six types of side chains using one-hot coding.

AAindex

AAindex is a database collecting various physicochemical and biochemical properties of amino acids [14]. It has been used in Kmal prediction previously. As in previous publications [6], [7], fifteen numerical index values were extracted for each amino acids, including physicochemical index, hydrophobicity, polarity, polarizability, hydration potential, accessibility reduction ratio, net charge, molecular weight, PK-N, PK-C, melting point, optical rotation, entropy of formation, heat capacity, and absolute entropy.

Model training and evaluation

In this study, random forest was used to build prediction models to compare the performance of various coding schemes. Random forest [15] is a widely used machine learning algorithm, which has successful applications in PTM site prediction [7], [8]. Briefly, the algorithm is a bagging-type ensemble of several decision trees by bootstrapping samples, and the final decision is made based on voting. The number of trees is a vital hyperparameter of the algorithm, and 200 trees were used in this study.

To measure the performance of models that use different features, four evaluation measures were used, including accuracy (ACC), precision (PRE), Sensitivity (SEN), and area under receiver operating characteristic curve (AUC).

$$ACC = \frac{true\ postive + true\ negative}{all\ samples}$$

$$PRE = \frac{true\ postive}{true\ postive + false\ positive}$$

$$SEN = \frac{true\ postive}{true\ postive + false\ nagetive}$$

Ten-fold cross-validation was used to evaluate the performance of each model. The final measure was the arithmetic mean from the cross-validation test.

Results

We used the proposed feature and a few commonly used features to encode the input individually. Then, we used random forest to build prediction models for each encoding scheme. We compared their prediction performance using ten-fold cross-validation.

When the proposed spatial feature was used to encode the input, two hyperparameters, N and M, needed to be determined. N defines how many neighboring residues to be considered, and M defines how many features were selected for the input encoding. We varied N and M in the range from 6 to 24, the performance is shown in Table 3. When N = 18 and M = 24, the method achieves the best AUC (0.66) and the best ACC (0.60), while PRE (0.58), and SEN (0.62) are all very close to the best (bold italic font in Table 3) in the whole spectrum.

Table 3. Prediction Performance for the proposed spatial feature

N	M	Acc	Pre	Sen	AUC
6	6	0.54	0.57	0.23	0.55
	12	0.55	0.57	0.40	0.57
	18	0.59	0.58	0.52	0.61
	24	0.57	0.58	0.57	0.61
12	6	0.50	0.52	0.50	0.53
	12	0.54	0.54	0.57	0.59
	18	0.57	0.57	0.60	0.63
	24	0.60	0.59	0.62	0.65
18	6	0.52	0.53	0.53	0.57
	12	0.57	0.57	0.59	0.61
	18	0.57	0.57	0.60	0.63
	24	0.60	0.58	0.62	0.66
24	6	0.56	0.56	0.55	0.59
	12	0.57	0.58	0.58	0.60
	18	0.59	0.57	0.62	0.64
	24	0.59	0.58	0.63	0.63

In previous Kmal prediction studies [5], [7], protein sequences were truncated into 25-residue segments and predictions were made regarding whether the lysine residue at the center of the segment was a Kmal site. We followed the same procedure when PSSM, AAindex, residue identity, and side chain property were used respectively to encode the input. When FEATURE was used to encode input, properties of six spheres were extracted, and each sphere resulted in 80 numeric values. Table 4 compares the performance of these five encoding schemes with that of the proposed spatial feature. The best values for each measure are shown in bold italic font. Since AUC gives a balanced assessment over both positive and negative classes, we will use AUC as the primary measure. Among all the encoding schemes, the proposed spatial feature and the FEATURE achieved the best AUC (0.66). The proposed spatial feature also achieved the best ACC (0.60), while its PRE (0.58) was very close to the best.

Although the overall performance of the proposed spatial feature and that of the FEATURE are very similar, the proposed feature has a clear advantage of lower input dimension. The proposed method has an input dimension of 24, while the FEATURE has a dimension of 480 in the input. Other methods have input dimensions from 144 to 480, much higher than that of the proposed method. Lower input dimension offers many benefits including easier interpretation of the prediction model, easier identification of properties that are crucial for the prediction, and faster computing.

Both the proposed feature and FEATURE methods used properties derived from protein structure. In comparison, the PSSM, AAindex, Residue Identity, and Side Chain Property methods only used properties derived from protein sequence. The fact that the proposed method and the FEATURE achieved better performance than the others indicates that proper structural conformations are needed for the lysine malonylation and structural information is crucial for the prediction of Kmal sites.

Table 4. Prediction Performance for various encoding schemes

Feature	Dimension of input	ACC	PRE	SEN	AUC
PSSM	480	0.59	0.56	0.68	0.61
AAindex	360	0.58	0.57	0.71	0.63
Residue Identity	480	0.59	0.59	0.68	0.61
Side Chain Property	144	0.60	0.58	0.71	0.62
FEATURE	480	0.59	0.59	0.67	0.66
The proposed spatial feature	24	0.60	0.58	0.62	0.66

Discussion and Conclusion

The identification of Kmal sites is crucial for understanding the disease mechanism and metabolic process [2]. Many methods have been proposed to exploit various features to achieve better performance [8]. Most of these features are derived from protein sequences. However, the protein structure is the foundation of many protein functions. The availability of more and more protein structures provides opportunities to perform function prediction using spatial features. Herein, we propose a novel feature based on spatial proximity and relative orientation. The feature uses distance and angle to capture the proximity and relative spatial orientation between amino acids, providing a succinct description of the structural conformation.

When 3D structural information is used for functional prediction by machine learning methods, a main challenge is how to encode the 3D structural information into 1D vectors that the machine learning methods can efficiently process. All previous methods concatenated structural features into a vector based on their order on the protein sequence or their proximity on the structure. However, the sequential order or structural proximity doesn't accurately reflect how the features are distributed in the 3D space. Therefore, the same spatial feature could be placed in different positions of the vector for different examples. However, in the vector

presentation, the same vector position of different examples is supposed to describe the same feature of different examples. This dilemma presents a hurdle for predicting function using structural information. In the proposed method, spatial features are put into an unordered bag, and then important features are selected. Examples are encoded into vectors using a set of selected spatial features, ensuring that all examples are encoded with the same order of spatial features.

The results presented here demonstrate the efficacy of the proposed spatial feature. The proposed spatial feature also has the advantages of low dimensionality, which makes it preferable for various prediction tasks.

References

[1] O. Nørregaard Jensen, "Modification-specific proteomics: characterization of post-translational modifications by mass spectrometry," Current Opinion in Chemical Biology, vol. 8, no. 1, pp. 33–41, Feb. 2004, doi: 10.1016/j.cbpa.2003.12.009.

[2] C. Peng et al., "The First Identification of Lysine Malonylation Substrates and Its Regulatory Enzyme," Molecular & Cellular Proteomics, vol. 10, no. 12, Dec. 2011, doi: 10.1074/mcp.M111.012658.

[3] Y. Du et al., "Lysine Malonylation Is Elevated in Type 2 Diabetic Mouse Models and Enriched in Metabolic Associated Proteins," Molecular & Cellular Proteomics, vol. 14, no. 1, pp. 227–236, Jan. 2015, doi: 10.1074/mcp.M114.041947.

[4] Y. Ma, M. Yang, X. Lin, X. Liu, H. Huang, and F. Ge, "Malonylome Analysis Reveals the Involvement of Lysine Malonylation in Metabolism and Photosynthesis in Cyanobacteria," J. Proteome Res., vol. 16, no. 5, pp. 2030–2043, May 2017, doi: 10.1021/acs.jproteome.7b00017.

[5] L.-N. Wang, S.-P. Shi, H.-D. Xu, P.-P. Wen, and J.-D. Qiu, “Computational prediction of species-specific malonylation sites via enhanced characteristic strategy,” Bioinformatics, p. btw755, Dec. 2016, doi: 10.1093/bioinformatics/btw755.

[6] Y. Xu, Y.-X. Ding, J. Ding, L.-Y. Wu, and Y. Xue, “Mal-Lys: prediction of lysine malonylation sites in proteins integrated sequence-based features with mRMR feature selection,” Scientific Reports, vol. 6, no. 1, Art. no. 1, Dec. 2016, doi: 10.1038/srep38318.

[7] Y. Zhang et al., “Computational analysis and prediction of lysine malonylation sites by exploiting informative features in an integrative machine-learning framework,” Briefings in Bioinformatics, Aug. 2018, doi: 10.1093/bib/bby079.

[8] Z. Chen et al., “Large-scale comparative assessment of computational predictors for lysine post-translational modification sites,” Brief Bioinform, vol. 20, no. 6, pp. 2267–2290, Nov. 2019, doi: 10.1093/bib/bby089.

[9] H. M. Berman et al., “The Protein Data Bank,” Nucleic Acids Res, vol. 28, no. 1, pp. 235–242, Jan. 2000, doi: 10.1093/nar/28.1.235.

[10] J. M. Dana et al., “SIFTS: updated Structure Integration with Function, Taxonomy and Sequences resource allows 40-fold increase in coverage of structure-based annotations for proteins,” Nucleic Acids Res, vol. 47, no. D1, pp. D482–D489, Jan. 2019, doi: 10.1093/nar/gky1114.

[11] Y. Huang, B. Niu, Y. Gao, L. Fu, and W. Li, “CD-HIT Suite: a web server for clustering and comparing biological sequences,” Bioinformatics, vol. 26, no. 5, pp. 680–682, Mar. 2010, doi: 10.1093/bioinformatics/btq003.

[12] S. F. Altschul et al., “Gapped BLAST and PSI-BLAST: a new generation of protein database search programs,” Nucleic Acids Res., vol. 25, no. 17, pp. 3389–3402, Sep. 1997, doi: 10.1093/nar/25.17.3389.

[13] I. Halperin, D. S. Glazer, S. Wu, and R. B. Altman, “The FEATURE framework for protein function annotation: modeling new functions, improving performance, and extending to novel applications,” BMC Genomics, vol. 9, no. Suppl 2, p. S2, Sep. 2008, doi: 10.1186/1471-2164-9-S2-S2.

[14] S. Kawashima, H. Ogata, and M. Kanehisa, “AAindex: Amino Acid Index Database.,” Nucleic Acids Res, vol. 27, no. 1, pp. 368–369, Jan. 1999.

[15] L. Breiman, “Random forests,” Machine learning, vol. 45, no. 1, pp. 5–32, 2001.

[16] P. Li, G. Pok, K. S. Jung, H. S. Shon, and K. H. Ryu, “QSE: A new 3-D solvent exposure measure for the analysis of protein structure,” Proteomics, vol. 11, no. 19, pp. 3793–3801, Oct. 2011, doi: 10.1002/pmic.201100189.

PAPER 2: COMPREHENSIVE ASSESSMENT OF SEQUENCE- AND STRUCTURE-BASED FEATURES FOR LYSINE POST-TRANSLATIONAL MODIFICATION SITES PREDICTION

Abstract

Predicting lysine post-translation modifications (PTMs) is increasingly important due to its crucial role in multiple biological functions and processes. To date, a large number of computational tools have been developed to utilize a large volume of sequencing resources to predict one or several types of modification sites. These proposed tools have usually combined features derived from sequences, e.g., physicochemical properties, evolutionary profile, and predicted secondary structures. However, protein structure also has deep influences on exerting biological functions, and very few attempts have been made to employ real structure in PTM prediction. In this study, we leveraged the abundant protein structure resources in Protein Data Bank (PDB), extracted structural characteristics using the existing methods and our newly proposed spatial feature, and evaluated both sequence- and structure-based features in four types of lysine PTM sites prediction. Our recently proposed spatial feature achieved the highest MCC value in multiple datasets, and sequence-based features showed competitive performance in terms of accuracy, AUC, and F1 measures. Even though other structure-based features did not outperform sequence-based features, these features derived from structures are still considered informative in PTM site prediction. With the development of protein structure production, we can expect structure-based features will play a more vital role in PTM site prediction.

Introduction

The complicated biological processes supporting cellular growth, reproduction, and survival, are mainly mediated by protein molecules. PTM is a type of specific and selective

covalent processing that modulates molecular interactions, protein localization, and stability [1]. The coding capacity of the human genome is about 30,000 genes, and alternative splicing and PTM expand human proteomes to over one million types of proteins [2]. PTMs modify amino acid side chains leading to change in physicochemical properties. This change, in turn, significantly affects the protein's structural and functional diversity [3]. Lysine is one of 20 standard amino acids and a hotspot for enzymatic and chemical PTMs. In addition to being the most modified amino acid, lysine has the most comprehensive types of modification [4]. The identification of lysine modification has a long history of over seventy years, and new modification types are still being identified. Recently, for example, methylation [5] and glutarylation have been identified [6]. Of these modification types, acetylation, methylation, ubiquitination, and malonylation have been intensely investigated, and the functional importance of these modifications encourages the development of more and quicker identification methods.

Acetylation was first reported on histones more than half decade ago [7]. In Acetylation, an acetyl functional group is introduced into the lysine residue, and the process is mainly performed by lysine acetyltransferases in species ranging from bacteria to mammals. This modification event can change DNA-protein interactions, transcriptional activity, and protein stability, leading to cancer, neuro, and cardiovascular diseases [8].

Malonylation is a recently identified lysine modification type where a malonyl group is attached to lysine [9]. The modified sites were involved in an enzyme of the glycolysis pathway, and a typical mice experimental model of type 2 diabetes indicated the lysine malonylation was associated with type 2 diabetes [10]. The presence of lysine malonylation is found not only in mammals but also in plants. The malonylated lysine was found in multiple subcellular

compartments and associated with several pathways, including carbon metabolism and biosynthesis of amino acids in wheat [11].

The first lysine modification type discovered was methylation, dating back to 1959 [5]. In methylation, a methyl group is added to a substrate or substitutes a portion of a lysine. One function of lysine methylation is modulating chromatin-based transcriptional control and epigenetics. Lysine methylation in histone is associated with euchromatin/heterochromatin and transcriptional activation/repression, and the methylation process is enzymatically reversible. This dynamic modulation also cross-regulates with other modification events, such as acetylation and ubiquitination [12].

Ubiquitination was first reported in the 1970s [13]. Unlike acetylation and methylation, the ubiquitination process attaches a large molecule to proteins [3]. The large molecule can be either a single 76 amino acid polypeptide ubiquitin molecule, multiple ubiquitin molecules, or other ubiquitin-related ubiquitin structures. Ubiquitination is associated with protein activation and/or inactivation, protein localization, and protein-protein interaction and function as critical regulators in multiple cellular processes, such as transcription, DNA repair, signal transduction, and cell-cycle control [14].

Due to the importance of lysine PTMs, many efforts have been dedicated to identifying lysine PTM events in substrates. Mass spectrometry, mass tagging, affinity tagging, and affinity enrichment are some popular strategies and techniques used to identify PTMs [1]. However, these experimental methods are time-consuming and labor-intensive, which hinders large-scale analysis on proteomics. Alternatively, machine learning methods can generalize characteristics from the existing datasets and make high-throughput predictions on future samples. A number of tools have been developed to predict various types of lysine PTMs.

Hou et al. [15] combined amino acid physicochemical property, transition probability, and position-specific composition from amino acid sequences and built an acetylation classifier with logistic regression. Li et al. [16] proposed a classifier model, SSPKA, to predict species-specific acetylation sites. SSPKA consists of features selected from the sequence, predicted secondary structure, functional annotation (domain, binding site, etc.), and functional features (gene ontology, KEGG path, etc.). Wuyun et al. [17] utilized sequence-based, physicochemical and biochemical properties and predicted structural features, and developed an SVM-based tool, KA-predictor, to predict species-specific lysine acetylation sites.

Wang et al. [18] proposed the first online species-specific malonylation sites predictor, MaloPred, by combining sequence-based, physicochemical properties, and evolutionary-derived features. Taherzadeh et al. [19] employed sequence-based, evolutionary information, physicochemical properties, and predicted structural features and proposed a lysine malonylation classifier, SPRINT-Mal. In addition to traditional machine learning methods, Chen et al. [20] proposed LEMP that integrated recurrent neural networks with word embedding and random forest with sequence-based features to predict lysine malonylation sites.

Shao et al. [21] mined sequential characteristics using Bi-profile Bayes feature extraction methods and proposed an SVM-based classifier, BPB-PPMS, to computationally identify lysine methylation sites. Contrasting to species-specific prediction, Lee et al. [22] split the identification problem by the identity of modified protein, histone or non-histone. MethyK, a web server consisting of two models, was proposed using sequence-based and predicted features. Deng et al. [23] developed a multiple-function predictor of GPS-MSP that can predict lysine and arginine methylation sites, and the predictor relies on a group-based prediction system that scores proteins biochemical properties against positive and negative samples.

UbiPred is the first tool dedicated to computationally predicting lysine ubiquitylation sites, which extracted informative physicochemical properties and implemented an SVM model [24]. Chen et al. [25] utilized local sequence similarity, physicochemical property, and amino acid composition feature and implemented several species-specific SVM models known as UbiProber. Conventionally, a training dataset consists of a number of experimentally validated modification sites as positive data and a similar amount of or more non-validated sites as negative data. Wang et al. [26] screened non-validated samples and incrementally selected effective negative samples into the negative dataset. They further extracted physicochemical property features from their curated dataset and developed a classifier known as ESA-UbiSite dedicated to identifying ubiquitination sites in humans.

Computational identification of PTM sites can be generalized into five steps: construct a valid benchmark dataset; extract features or effectively represent collected dataset; train a predicting model using machine learning algorithm; assess performance of the model; deploy the model using a webserver or standalone program [27]. Amid these steps, feature design is an extremely crucial part to develop a robust predictor. Features employed in PTM sites prediction include primary sequence-derived features, predicted protein structural features, protein physicochemical properties, protein position-specific scoring matrices, peptide similarity features, and protein functional annotations [28]. These features are directly or indirectly derived from protein sequences. The other type of protein representation is structural characteristics that also modulate protein functions. However, modification site prediction progressively advanced through mining sequential information rather than structural features due to multiple reasons. First, most PTM sites were deposited into protein sequence database, e.g., UniProt [29], and data retrieval can be complete through processing one or several databases. Sequences in these

databases were well maintained with careful human-curating, but entries in Protein Data Bank (PDB) [30] often contain variants and missing information, which hinders structure file parsing. Second, the interaction between a prediction webserver or standalone program and users is straightforward, which requires query sequences only and may also allow users to define some hyperparameters.

Although combining structural characteristics into PTM sites prediction is not convenient, protein structure should not be neglected because functional information is shared by proteins with similar structures [31]. For example, most phosphorylation sites were identified at the surface of the protein with higher percentage solvent accessibility than these of non-modified sites [32]. Arginine and lysine methylation sites had a significantly lower convex hull of protein surfaces value, which indicates methylation sites were located at the first convex hull of the protein surfaces and prone to contact other proteins [33]. The evidence suggests structural features should be combined to facilitate identifying PTM sites. Besides, the entries collected in PDB are exponentially increasing every year, and protein structure predicted by computational methods provides an alternative way to obtain protein structure [34], indicating these aforementioned difficulties are being resolved. Even though structural resources are not as abundant and accessible as sequential information, we aimed to utilize current tools to exploit protein structures in PTM predictions.

Material and Methods

Dataset construction

To establish feature performance comparison, several benchmark datasets were constructed. Datasets from PTM-ssMP [35] were downloaded. The datasets consist of lysine acetylation, ubiquitination, methylation, and some modification sites integrated from other

databases, e.g., dbPTM [36]. Redundancy was removed using CH-HIT [37] to ensure that pairwise similarity among protein sequences was less than 40 percent. Originally protein sequences with PTM sites were collected from UniProt, and we used SIFTS [38] to map UniProt entries to PDB entries. SIFTS is a project initiated by European Bioinformatics Institute to provide residue level mappings between PDB structure and UniProt sequence. In addition to the three datasets, the lysine malonylation dataset, which was evaluated in our recent study [39], was also included in this study.

Feature extraction

In this study, two types of features, sequence- and structure-based features, were evaluated (Table 5). Sequence-based features included amino acid identity, amino acid side chain properties, AAindex, Position-Specific Scoring Matrices (PSSM), predicted secondary structures, and predicted disorder scores. A 25-residue sequence segment with central residue lysine (K) of each site was employed to extract sequence-based features. The other type of feature, structure-based feature, consists of Half sphere exposure (HSE), residue depth, DSSP, and an our recently proposed spatial structure.

Residue identities

The identity of a residue was encoded as a 20-dimensional feature vector, and a 25-AA segments (12 upstream and downstream AA) was represented using a 480-dimensional numerical vector.

Side chain properties

According to the characteristics of the side chain in each residue, residues can be categorized into six classes: hydrophobic – aliphatic (Ala, Lle, Leu, Met, and Val), hydrophobic – aromatic (Phe, Trp, and Tyr), polar neutral (Asn, Cys, Gln, Ser, and Thr), electrically charged –

acidic (Asp and Glu), electrically charged – basic (Arg, His, and Lys), and unique amino acids (Gly and Pro). Similar to residue identity, the 25-AA segment was retrieved, and a 144-dimensional numerical vector was generated to describe sequence segment side-chain properties.

PSSM

To incorporate evolutionary information of protein sequences, PSSM was utilized to indicate conservation scores at each position. PSSM was generated by running PSI-BLAST [40] against the UniProt *uniref50* database with three iterations and an *e*-value at 0.0001.

AAindex

AAindex [41] that a database archives various physicochemical properties of AAs was used to extract physicochemical properties of residues surrounding the modification sites. Fifteen numerical index values were extracted for each amino acid, including physicochemical index, hydrophobicity, polarity, polarizability, hydration potential, accessibility reduction ratio, net charge, molecular weight, PK-N, PK-C, melting point, optical rotation, the entropy of formation, heat capacity, and absolute entropy. Thus, a 360-dimensional numerical vector was generated by iterating the sequence segment.

Predicted protein structure

Protein structures can be predicted by mining sequence evolution profiles. SPIDER3-Single was employed to predict protein structure using sequence segment. SPIDER3-Single [42] is an updated structure predictor based on SPIDER3[43]. This structure predictor takes PSI-BLAST and HHBlits [44] sequence profiles together with seven physiochemical properties from the sequence and feeds into long short-term memory bidirectional recurrent neural networks. SPIDER3-Single produces solvent accessible surface area, three-state and eight-state secondary

structure, main-chain angles (backbone ϕ and ψ torsion angles and Cα-atom-based θ and τ angles), half-sphere exposure, and contact number.

Predicted disorder score

Another predicted structure feature adopted in this study is the intrinsic disorder value, which reflects the possibility of a protein to fold into a well-defined and rigid structure. We used SPOT-Disorder2 to predict protein intrinsic disorder score. SPOT-Disorder2 is a recently proposed deep learning-based disorder predictor [45]. The model includes evolutionary profiles consisting of PSSM and HHblits and sequentially feeds into inception paths, residual connections, Squeeze-and-Excitation, LSTM, and fully connected layer segments.

HSE

HSE is a two-dimensional measure of a residue's solvent exposure [46]. A sphere is defined around the interested residue at a given radius and a plane perpendicular to the $C_\alpha - C_\beta$ vector splits the sphere into two half spheres. The direction of the $C_\alpha - C_\beta$ vector determines up and down half-sphere, and the number of C_α atoms in each sphere reflect the residue's solvent exposure. Another variant of HSE is that instead of using $C_\alpha - C_\beta$ vector, the pseudo $C_\alpha - C_\beta$ vector based on three consecutive C_α atoms are adopted. In this study, we included structure features generated by both methods using the implementations in Bio.PDB package [47].

Residue depth

Another structural feature is residue depth that describes the average distance of the atoms of a residue from the solvent-accessible surface. The solvent-accessible surface was constructed by the program MSMS [48], and the distance value was calculated by Bio.PDB.

DSSP

DSSP [49] parses protein structure profiles and calculates geometrical features and solvent exposure. It yields a relative accessible solvent area, bond angle, torsion angle, and hydrogen bond energies.

Spatial feature

We recently proposed a novel spatial structure feature and validated it in lysine malonylation dataset [39]. Briefly, N neighbor residues surrounded by the interested residue were retrieved. A triplet (t, d, θ) consisting of type, distance, and angle was used to define the center residue and each of N neighbors. Similar triplet was searched in all other samples if have, and non-supervised clustering analysis (K-means clustering employed) categorized these triplet groups into several clusters within the same amino acid types. Clusters were ranked in terms of cluster consistency, and the final feature dataset is comprised of triplets from the first M clusters. Besides, we extended the spatial feature from a single angle to three angles as follows. The single angle is a dihedral angle between $C_{\alpha 1} - R_1$ and $C_{\alpha 1} - R_2$. The single dihedral angle from two vectors cannot absolutely define the position in 3D space, and thus we added two more dihedral angles, between $C_{\alpha 1} - R_1$ and $C_{\alpha 1} - C_{\alpha 2}$ and between $C_{\alpha 1} - R_1$ and $C_{\alpha 1} - C_{Carbonyl-2}$. With the distance and three angles, the spatial relationship between the center residue and another neighbor residue is unique. A program package implemented by Python is developed and available in GitHub (https://github.com/lyjspx/A-Novel-Protein-Structural-Feature). The package has a number of functionalities: Uniport ID and PDB ID mapping, Residue number and PDB number mapping in PDB entry, PSSM retrieval based on PDB ID, and spatial feature extraction. PSSM and spatial feature calculated prior will be stored in a Sqlite3 database for the

sake of efficient retrieval. Computing techniques, such as multiprocessing and vectorization were employed to accelerate computation progress.

Table 5. Summary of sequence-based and structures evaluated.

Feature source	Type	Name	Source
Sequence	Sequence information	Residue Identity	
	Biochemical	AAindex	
		Side chain property	
	Evolution-based	PSSM	
	Predicted structure	Secondary structure (8-state)	SPIDER3-Single
		Secondary structure (8-state) in probability	SPIDER3-Single
		Secondary structure (3-state)	SPIDER3-Single
		Secondary structure (3-state) in probability	SPIDER3-Single
		Accessible Surface Area	SPIDER3-Single
		Half Sphere Exposure (α −up and -down)	SPIDER3-Single
		Contact Number	SPIDER3-Single
		Backbone torsion angle $\phi\ and\ \psi$	SPIDER3-Single
		θ: angle between $C\alpha_{i-1} - C\alpha_i - C\alpha_{i+1}$	SPIDER3-Single
		τ: angle between $C\alpha_i - C\alpha_{i+1}$	SPIDER3-Single
		Intrinsic disorder	SPOT-Disorder-Single
Structure		Half-Sphere Exposure (up, down, and angle) based on the approximate $C\alpha - C\beta$	BioPython
		Half-Sphere Exposure (up, down, and angle) based on the real $C\alpha - C\beta$	BioPython

Table 5. Summary of sequence-based and structures evaluated (continued).

Feature source	Type	Name	Source
		Residue Depth (the average distance of the atoms of a residue from the solvent accessible surface)	MSMS and BioPython
		Secondary Structure	DSSP and BioPython
		Relative ASA	DSSP and BioPython
		Backbone torsion angle ϕ and ψ	DSSP and BioPython
		Hydrogen bonds energy (NH–>O_1 and O–>NH_1)	DSSP and BioPython
		Spatial feature	In house program

Model training

Most of the lysine PTM predictors were proposed based on some well-established machine algorithms [28]. Four machine learning algorithms were employed to evaluate feature performance in this study: random forest, support vector machine (SVM), gradient boost classifier, and K-nearest neighbor. Random forest is a popular machine learning algorithm in classification and regression [50]. The algorithm essentially is an ensembled decision tree with resampling, and we trained the random forest models with 500 trees in this study. SVM classifies samples by searching for an optimal hyperplane and can accommodate high-dimensional data through the use of kernel functions [51]. We used an SVM with the Gaussian radial basis kernel in this study. Gradient Boost Classifier is a decision tree-based algorithm with iterative optimization [52]. Another classification algorithm used in this study is K-nearest neighbor, which categorizes samples into two classes by nearest neighbors' voting. Given the limited size of PTM datasets we have, 5-fold cross-validation was used to evaluate the performance of models. All model training and evaluation were performed using the scikit-learn toolkit [53].

Evaluation metrics

Six evaluation measures were used to evaluate performance, including accuracy (ACC), sensitivity (SN), precision (PRE), the area under curves (AUC), F1 score (F1), and Matthew's correlation coefficient (MCC).

$$ACC = \frac{TP + TN}{TP + FP + TN + FN}$$

$$SN = \frac{TP}{TP + FN}$$

$$PRE = \frac{TP}{TP + FP}$$

$$F1 = 2 \times \frac{TP}{2TP + FP + FN}$$

$$MCC = \frac{(TP \times TN) - (FN \times FP)}{\sqrt{(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)}}$$

Results

Constructed datasets

Table 6 shows the four datasets we obtained from two sources, PTM-ssMP and Kmal-sp. Because most UniProt entries didn't have corresponding PDB entries, the number of mapped PDB entries was about one tenth of that of the original UniProt entries. If a UniProt entry was mapped to multiple PDB entries, a single PDB entry was randomly retained for the sake of controlling redundancy. The resulting datasets consist of from 96 to 5733 samples. Because PTM-ssMP provided negative samples in this data portal, we adopted the same mapping strategy on the negative samples. Kmal-sp did not include negative samples, and negative sites of the malonylation dataset were randomly picked within non-positive lysine sites of PDB entries.

Table 6. Summary information of constructed datasets

Modification type	PTM-ssMP (UniProt entries)	PDB entries	Source
Acetylation	9067/9067*	1532/858	PTM-ssMP
Methylation	544/544	56/40	PTM-ssMP
Ubiquitination	23243/23243	3691/2042	PTM-ssMP
Malonylation	9760/NA	1036/1036	Kmal-sp

*The two numbers represent the numbers of positive and negative samples in the dataset.

Sequence analysis

The occurrence frequencies of sequences at each position were analyzed by Two Sample Logo [54], and flanking sites indicating differential patterns with t-test ($P< 0.05$) were visualized (Figure 1). Due to the small size of the methylation dataset, the sequence pattern is extremely sparse, and single amino acid type enriched and depleted in several sites. For example, valine (V) enriched at positions 3 and 8 and depleted at position 10. For the rest three datasets, differences (enrichment and depletion) at each position between positive and negative samples

were less than ten percent. Arginine (R) enriched in multiple positions of malonylation datasets, whereas more depletion can be observed in acetylation and ubiquitination datasets. In the ubiquitination dataset, ubiquitinated lysine and flanking lysine residues were mutually exclusive, as the depletion of lysine at positions 9 – 17.

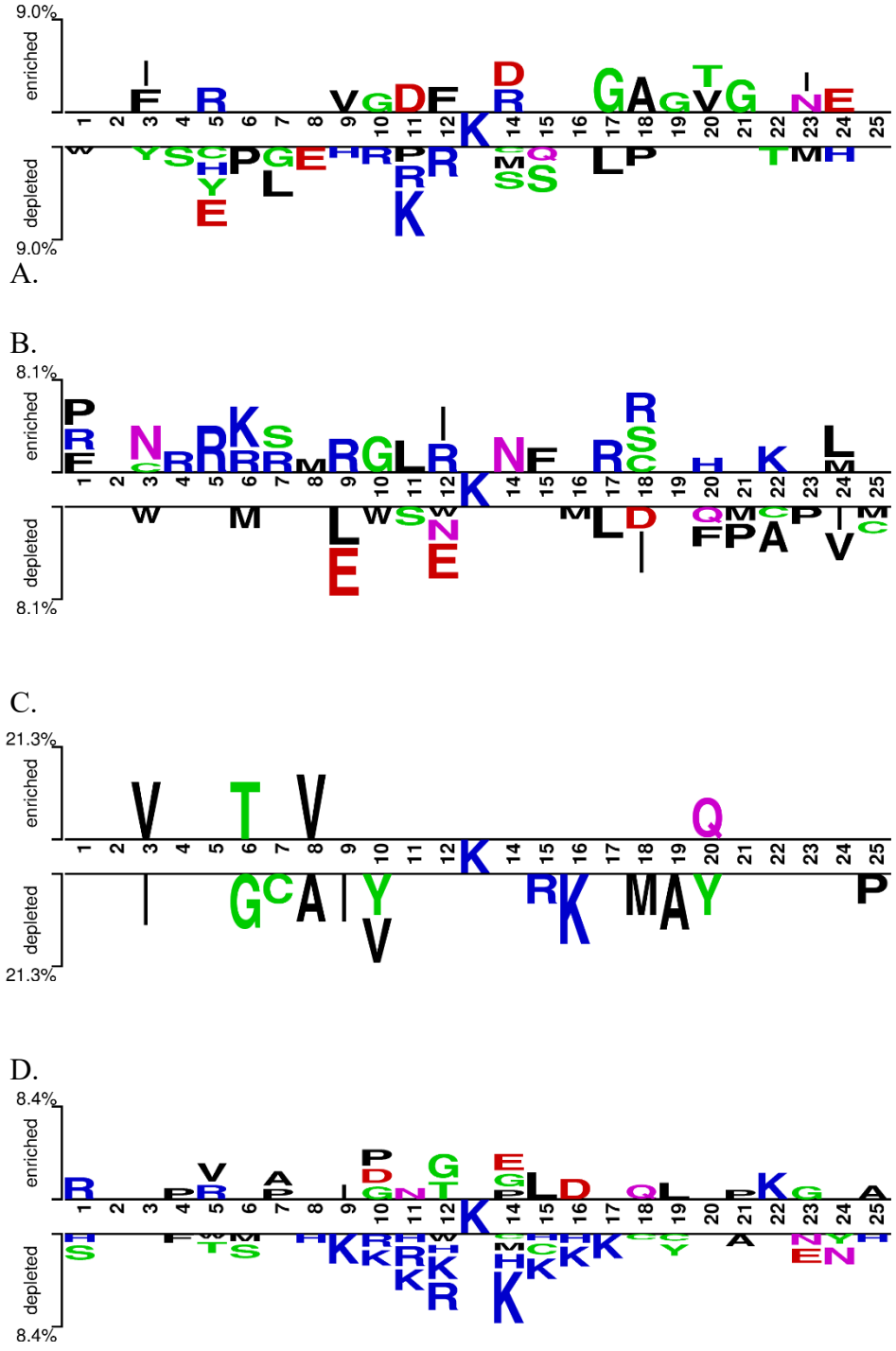

Figure 1. Sequence motif conservation analysis of lysine acetylation (A), malonylation (B), methylation (C) and ubiquitination (D) datasets.

Performance evaluation of different features

In our study, four state-of-the-art machine learning algorithms were employed to evaluate how different features contribute the prediction of PTM sites. The test results indicated random forest outperformed other algorithms in most cases. Therefore, we will only report the results obtained by random forest in Table 7, and the testing results of other methods were included in supplemental materials. For our proposed spatial feature, a large number of hyperparameter combinations were tested, and the top five combinations based on the MCC value achieved are shown in Table 7. Besides, the size of the methylation dataset is relatively small, which led to unstable prediction results because of potential sampling error in cross-validation. Thus, the feature performance comparisons are mainly focused on the other three datasets.

The performance of the same feature varied across datasets. We did not observe a single encoding scheme that outperformed others across all datasets. Our spatial features achieved the best performance in terms of MCC value but not in other comprehensive evaluation measures (AUC and F1). The spatial features achieved the highest MCC values with different hyperparameter (N, M, and K) combinations in different datasets. Another example, AAindex showed good performance in malonylation and ubiquitination site predictions, but the predictive power decreased in acetylation prediction.

In all datasets sequence-based feature encoding schemes performed better than all structure derived features except our spatial feature in terms of three comprehensive evaluation measures (AUC, F1, and MCC). PSSM, identity, side-chain property, and predicted structure features output by SPIDER3-Single each achieved best results in some of the datasets, and our spatial feature outperformed other features. Even though AAindex did achieve the best result in any of the datasets, it consistently achieved high-performance rank in all datasets.

Features derived from sequences have been widely proven to be effective in PTM prediction. Structure-based features can also be considered informative features. For instance, residue depth (RD) achieved 0.61, 0.57, and 0.15 for AUC, F1, and MCC, respectively. Two types of half-sphere exposure feature exhibited differences in all datasets, and a minor advantage was achieved by HSE-CA in malonylation and ubiquitination.

Table 7. The predictive performance of features by five-fold cross-validation on four datasets with random forest model.

	Acc	Pre	Sen	AUC	F1	MCC
Acetylation						
AAindex	0.647	0.650	0.980	0.574	0.783	0.050
HSE-CA	0.554	0.631	0.729	0.481	0.680	-0.026
HSE-CB	0.589	0.654	0.781	0.539	0.707	0.055
DSSP	0.591	0.641	0.811	0.498	0.713	0.014
Disorder	0.549	0.652	0.656	0.514	0.654	0.012
PSSM	**0.662**	**0.671**	0.985	0.563	**0.798**	-0.019
RD	0.564	0.638	0.743	0.494	0.689	-0.004
Identity	0.651	0.654	0.978	**0.586**	0.785	0.084
Side Chain Property	0.644	0.648	**0.983**	0.571	0.785	0.027
Spider3	0.649	0.652	0.976	0.545	0.783	0.054
Spatial_N6_M24_K2	0.630	0.662	0.885	0.542	0.754	**0.098**
Spatial_N6_M18_K2	0.632	0.661	0.867	0.547	0.747	0.094
Spatial_N12_M24_K3	0.635	0.651	0.925	0.537	0.767	0.090
Spatial_N12_M18_K2	0.633	0.649	0.926	0.560	0.766	0.072
Spatial_N6_M12_K2	0.615	0.656	0.837	0.526	0.732	0.070
Malonylation						
AAindex	**0.607**	0.594	**0.710**	0.649	0.639	0.214
HSE-CA	0.552	0.554	0.542	0.594	0.548	0.110
HSE-CB	0.532	0.527	0.523	0.549	0.532	0.059
DSSP	0.558	0.557	0.552	0.613	0.557	0.124
Disorder	0.576	0.581	0.568	0.589	0.576	0.156
PSSM	0.578	0.556	0.636	0.630	0.603	0.173
RD	0.570	0.575	0.566	0.612	0.568	0.151
Identity	0.587	0.581	0.660	0.633	0.618	0.193
Side Chain Property	0.606	0.590	0.705	0.637	**0.649**	0.201
Spider3	0.583	0.577	0.679	0.624	0.615	0.153
Spatial_N18_M24_K3	0.605	**0.622**	0.651	**0.657**	0.639	**0.217**
Spatial_N18_M18_K2	0.593	0.586	0.640	0.627	0.609	0.183
Spatial_N24_M12_K2	0.593	0.584	0.639	0.629	0.612	0.181
Spatial_N18_M24_K2	0.595	0.584	0.643	0.629	0.603	0.174
Spatial_N18_M24_K4	0.596	0.594	0.639	0.632	0.616	0.174

Table 7. The predictive performance of features by five-fold cross-validation on four datasets with random forest model (continued).

	Acc	Pre	Sen	AUC	F1	MCC
Methylation						
AAindex	0.595	0.618	0.809	0.512	0.691	0.043
HSE-CA	0.588	0.653	0.762	0.502	0.686	-0.034
HSE-CB	0.667	0.697	0.790	0.636	0.747	0.280
DSSP	0.601	0.682	0.859	0.286	0.765	-0.116
Disorder	0.553	0.629	0.580	0.571	0.608	0.095
PSSM	**0.701**	0.682	**0.940**	0.682	**0.797**	-0.039
RD	0.542	0.638	0.669	0.500	0.623	-0.049
Identity	0.615	0.625	0.879	0.515	0.744	0.085
Side Chain Property	0.541	0.593	0.826	0.468	0.681	0.004
Spider3	0.573	0.600	0.758	0.580	0.672	0.079
Spatial_3Angle_N24_M24_K4	0.661	0.727	0.740	**0.690**	0.743	**0.303**
Spatial_3Angle_N6_M12_K4	0.661	0.672	0.792	0.606	0.744	0.296
Spatial_3Angle_N18_M24_K4	0.629	**0.745**	0.676	0.637	0.717	0.282
Spatial_3Angle_N18_M6_K3	0.681	0.718	0.774	0.627	0.747	0.273
Spatial_3Angle_N24_M6_K3	0.650	0.712	0.819	0.638	0.743	0.272
Ubiquitination						
AAindex	0.653	0.657	0.979	0.594	0.788	0.106
HSE-CA	0.579	**0.665**	0.736	0.514	0.697	0.024
HSE-CB	0.615	0.661	0.859	0.511	0.746	-0.004
DSSP	0.551	0.648	0.672	0.494	0.659	-0.007
Disorder	0.577	0.632	0.792	0.512	0.707	0.002
PSSM	**0.657**	0.662	0.963	0.568	0.785	0.048
RD	0.589	0.661	0.778	0.501	0.713	0.002
Identity	0.649	0.651	0.982	**0.596**	0.784	0.050
Side Chain Property	0.653	0.653	**0.990**	0.585	0.787	0.073
Spider3	0.652	0.653	**0.990**	0.591	**0.788**	0.072
Spatial_N12_M24_K2	0.654	0.659	0.953	0.563	0.780	**0.126**
Spatial_N6_M24_K2	0.636	0.661	0.894	0.568	0.760	0.096
Spatial_N18_M24_K2	0.648	0.651	0.961	0.558	0.777	0.085
Spatial_N18_M18_K2	0.647	0.653	0.959	0.555	0.776	0.085
Spatial_N24_M24_K2	0.648	0.651	0.978	0.554	0.782	0.077

Spatial characteristics underneath samples

Our proposed spatial feature achieved the highest MCC value in all four PTM datasets, indicating its superior predictive power. To further elucidate the mechanism of this method, we used the spatial features learned in the malonylation dataset as an example to visualize the discoveries. Figure 2 shows the top four spatial features, namely LEU-1, LEU-2, ALA-1, and ALA-2, that our method extracted when using parameters N18M24K2 (Figure 2). Residues with feature LEU-1 has an average distance at 7.29 ranging from 2.86 to 25.81, and the angles are averaged at 137.36 ranging from 106.80 to 171.89. For a given feature, positive and negative samples only show minor difference in the values of angle that feature takes. Average angles of positive and negative samples and in LEU-1 are 138.53 and 136.17, respectively, and similar patterns were observed in the angles of the features. The angles between two features with the same amino acid type are apparently different. For example, the angle of LEU-1 averaging at 137.36 is much higher than that of LEU-2 (75.42). In addition to the angle difference between features, in our preliminary experiments, angle and distance were evaluated individually, and angle achieved higher performance than distance, which suggests the angle portion of the features contributed more predicting power than the distance.

Since positive and negative examples had similar values in the features, then how did the feature provide strong predicting power for PTM sites prediction? We analyzed the presence to absence ratio of a feature in the positive and negative samples separately, and found that there are significant difference between them (Figure 3). Take LEU-2 as an example: the presence to absence ratio of this feature is 1.517 in the negative set and 1.827 in the positive set. The difference is significant in Pearson's Chi-squared test ($P< 0.05$). Thus, this feature is enriched in the positive set. Figure 3 visualizes the presence to absence ratios of the top four features. Thus,

the predicting power of the proposed spatial features resides in their enrichment or depletion in the PTM sites.

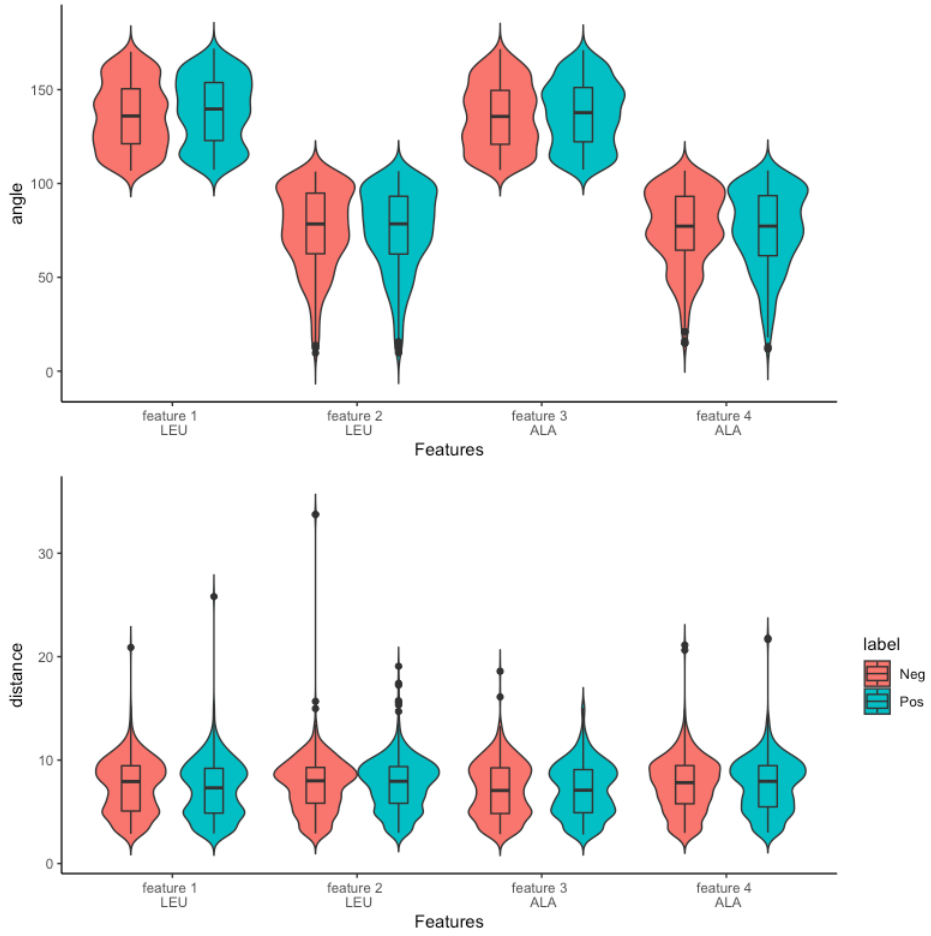

Figure 2. Angle and distance distribution of top four clusters of the proposed spatial feature (N18M24K2) in lysine malonylation dataset.

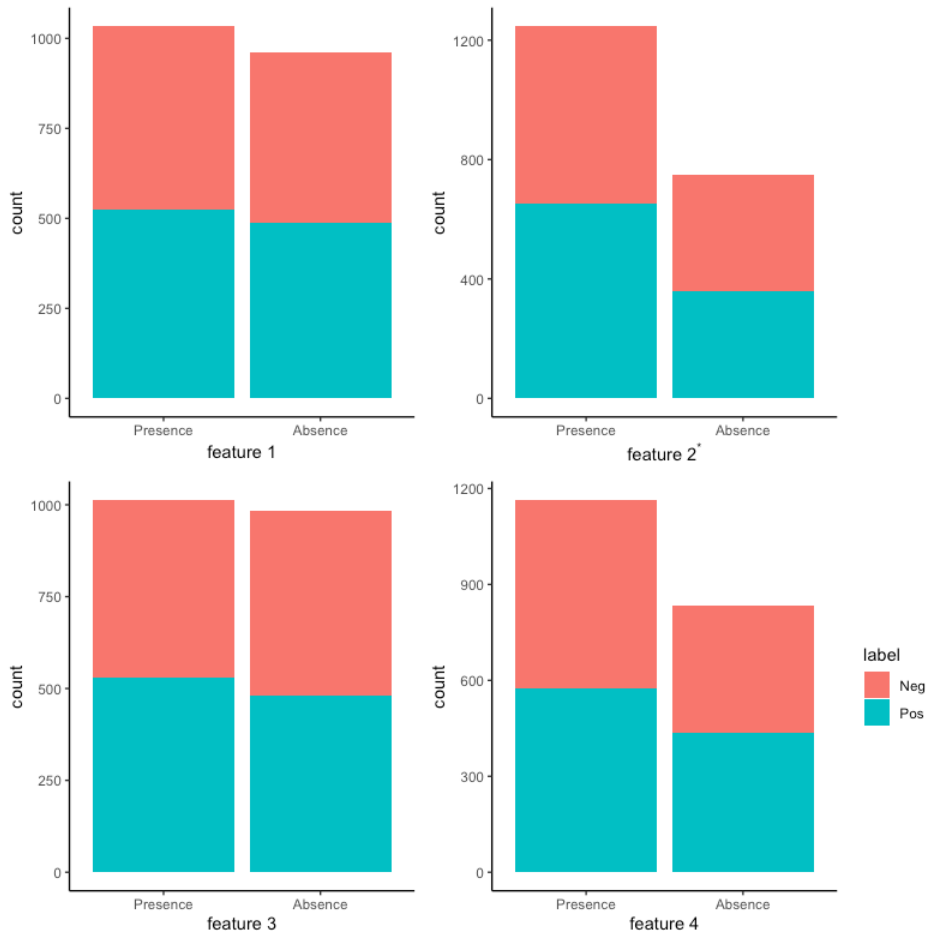

Figure 3. Presence and absence of neighbors in top four clusters of the proposed spatial feature (N18M24K2) in lysine malonylation dataset. * means statistical significance in Pearson's Chi-Square test of independence (P<0.05).

Discussion

Identifying PTM sites is crucial in understanding the cellular regulation mechanism. By 2019, 49 lysine PTM site computational prediction tools have been proposed [28]. These tools all exploited sequences containing PTM sites extracted features with/without feature selection and fed features into one or more machine learning algorithms. These tools mostly collected samples from one or more databases and employed state-of-the-art machine learning algorithms. Thus, among these procedures, designing new feature encoding schemes is critical to boosting predictor performance.

Zheng et al. [33] attempted to combine structural features and sequential features to predict lysine and arginine methylation sites, but their final model did not include structural features due to a lack of protein structures with annotated PTM sites. Currently, PTM-SD does provide the service of collecting structurally resolved and annotated PTM sites [55], but the size of their dataset is still too small to train predicting models. For example, as of 2021 only 100 lysine acetylation modification sites were collected, and, additionally, there is a high level of redundancy in the data set. We used SFITS to map protein sequences in UniProt to protein structures in PDB, which provided the best solution to obtain protein structures with PTM annotations. Since the majority of proteins in UniProt don't have corresponding structures in PDB, patterns observed in datasets derived from UniProt may be different than those observed in datasets derived from PDB. For instance, the malonylation dataset originally collected in kmal-sp showed lysine (K) and arginine (R) co-localized with modified lysine sites and Serine (S) and Glutamic (E) mutually exclusive to modified sites [56]. In our resulting malonylation dataset, the pattern is weak or diminished. Another dataset-related issue is the choice of negative samples. Using non-validated sites as the negative samples was critiqued, and refined negative samples increased ubiquitination accuracy by 0.15 in a previous study [26]. In structural feature prediction, how to select negative samples in UniProt or PDB entries is still an open question.

We have demonstrated that all real structure based-features were informative even though some of them were not competitive to sequence-based features (Table 7). However, the predicted structure, SPIDER3-Single, achieved good performance. This good performance probably results from the fact that the current structure prediction was implemented based on multiple sequence alignment or other derivatives [42], and thus evolutionary information was implicitly embedded in the features. Our proposed spatial feature achieved competitive performance compared to

other sequence-based features. The spatial feature essentially extracts clusters of neighbors sharing similar distance and angle to the center residue and aligns these clusters in order. We showed a significant difference in presence and absence within one cluster, but it should be pointed out that modification sites interact with many other neighbors simultaneously, an interaction that may not be captured in single cluster plots (Figure 3). In addition to a single angle, we also attempted to combine three angles into features to precisely depict neighbors' spatial position. The performance with three angles is not as good as a single angle, and the reason probably results from a high correlation (>0.95) between angles which may be reductant for models. So far, the application of structure-based features is limited due to sample size, artificial intelligence structure prediction (e.g., AlphaFold) will boost the size of the protein structure pool, and the usage of these features will be much expanded.

References

[1] O. Nørregaard Jensen, "Modification-specific proteomics: characterization of post-translational modifications by mass spectrometry," *Current Opinion in Chemical Biology*, vol. 8, no. 1, pp. 33–41, Feb. 2004, doi: 10.1016/j.cbpa.2003.12.009.

[2] C. T. Walsh, S. Garneau-Tsodikova, and G. J. Gatto, "Protein Posttranslational Modifications: The Chemistry of Proteome Diversifications," *Angewandte Chemie International Edition*, vol. 44, no. 45, pp. 7342–7372, 2005, doi: 10.1002/anie.200501023.

[3] C. Azevedo and A. Saiardi, "Why always lysine? The ongoing tale of one of the most modified amino acids," *Advances in Biological Regulation*, vol. 60, pp. 144–150, Jan. 2016, doi: 10.1016/j.jbior.2015.09.008.

[4] R. Bischoff and H. Schlüter, “Amino acids: Chemistry, functionality and selected non-enzymatic post-translational modifications,” *Journal of Proteomics*, vol. 75, no. 8, pp. 2275–2296, Apr. 2012, doi: 10.1016/j.jprot.2012.01.041.

[5] R. P. Ambler and M. W. Rees, “Epsilon-N-Methyl-lysine in bacterial flagellar protein,” *Nature*, vol. 184, pp. 56–57, Jul. 1959, doi: 10.1038/184056b0.

[6] M. Tan *et al.*, “Lysine glutarylation is a protein posttranslational modification regulated by SIRT5,” *Cell Metab*, vol. 19, no. 4, pp. 605–617, Apr. 2014, doi: 10.1016/j.cmet.2014.03.014.

[7] V. G. Allfrey, R. Faulkner, and A. E. Mirsky, “Acetylation and methylation of histones and their possible role in the regulation of rna synthesis*,” *Proc Natl Acad Sci U S A*, vol. 51, no. 5, pp. 786–794, May 1964.

[8] C. Choudhary, B. T. Weinert, Y. Nishida, E. Verdin, and M. Mann, “The growing landscape of lysine acetylation links metabolism and cell signalling,” *Nat Rev Mol Cell Biol*, vol. 15, no. 8, pp. 536–550, Aug. 2014, doi: 10.1038/nrm3841.

[9] C. Peng *et al.*, “The First Identification of Lysine Malonylation Substrates and Its Regulatory Enzyme,” *Molecular & Cellular Proteomics*, vol. 10, no. 12, Dec. 2011, doi: 10.1074/mcp.M111.012658.

[10] Y. Du *et al.*, “Lysine Malonylation Is Elevated in Type 2 Diabetic Mouse Models and Enriched in Metabolic Associated Proteins,” *Molecular & Cellular Proteomics*, vol. 14, no. 1, pp. 227–236, Jan. 2015, doi: 10.1074/mcp.M114.041947.

[11] J. Liu *et al.*, “Systematic analysis of the lysine malonylome in common wheat,” *BMC Genomics*, vol. 19, no. 1, p. 209, Mar. 2018, doi: 10.1186/s12864-018-4535-y.

[12] W. K. Paik, D. C. Paik, and S. Kim, “Historical review: the field of protein methylation,” *Trends in Biochemical Sciences*, vol. 32, no. 3, pp. 146–152, Mar. 2007, doi: 10.1016/j.tibs.2007.01.006.

[13] G. Goldstein, M. Scheid, U. Hammerling, D. H. Schlesinger, H. D. Niall, and E. A. Boyse, “Isolation of a polypeptide that has lymphocyte-differentiating properties and is probably represented universally in living cells.,” *Proc Natl Acad Sci U S A*, vol. 72, no. 1, pp. 11–15, Jan. 1975.

[14] O. Kerscher, R. Felberbaum, and M. Hochstrasser, “Modification of proteins by ubiquitin and ubiquitin-like proteins,” *Annu Rev Cell Dev Biol*, vol. 22, pp. 159–180, 2006, doi: 10.1146/annurev.cellbio.22.010605.093503.

[15] T. Hou *et al.*, “LAceP: Lysine Acetylation Site Prediction Using Logistic Regression Classifiers,” *PLOS ONE*, vol. 9, no. 2, p. e89575, Feb. 2014, doi: 10.1371/journal.pone.0089575.

[16] Y. Li *et al.*, “Accurate in silico identification of species-specific acetylation sites by integrating protein sequence-derived and functional features,” *Sci Rep*, vol. 4, no. 1, p. 5765, May 2015, doi: 10.1038/srep05765.

[17] Q. Wuyun, W. Zheng, Y. Zhang, J. Ruan, and G. Hu, “Improved Species-Specific Lysine Acetylation Site Prediction Based on a Large Variety of Features Set,” *PLOS ONE*, vol. 11, no. 5, p. e0155370, May 2016, doi: 10.1371/journal.pone.0155370.

[18] L.-N. Wang, S.-P. Shi, H.-D. Xu, P.-P. Wen, and J.-D. Qiu, “Computational prediction of species-specific malonylation sites via enhanced characteristic strategy,” *Bioinformatics*, p. btw755, Dec. 2016, doi: 10.1093/bioinformatics/btw755.

[19] G. Taherzadeh, Y. Yang, H. Xu, Y. Xue, A. W.-C. Liew, and Y. Zhou, "Predicting lysine-malonylation sites of proteins using sequence and predicted structural features," *J. Comput. Chem.*, vol. 39, no. 22, pp. 1757–1763, Aug. 2018, doi: 10.1002/jcc.25353.

[20] Z. Chen, N. He, Y. Huang, W. T. Qin, X. Liu, and L. Li, "Integration of A Deep Learning Classifier with A Random Forest Approach for Predicting Malonylation Sites," *Genomics, Proteomics & Bioinformatics*, vol. 16, no. 6, pp. 451–459, Dec. 2018, doi: 10.1016/j.gpb.2018.08.004.

[21] J. Shao, D. Xu, S.-N. Tsai, Y. Wang, and S.-M. Ngai, "Computational Identification of Protein Methylation Sites through Bi-Profile Bayes Feature Extraction," *PLOS ONE*, vol. 4, no. 3, p. e4920, Mar. 2009, doi: 10.1371/journal.pone.0004920.

[22] T.-Y. Lee, C.-W. Chang, C.-T. Lu, T.-H. Cheng, and T.-H. Chang, "Identification and characterization of lysine-methylated sites on histones and non-histone proteins," *Computational Biology and Chemistry*, vol. 50, pp. 11–18, Jun. 2014, doi: 10.1016/j.compbiolchem.2014.01.009.

[23] W. Deng, Y. Wang, L. Ma, Y. Zhang, S. Ullah, and Y. Xue, "Computational prediction of methylation types of covalently modified lysine and arginine residues in proteins," *Briefings in Bioinformatics*, vol. 18, no. 4, pp. 647–658, Jul. 2017, doi: 10.1093/bib/bbw041.

[24] C.-W. Tung and S.-Y. Ho, "Computational identification of ubiquitylation sites from protein sequences," *BMC Bioinformatics*, vol. 9, no. 1, p. 310, Jul. 2008, doi: 10.1186/1471-2105-9-310.

[25] X. Chen, J.-D. Qiu, S.-P. Shi, S.-B. Suo, S.-Y. Huang, and R.-P. Liang, "Incorporating key position and amino acid residue features to identify general and species-specific

Ubiquitin conjugation sites," *Bioinformatics*, vol. 29, no. 13, pp. 1614–1622, Jul. 2013, doi: 10.1093/bioinformatics/btt196.

[26] J.-R. Wang, W.-L. Huang, M.-J. Tsai, K.-T. Hsu, H.-L. Huang, and S.-Y. Ho, "ESA-UbiSite: accurate prediction of human ubiquitination sites by identifying a set of effective negatives," *Bioinformatics*, p. btw701, Jan. 2017, doi: 10.1093/bioinformatics/btw701.

[27] K.-C. Chou, "Artificial Intelligence (AI) Tools Constructed via the 5-Steps Rule for Predicting Post-Translational Modifications," *Trends in Artificial Intelligence*, vol. 3, no. 1, Art. no. 3, Aug. 2019, doi: Artificial Intelligence (AI) Tools Constructed via the 5-Steps Rule for Predicting Post-Translational Modifications.

[28] Z. Chen *et al.*, "Large-scale comparative assessment of computational predictors for lysine post-translational modification sites," *Brief Bioinform*, vol. 20, no. 6, pp. 2267–2290, Nov. 2019, doi: 10.1093/bib/bby089.

[29] The UniProt Consortium, "UniProt: a worldwide hub of protein knowledge," *Nucleic Acids Research*, vol. 47, no. D1, pp. D506–D515, Jan. 2019, doi: 10.1093/nar/gky1049.

[30] H. M. Berman *et al.*, "The Protein Data Bank," *Nucleic Acids Res*, vol. 28, no. 1, pp. 235–242, Jan. 2000, doi: 10.1093/nar/28.1.235.

[31] C. A. Orengo, A. E. Todd, and J. M. Thornton, "From protein structure to function," *Current Opinion in Structural Biology*, vol. 9, no. 3, pp. 374–382, Jun. 1999, doi: 10.1016/S0959-440X(99)80051-7.

[32] E. Vandermarliere and L. Martens, "Protein structure as a means to triage proposed PTM sites," *PROTEOMICS*, vol. 13, no. 6, pp. 1028–1035, 2013, doi: https://doi.org/10.1002/pmic.201200232.

[33] 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, Art. no. 1, Apr. 2020, doi: 10.1038/s41598-020-62883-2.

[34] M. AlQuraishi, “AlphaFold at CASP13,” *Bioinformatics*, vol. 35, no. 22, pp. 4862–4865, Nov. 2019, doi: 10.1093/bioinformatics/btz422.

[35] Y. Liu, M. Wang, J. Xi, F. Luo, and A. Li, “PTM-ssMP: A Web Server for Predicting Different Types of Post-translational Modification Sites Using Novel Site-specific Modification Profile,” *Int. J. Biol. Sci.*, vol. 14, no. 8, pp. 946–956, 2018, doi: 10.7150/ijbs.24121.

[36] K.-Y. Huang *et al.*, “dbPTM in 2019: exploring disease association and cross-talk of post-translational modifications,” *Nucleic Acids Res*, vol. 47, no. D1, pp. D298–D308, Jan. 2019, doi: 10.1093/nar/gky1074.

[37] Y. Huang, B. Niu, Y. Gao, L. Fu, and W. Li, “CD-HIT Suite: a web server for clustering and comparing biological sequences,” *Bioinformatics*, vol. 26, no. 5, pp. 680–682, Mar. 2010, doi: 10.1093/bioinformatics/btq003.

[38] J. M. Dana *et al.*, “SIFTS: updated Structure Integration with Function, Taxonomy and Sequences resource allows 40-fold increase in coverage of structure-based annotations for proteins,” *Nucleic Acids Res*, vol. 47, no. D1, pp. D482–D489, Jan. 2019, doi: 10.1093/nar/gky1114.

[39] Y. Liu and C. Yan, “A Novel Spatial Feature For Predicting Lysine Malonylation Sites Using Machine Learning,” in *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Dec. 2020, pp. 76–79. doi: 10.1109/BIBM49941.2020.9313184.

[40] S. F. Altschul *et al.*, “Gapped BLAST and PSI-BLAST: a new generation of protein database search programs,” *Nucleic Acids Res.*, vol. 25, no. 17, pp. 3389–3402, Sep. 1997, doi: 10.1093/nar/25.17.3389.

[41] S. Kawashima, H. Ogata, and M. Kanehisa, “AAindex: Amino Acid Index Database.,” *Nucleic Acids Res*, vol. 27, no. 1, pp. 368–369, Jan. 1999.

[42] R. Heffernan, K. Paliwal, J. Lyons, J. Singh, Y. Yang, and Y. Zhou, “Single-sequence-based prediction of protein secondary structures and solvent accessibility by deep whole-sequence learning,” *Journal of Computational Chemistry*, vol. 39, no. 26, pp. 2210–2216, 2018, doi: https://doi.org/10.1002/jcc.25534.

[43] R. Heffernan, Y. Yang, K. Paliwal, and Y. Zhou, “Capturing non-local interactions by long short-term memory bidirectional recurrent neural networks for improving prediction of protein secondary structure, backbone angles, contact numbers and solvent accessibility,” *Bioinformatics*, vol. 33, no. 18, pp. 2842–2849, Sep. 2017, doi: 10.1093/bioinformatics/btx218.

[44] M. Remmert, A. Biegert, A. Hauser, and J. Söding, “HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment,” *Nature Methods*, vol. 9, no. 2, Art. no. 2, Feb. 2012, doi: 10.1038/nmeth.1818.

[45] J. Hanson, K. K. Paliwal, T. Litfin, and Y. Zhou, “SPOT-Disorder2: Improved Protein Intrinsic Disorder Prediction by Ensembled Deep Learning,” *Genomics, Proteomics & Bioinformatics*, vol. 17, no. 6, pp. 645–656, Dec. 2019, doi: 10.1016/j.gpb.2019.01.004.

[46] T. Hamelryck, “An amino acid has two sides: A new 2D measure provides a different view of solvent exposure,” *Proteins*, vol. 59, no. 1, pp. 38–48, Feb. 2005, doi: 10.1002/prot.20379.

[47] T. Hamelryck and B. Manderick, “PDB file parser and structure class implemented in Python,” *Bioinformatics*, vol. 19, no. 17, pp. 2308–2310, Nov. 2003, doi: 10.1093/bioinformatics/btg299.

[48] M. F. Sanner, A. J. Olson, and J. C. Spehner, “Reduced surface: an efficient way to compute molecular surfaces,” *Biopolymers*, vol. 38, no. 3, pp. 305–320, Mar. 1996, doi: 10.1002/(SICI)1097-0282(199603)38:3%3C305::AID-BIP4%3E3.0.CO;2-Y.

[49] W. G. Touw *et al.*, “A series of PDB-related databanks for everyday needs,” *Nucleic Acids Research*, vol. 43, no. D1, pp. D364–D368, Jan. 2015, doi: 10.1093/nar/gku1028.

[50] L. Breiman, “Random forests,” *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[51] A. J. Smola and B. Schölkopf, “A tutorial on support vector regression,” *Statistics and Computing*, vol. 14, no. 3, pp. 199–222, Aug. 2004, doi: 10.1023/B:STCO.0000035301.49549.88.

[52] J. H. Friedman, “Greedy function approximation: A gradient boosting machine.,” *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, Oct. 2001, doi: 10.1214/aos/1013203451.

[53] F. Pedregosa *et al.*, “Scikit-learn: Machine Learning in Python,” *Journal of Machine Learning Research*, vol. 12, no. 85, pp. 2825–2830, 2011.

[54] V. Vacic, L. M. Iakoucheva, and P. Radivojac, “Two Sample Logo: a graphical representation of the differences between two sets of sequence alignments,” *Bioinformatics*, vol. 22, no. 12, pp. 1536–1537, Jun. 2006, doi: 10.1093/bioinformatics/btl151.

[55] P. Craveur, J. Rebehmed, and A. G. de Brevern, “PTM-SD: a database of structurally resolved and annotated posttranslational modifications in proteins,” *Database (Oxford)*, vol. 2014, May 2014, doi: 10.1093/database/bau041.

[56] Y. Zhang *et al.*, “Computational analysis and prediction of lysine malonylation sites by exploiting informative features in an integrative machine-learning framework,” *Briefings in Bioinformatics*, Aug. 2018, doi: 10.1093/bib/bby079.