Briefings in Bioinformatics, 22(4), 2021, 1–10

https://doi.org/10.1093/bib/bbaa255
Problem Solving Protocol

Deep-Kcr: accurate detection of lysine crotonylation sites using deep learning method

Hao Lv, Fu-Ying Dao, Zheng-Xing Guan, Hui Yang, Yan-Wen Li and Hao Lin

Corresponding authors: Yan-Wen Li. E-mail: liyw085@nenu.edu.cn; Hao Lin. Tel.: +86-13678168394; E-mail: hlin@uestc.edu.cn

Abstract

As a newly discovered protein posttranslational modification, histone lysine crotonylation (Kcr) involved in cellular regulation and human diseases. Various proteomics technologies have been developed to detect Kcr sites. However, experimental approaches for identifying Kcr sites are often time-consuming and labor-intensive, which is difficult to widely popularize in large-scale species. Computational approaches are cost-effective and can be used in a high-throughput manner to generate relatively precise identification. In this study, we develop a deep learning-based method termed as Deep-Kcr for Kcr sites prediction by combining sequence-based features, physicochemical property-based features and numerical space-derived information with information gain feature selection. We investigate the performances of convolutional neural network (CNN) and five commonly used classifiers (long short-term memory network, random forest, LogitBoost, naive Bayes and logistic regression) using 10-fold cross-validation and independent set test. Results show that CNN could always display the best performance with high computational efficiency on large dataset. We also compare the Deep-Kcr with other existing tools to demonstrate the excellent predictive power and robustness of our method. Based on the proposed model, a webserver called Deep-Kcr was established and is freely accessible at http://lin-group.cn/server/Deep-Kcr.

Key words: histone lysine crotonylation; computational prediction; feature encoding schemes; deep learning

Introduction

Protein posttranslational modifications (PTMs), in which amino acid residues in a protein are covalently modified, have been increasingly recognized to play a crucial role in diverse biological processes, such as DNA replication, cell differentiation and organismal development [1–4]. Abnormal PTMs can lead to various pathological conditions, such as developmental defects and malignant transformation [5]. With the advancement of modern proteomics technologies, the following short-chain lysine (K) acylations have been identified: acetylation, propionylation, butyrylation, 2-hydeoxyisobutyrylation, succinylation, malonylation, glutarylation, crotonylation and β-hydroxybutyrylation [2, 6–12]. Although some of these acylations have been studied in depth, especially for acetylation, little is known about the cellular regulation and functional relevance of newly identified acylations.

Histone lysine crotonylation (Kcr) is a newly discovered PTM that exists in several types of eukaryotic genomes (Figure 1)

Hao Lv is a PhD candidate of the Center for Informational Biology at the University of Electronic Science and Technology of China. His research interests include bioinformatics.
Fu-Ying Dao is a PhD candidate of the Center for Informational Biology at the University of Electronic Science and Technology of China. Her research interests include bioinformatics.
Zheng-Xing Guan is a master student of the Center for Informational Biology at the University of Electronic Science and Technology of China. His research interests include bioinformatics.
Hui Yang is a PhD candidate of the Center for Informational Biology at the University of Electronic Science and Technology of China. Her research interests include bioinformatics.
Yan-Wen Li is an Associate Professor at Northeast Normal University. Her research field is bioinformatics.
Hao Lin is a Professor at the Center for Informational Biology at the University of Electronic Science and Technology of China. His research is in the areas of bioinformatics and system biology.
Submitted: 13 August 2020; **Received (in revised form):** 31 August 2020

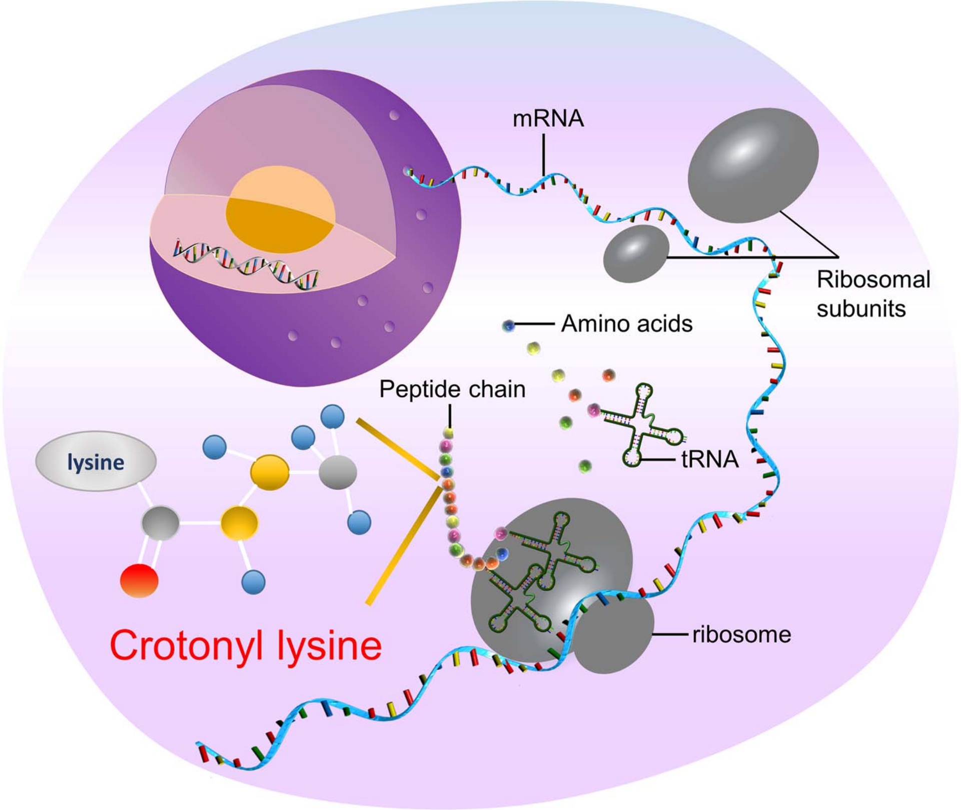

Figure 1. Schematic diagram of the process of histone lysine crotonylation modification.

[2, 13]. Histone Kcr may exert charge-based cis-effects on chromatin fibers by neutralizing the positive charge of the ε-amino group of K. Furthermore, the increase in bulkiness and rigidity of crotonyl may lead to an enhanced effect [14]. Therefore, histone Kcr is specifically marked as either an active promoter or potential enhancer in both human somatic and mouse male germ cell genomes [2]. Although these discoveries have prompted interest in the functional consequence of Kcr, the regulatory role and enzymes responsible for Kcr in the cellular environment remain unclear [15].

To fill this knowledge gap, several proteomics technologies—such as stable isotope labeling by amino acids in cell culture labeling, high-performance liquid chromatography fractionation, affinity enrichment and high-resolution liquid chromatography–tandem mass spectrometry—have been developed to detect Kcr sites [5]. However, due to the time-consuming and labor-intensive characteristics of these experimental methods, it is difficult to widely popularize it in large-scale species. Therefore, computational methods are needed to guide the interrogation of the Kcr sites of interest.

During the past few years, some computational approaches have been developed for the prediction of Kcr sites. Huang and Zeng [16] first established a discrete hidden Markov model called CrotPred to identify Kcr sites and achieved the areas under receiver operating characteristic (ROC) curves (AUCs) of 0.7823. Later on, Qiu *et al.* [17] proposed a novel approach based on position weight amino acid compositions and support vector machine (SVM) to predict Kcr sites. They obtained the AUC of 0.9217. Subsequently, Ju and He [18] developed CKSAAP_CrotSite for the identification of Kcr sites by using the composition of *k*-spaced amino acid pairs as the input coding and using the SVM as the classifier. The AUC of 0.9914 was reported. By incorporating five tiers of amino acid pairwise couplings into the general pseudo amino acid composition, Qiu *et al.* [19] built another computational tool based on ensemble random forest algorithm—named iKcr-PseEns—to predict Kcr sites. Their model produced the AUC of 0.9753. More recently, Malebary *et al.* [20] developed a new computational predictor, called iCrotoK-PseAAC, which incorporates various position and composition relative features along with statistical moments into PseAAC to identify Kcr sites. The accuracy reached to 0.9917.

Although the models mentioned above reported good performances on the identification of Kcr sites, some problems remain. (i) With the development of experimental technology, the resolution and sensitivity of sequence coverage for peptide mapping in histones have been significantly improved. However, the datasets used in the previous models have not been improved. (ii) While these have stimulated research on Kcr sites for prediction and validation, they did not systematically analyze and assess available features and/or machine learning methods in predicting Kcr sites.

Based on these ideas, the following key contributions of the present study can be documented and added to the field of bioinformatics: (i) a new benchmark dataset with high resolution and sensitivity was built to train a specific model. (ii) A novel convolutional neural network (CNN)-based model, called Deep-Kcr, was constructed for the prediction of Kcr sites by which significant improvements can be achieved in comparison with other existing tools. (iii) The performances of six different feature encoding schemes were assessed, and the effectiveness for improving the predictive ability of the model was confirmed. The frame of this work is exhibited in Figure 2.

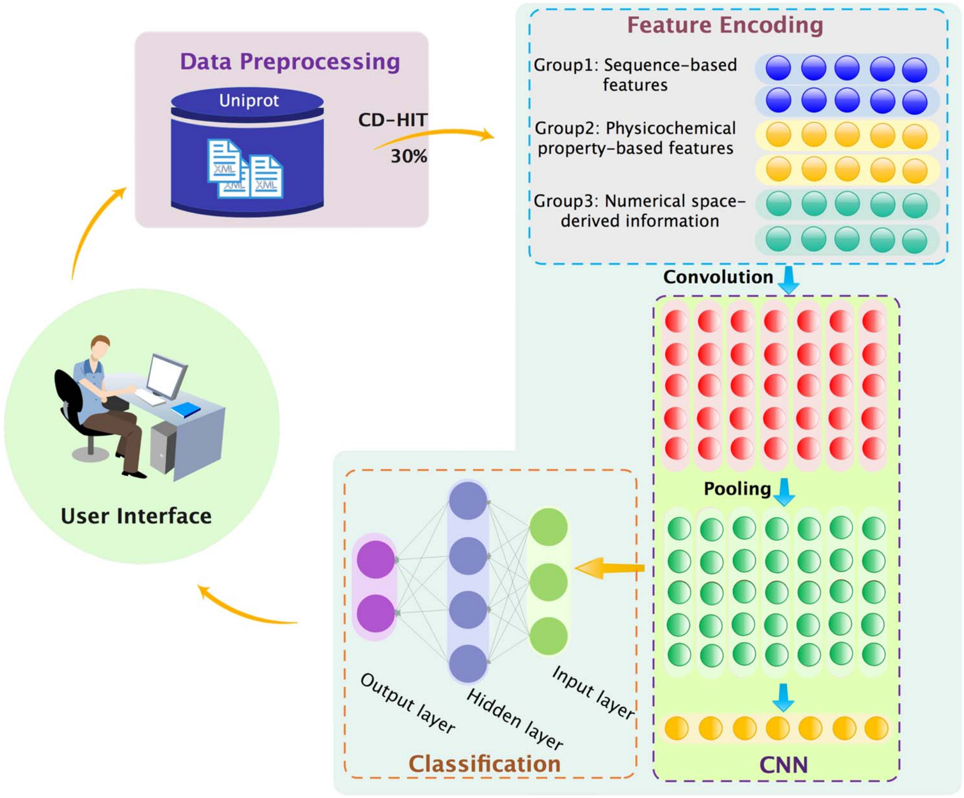

Figure 2. Overall framework of Deep-Kcr.

Materials and methods

Benchmark dataset construction

In this study, the dataset that was originally produced in previous literature was implemented [5].The dataset included 14 311 Kcr sites across 3734 histone proteins in the HeLa cell. After downloading all of the above-mentioned protein sequences from the UniProt database [21] using their UniProt IDs, the CD-HIT program [22] was used to get rid of redundant samples by setting the threshold of sequence identity to 30%. Then, the processed sequences were truncated into 31-residue-long sequence segments with K located at the center. A segment was defined as a positive sample if its central K was crotonylation; otherwise, it was defined as a negative sample. As a result, a great number of negative samples were obtained. To balance the positive and negative data, we randomly select sequences with the same number of positive samples from the non-redundant negative samples. After doing all of this, 9964 positive samples and 9964 negative samples were obtained. The data can be freely downloaded at http://lin-group.cn/server/Deep-Kcr/download.html. The non-redundant dataset was randomly divided into the training dataset and independent testing dataset according to the ratio of 7:3.

Feature encoding schemes

Adopting an effective feature encoding approach is a crucial step to produce a well-performing predictor [23, 24]. The following six types of feature encoding schemes were used to transform protein fragments into feature vectors: (i) composition of k-space amino acid pairs (CKSAAP) and position weight amino acid composition (PWAA) were employed to extract sequence-based features; (ii) amino acid index (AAindex), composition, transition and distribution (CTD) of physicochemical parameters and encoding based on grouped weight (EBGW) were selected to extract physicochemical property-based features and (iii) Word2Vec was used to extract numerical space-derived information.

Group 1: sequence-based features

CKSAAP. The CKSAAP scheme reflects short-range interactions between amino acid pairs and has been widely used in bioinformatics [25–27]. Taking k= 0 as an example, there are 400 0-spaced residue pairs (i.e. AA, AC, AD, ..., YY). The feature vector can be calculated using the following equation:

$$\left(\frac{N_{AA}}{N_{Total}},\frac{N_{AC}}{N_{Total}},\frac{N_{AD}}{N_{Total}},\cdots,\frac{N_{YY}}{N_{Total}}\right)_{400}, \quad (1)$$

where N_{Total} is the length of the total composition residues (for instance, if the protein fragment residue with a length L is 31 and $k=0,1,2,3,4$ and 5, then $N_{\text{Total}}=L-k-1$ will be 30, 29, 28, 27, 26 and 25, respectively). N_{AA}, N_{AC}, N_{AD}, $\cdots$, N_{YY} represent the frequency of the amino acid pair within the fragment. Considering that the CKSAAP scheme was performed over k = 0, 1, 2, 3, 4 and 5 in the present study, the total dimension of the CKSAAP-based feature vector is $400 \times 6=2400$.

PWAA. Given an amino acid residue $a_i\left(i\ =\ 1,2,\cdots,20\right)$, the position information of a_i in the sequence fragment P with $2L+1$

amino acids can be expressed as the following equation:

$$C_i = \frac{1}{L(L+1)} \sum_{j=-L}^{L} x_{i,j} \left(j + \frac{|j|}{L} \right) (j = -L, \cdots, 0, \cdots, L), \quad (2)$$

where L indicates the number of upstream residues or downstream residues from the central site in the protein sequence fragment P, $x_{i,j} = 1$ if a_i is the jth position residue in protein sequence fragment P, otherwise $x_{i,j} = 0$. Finally, the dimension of the PWAA-based feature vector is 20.

Group 2: physicochemical property-based features

AAindex. The primary physicochemical properties based on the amino acid indices were extracted from the AAindex database [28]. For a fragment (with a length of 31 residues in this study), the amino acid at each position is represented by 12 values according to physicochemical and biochemical properties: net charge; normalized frequency of alpha-helix; alpha helix propensity of position 44 in T4 lysozyme; composition of amino acids (AA) in intracellular proteins; AA composition of membrane of multi-spanning proteins; the volume of crystallographic water; information value for accessibility; transfer energy, organic solvent/water; AA composition of membrane proteins; entropy of formation; conformational preference for all beta-strands and optimized relative partition energies. Accordingly, the dimension of the AAindex-based feature vector is 31 × 12 = 372.

CTD. The CTD scheme was originally proposed by Dubchak *et al.* [29] to predict the protein-folding classes. All amino acids were divided into three classes: polar, neutral and hydrophobic. Details about the division of the amino acids are provided in Table S1. Each amino acid was then encoded by 1, 2 or 3, according to the class it belonged to. The compositional (C) describes the global percent composition of 20 native amino acids, which is defined as

$$C = {}^{n_s}\!/_{L} \ (s = 1, 2, 3), \quad (3)$$

where n_s is the number of s in the encoded sequence, and L is the length of the protein fragment sequence.

Transition (T) characterizes the percent frequency with amino acids from one type of native amino acid followed by another type, which can be calculated by

$$T = \frac{n_{xy} + n_{yx}}{L - 1} \ (\mathrm{xy} = [12], [13], [23]), \quad (4)$$

where n_{xy} is the number of dipeptides encoded as 'xy' and 'yx', respectively.

Distribution (D) measures the respective locations of the first, 25, 50, 75 and 100% of each type of 20 native amino acids, which for the descriptor E_i is defined as

$$E_i 1D_x = \frac{P_1}{L}; E_i 25D_x = \frac{P_{25}}{L}; E_i 50D_x = \frac{P_{50}}{L}; E_i 75D_x = \frac{P_{75}}{L}; E_i 100D_x = \frac{P_{100}}{L} \ (i = 1, 2, \cdots, 7; x = 1, 2, 3), \quad (5)$$

where P_1, P_{25}, P_{50}, P_{75} and P_{100} measure the position of the first residue, 25, 50, 75 and 100% occurrence of x, respectively. Based on the seven physicochemical properties listed in Table S1, the dimension of the CTD-based feature vector is 147.

EBGW. The EBGW scheme [30] groups 20 amino acids into seven groups according to their hydrophobicity and charge characteristics (Table S2). For a protein sequence P, it can be transformed into three binary sequences

$$\begin{cases} H_1(P_j) = \begin{cases} 1 \ if P_j \in H_1 \\ 0 \ if P_j \notin H_1 \end{cases} \\ H_2(P_j) = \begin{cases} 1 \ if P_j \in H_2 \\ 0 \ if P_j \notin H_2 \end{cases} \\ H_3(P_j) = \begin{cases} 1 \ if P_j \in H_3 \\ 0 \ if P_j \notin H_3 \end{cases} \end{cases} . \quad (6)$$

Every binary sequence can be divided into a J subsequence increasing in length. For instance, for H_1, the jth subsequence is defined as the following:

$$X_1(j) = \frac{\mathrm{Sum}(j)}{D(j)}, \quad (7)$$

where Sum(j) is the number of 1 in the jth subsequence, $D(j) = \mathrm{int}\left(j \times L/J\right)$ refers to the length of the jth subsequence and the function $\mathrm{int}()$ rounds a decimal to the closest integer. The EBGW scheme defines a protein sequence as the $3 \times J$-dimension vector. Here, J is selected as 1, 2, 3, 4 and 5. Accordingly, the dimension of the EBGW-based feature vector is 3 × 15 = 45.

Group 3: numerical space-derived information

Word2Vec. Word2Vec is a natural language processing (NLP) system that utilizes the neural network model to create a distributed representation of words in a corpus [31]. In such a system, words that appear in the same documents share semantic meaning, meaning that words are embedded in a continuous vector space where 'semantically similar' words have closer distances. The Word2Vec includes two model architectures: one is the continuous bag-of-words (CBOW) model, and the other is the continuous skip-gram model. The basic principle of the CBOW model uses a word's context words in a surrounding window to predict the word, while the basic principle of the skip-gram model uses a word to predict its surrounding words [32]. Algorithmically, the structure of the CBOW model exhibits the advantage of uniformly organizing the information distributed in the dataset. Therefore, the CBOW model was implemented for training the contextual similarity of protein sequences in the present work, and its architecture and parameters are exhibited in https://github.com/linDing-group/Deep-Kcr. Eventually, the dimension of the Word2Vec-based feature vector is 200.

Deep neural network

Recently, as a cutting-edge technique, deep learning-based methods, which include CNNs, have been successfully applied in the field of bioinformatics for the prediction of protein modification sites [33–35], RNA modification sites [36] and so on. The architecture of the designed CNN is contained by the input layer, a single convolution layer, a max-pooling layer and an output layer. The one-dimensional convolution layer defines 32 filters of a height 3 and a stride size of 1. For activation, the rectified linear unit (ReLu) was used followed by a convolution operation. To reduce computational complexity and avoid overfitting, the one-dimensional pooling layer defines the pool size as 2 and the strides as 2. Besides, the feature vectors were transferred to four consecutive, fully connected layers after the flattening operation. In these fully connected layers, the first three layers had 32, 16 and 8 nodes, respectively, followed by the ReLu activation function and regularization function. The fourth layer had only one binary node, followed by a Softmax activation function to predict Kcr sites (Yes/No). Details on the tensor processing of each layer are shown in https://github.com/linDing-group/Deep-Kcr.

Performance evaluation

To evaluate the prediction performance of the proposed method, a 10-fold cross-validation test was performed [37–41]. ROC curves were calculated and plotted based on specificities and sensitivities by taking different thresholds.

$$\begin{cases} \text{Specificity} = \frac{\text{TN}}{\text{FP+TN}} \\ \text{Sensitivity} = \frac{\text{TP}}{\text{TP+FN}} \end{cases}. \quad (8)$$

Where TP is the number of positive samples that are correctly classified in the prediction, and TN indicates the number of negative samples that are correctly classified by predictors. FP and FN represent the numbers of positive or negative samples that are classified by mistake, respectively. Additionally, AUCs were also calculated based on the trapezoidal approximation [42].

Results

Analysis of sequence composition, structure and function based on a benchmark dataset

Based on the curated dataset, the graphical sequence logo (t-test, $P < 0.05$) was generated by Two Sample Logo [43] to identify distinct patterns or conserved sequence motifs between the Kcr site-containing sequences and the background sequences (Figure 3A) [44]. As presented in Figure 3A, the following points were observed. (i) The most apparent feature of a Kcr site-containing sequences is the abundance of hydrophobic amino acids such as valine (V), leucine (L) and isoleucine (I) in both the upstream and downstream. Furthermore, charged and polar amino acids such as K, glutamic acid (E) and serine (S) were depleted at almost all positions, which displays a significant difference in amino acid compositions between positive and negative samples. Simultaneously, it suggests that the physicochemical properties of these specific amino acids around the Kcr site might, therefore, be used as informative features for building a model to predict Kcr sites. (ii) The negatively charged glutamic acid (E) and aspartic acid (D) were enriched from the position of −3 to +3, whereas positively charged K were remarkably depleted at varying positions from −4 to 4 surrounding the non-Kcr site. This indicates that the electric field conditions formed by negatively charged amino acid residues are more conducive to Kcr. (iii) A conserved motif, namely RRVDD/DDVRR, was significantly enriched in the flanking regions on both sides from positions −6 to −2 and 2 to 6. This mirror structure may promote the binding of crotonyl-coenzyme A hydratase to K residues. However, two motifs (KxxK) and (PxK) show a stronger preference surrounding the modified lysine in non-histone, where ‘x’ represent any amino acid [45].

To determine whether there is a structure preference for Kcr, the NetSurfP [46] was implemented for structure analysis on the positive and negative samples. The result showed that approximately 50% of Kcr sites were found in α helices, 10% were located in β strands and the remaining 40% were seen in disordered coils (Figure 3B). The distribution pattern of Kcr exhibited no significant difference from that of non-Kcr protein K residues, suggesting that the structural information of the Kcr protein should not be selected as the encoding strategy due to the power of poverty to discriminate the positive and negative samples.

To analyze the functional distribution of the Kcr, the enrichment analysis of Gene Ontology (GO) terms was performed to choose statistically significant results. Previous studies have shown that histone-modified crotonyl groups have extremely structures compared with acetyl groups, in which both types of modifications share the same enzyme system [13, 14]. Therefore, it is assumed that the crotonyl group and acetyl group colocalize on a certain K residue. To test this prediction, the Kcr distribution was compared with previously obtained results for lysine acetylation (Kac) distribution exhibited in the PLMD database [47–49]. The result showed that histone crotonylation occupies 3760 of the same locations as acetylation because there is an apparent overlap between Kcr and Kac peaks, and 6187 histone crotonylations occur independently of acetylation. Next, the functional analysis of the colocalized Kcr sites and individual Kcr sites was investigated (Figure 3C–F). As shown in Figure 3C and D, for colocalized Kcr sites, they are all statistically enriched in RNA catabolic process, RNA splicing and translational initiation for biological processes. However, for individual Kcr sites, they are all significantly enriched in Golgi vesicle transport, nuclear transport, nucleocytoplasmic transport, ribosome biogenesis and RNA localization. For cellular components, colocalized Kcr sites tend to be located within the cytoplasm, ribosomes and spliceosomal complex, whereas individual Kcr sites can be found in cellular structures related to connecting intracellular and extracellular matrices (ECM), such as the nuclear envelope and focal adhesion. For molecular functions, both colocalized Kcr sites and individual Kcr sites are mainly involved in cell adhesion molecule binding and cadherin binding. These results suggested that colocalized Kcr sites and individual Kcr sites play a key role in the different stages of biological processes. The former performs important functions in the process of transcription and translation, while the latter mainly mediates material transportation and signal transduction between the nucleus and the cytoplasm.

To support the above analysis, the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway was performed to investigate the pathways in which Kcr was likely to be involved with. As shown in Figure 3E and F, except for some pathways related to diseases, the colocalized sites that significantly enriched the pathways of spliceosome and ribosome imply a potential role of Kcr in protein synthesis, while individual Kcr sites tend to participate in the pathways of endocytosis and RNA transport. These results are consistent with the GO annotation. In short, in HeLa cells, the same histone K residue can be simultaneously marked by both histone Kcr and Kac. In addition, many Kcr sites lack histone Kac, indicating that crotonylation can occur independently of Kac and play a role in different gene expression regulation processes.

Performance evaluation of different feature encoding schemes

The performances of six different feature encoding schemes (categorized into three major groups) were assessed using CNN based on 10-fold cross-validation (Figure 4). Generally, no single group of features outperforming other groups was observed. The

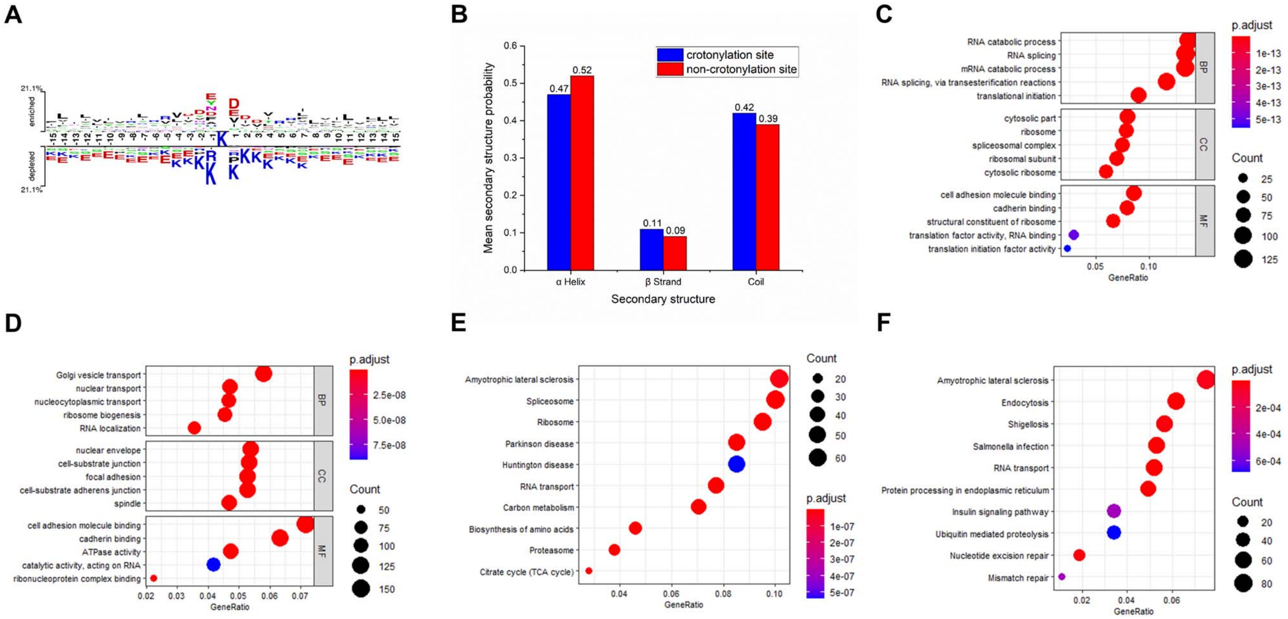

Figure 3. Analysis of sequence composition, structure and function based on benchmark dataset. (A) Sequence conservation pattern illustration generated by Two Sample Logo to show the amino acids with statistically significant differences surrounding the Kcr sites. Only residues significantly enriched or depleted (t-test, $P < 0.05$) flanking the centered Kcr sites are shown. (B) Distribution of Kcr site and non-Kcr site in structured regions of proteins. (C, D) Bubble chart showing GO terms of the colocalized Kcr sites and individual Kcr sites in the category of biological process (BP), molecular function (MF) and cellular component (CC). (E, F) Bubble chart showing KEGG pathway associated with colocalized Kcr sites and individual Kcr sites, respectively.

AAindex and CTD features based on physicochemical properties, CKSAAP features based on sequence information and Word2Vec features based on evolutionary information are the four most important types of features that produced a satisfactory performance. As expected, the PWAA features based on sequence information and EBGW features based on physicochemical properties showed poor performance. A possible reason for this is that the number of features generated by PWAA and EBGW is too small to extract enough useful patterns and characteristics. This implies the necessity of exploiting effective combinations of feature types from various aspects instead of using individual feature types alone, which have limited the predictive power of the trained models [50].

To further improve the predictive performance of the models, the extracted features were combined to form a 3184-dimensional feature set. As shown in Figure 4A, the model trained on the fusion feature set achieved better performance (AUC = 0.8671) compared with the models trained on the original feature sets (all AUCs are below 0.8484), indicating that the feature fusion strategy is effective in the prediction of Kcr sites to produce significant performance improvement. However, heterogeneous features may lead to a ‘feature disaster’ that undermine model performance. Therefore, the Information Gain (IG) [51] feature selection method was applied to the full feature sets, and the ranked attributes were obtained. Different feature sets that contained the top-ranked features were then created and tested, ranging from the top 50 to the top 750 features, with steps of 50. It was found that as the feature dimension increased, the performance of the model improved. Moreover, when the IG was selected to be greater than 250, the performance of the model tended to be stable and reached the peak when the IG was equal to 450, which corresponded to 0.8846 of AUC (Figure 4B). The results showed that the information gain feature selection method removes redundant features and improves the predictive ability of the model, providing useful insights when considering building a better prediction model for histone K modification sites.

Next, the feature importance and its contribution were further analyzed to find which feature was more valuable for the model performance after feature selection (Figure 4C–E). As shown in Figure 4C and D, the features contained in the optimal feature set are 40% Word2Vec, 21% CTD and 21% AAindex, suggesting their critical importance for Kcr site prediction. It is worth noting that although the optimal feature set contains only 9% EBGW- and 2% PWAA-based features, these two features are still very valuable based on the ratio of the selected dimension to the original dimension [0.89 (40/45) of EBGW and 0.45 (9/20) of PWAA] (Figure 4E). This finding indicated a significant contribution to the prediction of Kcr sites and was consistent with previous prediction studies of histone K modification sites for which EBGW- and PWAA-based features are often considered essential [52, 53]. Apart from these, there is too much redundant information included in CKSAAP-based features, so it only makes a subtle contribution when 0.01 (32/2400) features are selected and included in the optimal feature set.

From the above analysis, the feature encoding schemes used in the present study are effective for improving the predictive ability of the model. Furthermore, the IG feature selection method can fully consider the importance and contribution of each feature in the case of performance improvement.

Prediction performance with different classifiers

To test the validity of the optimal feature set in different classifiers, five other common classifiers were used to predict Kcr sites: long short-term memory network (LSTM), random forest (RF), LogitBoost (LB), naive Bayes (NB) and logistic regression (LR). To evaluate the prediction performance and robustness objectively, 10-fold cross-validation was performed (Figure 5). As shown in Figure 5, the AUC value of the CNN is 0.8846, which

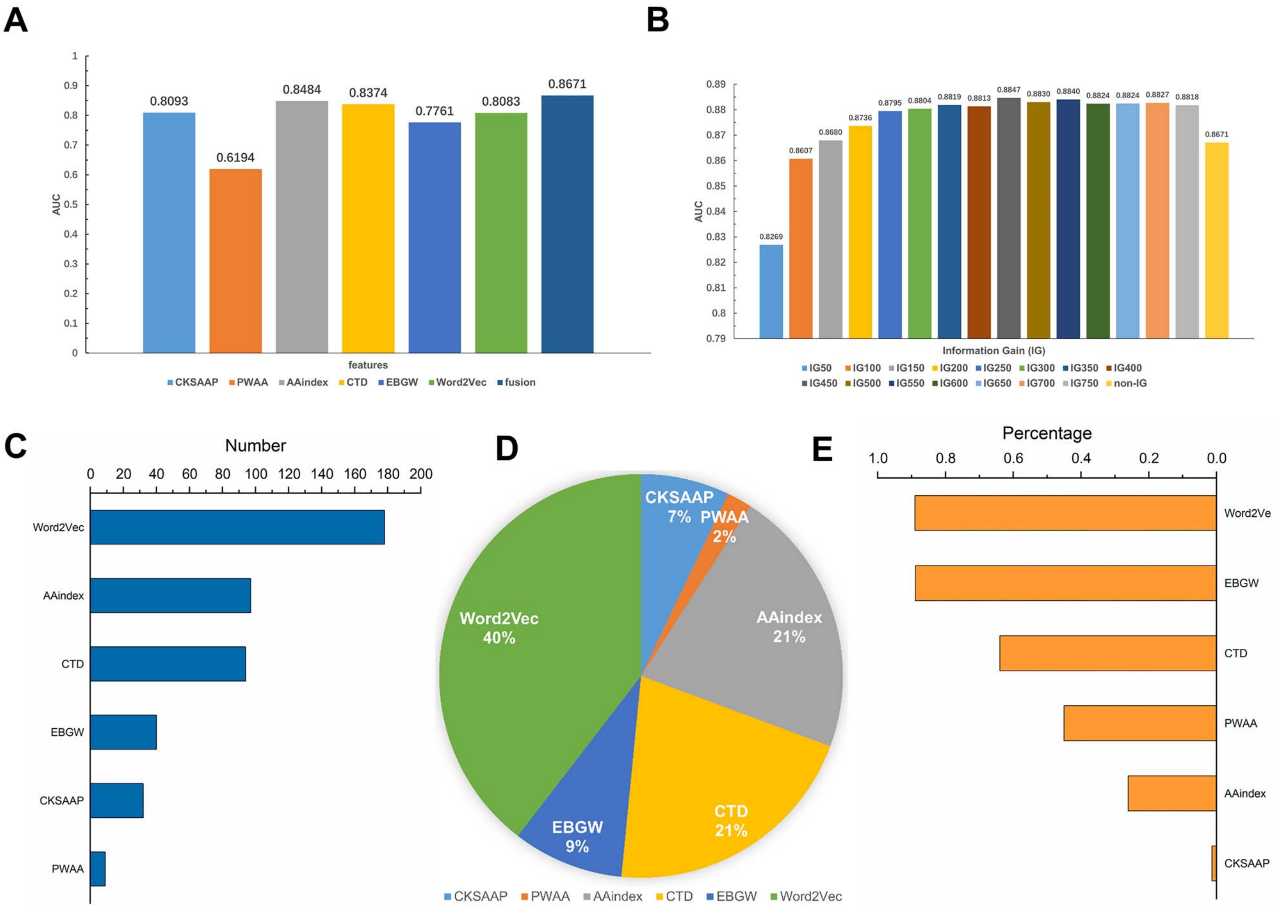

Figure 4. Analysis of single feature, fusion features and optimal feature set. (A) AUCs based on the single feature. (B) AUCs based on the fusion features optimized by IG; IG50, IG100, ···, IG750 represent the first 50, first 100, ···, first 750 features after IG optimized the original feature set. (C, D) The number and proportion of the types of features selected in the optimal feature set. (E) The ratio of selected dimension to original dimension in the optimal feature set.

is higher than those of LSTM, RF, LB, NB and LR at 0.3114, 0.0197, 0.0566, 0.1198 and 0.0348, respectively. This result suggested that CNN could achieve both the best performance and computational efficiency when being applied to the largest dataset.

Next, the performance of the single classifier on the independent dataset was investigated. As shown in Figure 5, the performance is highly consistent with the results from the 10-fold cross-validation as mentioned above, where the AUC of CNN is 0.8591, which is higher than those of LSTM, RF, LB, NB and LR at 0.3480, 0.0469, 0.0444, 0.1424 and 0.1847, respectively. This further demonstrates the stability and generalization ability of the constructed model.

Prediction performance comparison with existing tools

It is necessary to compare the proposed method with other existing state-of-the-art tools. To make an equal and objective evaluation of the performance, the three most popular and recently published tools were selected: the position weight-based method [17], CKSAAP_CrotSite [18] and LightGBM-CroSite [54]. All of these methods use the same window side as the method proposed in the present study, making the comparison both fair and accurate. We first conducted the comparison on training dataset with 10-fold cross-validation. As shown in Figure 6, the Deep-Kcr is superior to other tools, except for LightGBM-CroSite. However, previous research has shown that the testing result on the independent set test is more reliable and should be given more weight when evaluating the performance of different methods [44]. Therefore, an independent set test was also conducted to ensure the same evaluation criterion and to avoid potential bias. It should be noted that the position weight-based method and LightGBM-CroSite did not provide a webserver or software package. Although the CKSAAP_CrotSite provided a webserver, it does not work at present. For fair comparison, we rebuilt the models of these three tools, and the corresponding performances were obtained. As shown in Figure 6, the AUC produced by the proposed method outperformed the other methods, which is higher than those of position weight-based method, CKSAAP_CrotSite and LightGBM-CrotSite at 0.2865, 0.3478 and 0.3497, respectively. This result indicated that the proposed model provides an excellent predictive power compared with existing tools.

Discussion

In this work, a Deep-Kcr method based on CNN that also combined sequence-based features, physicochemical property-based features and numerical space-derived information with the information gain feature selection method was proposed to identify histone Kcr sites. The prediction performances of different feature encoding schemes, different classifiers and comparison with existing tools showed the robustness and

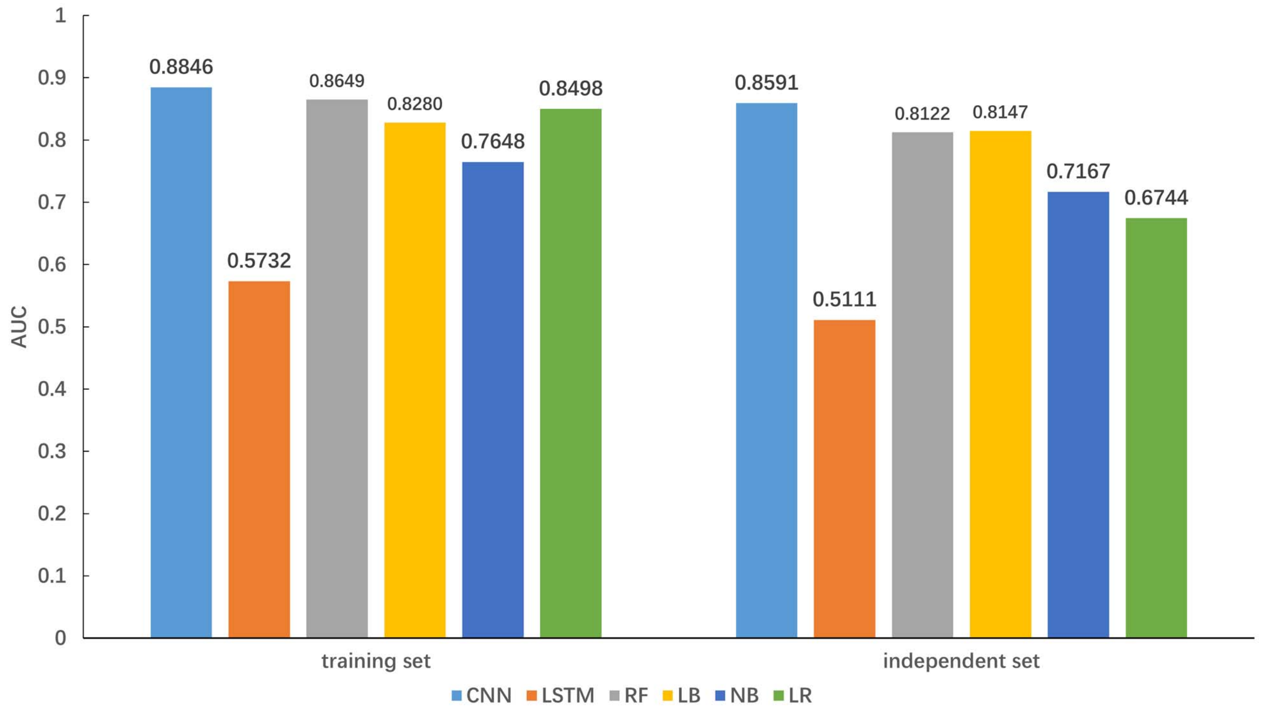

Figure 5. Comparison of the performance on different classifiers based on training set and independent set.

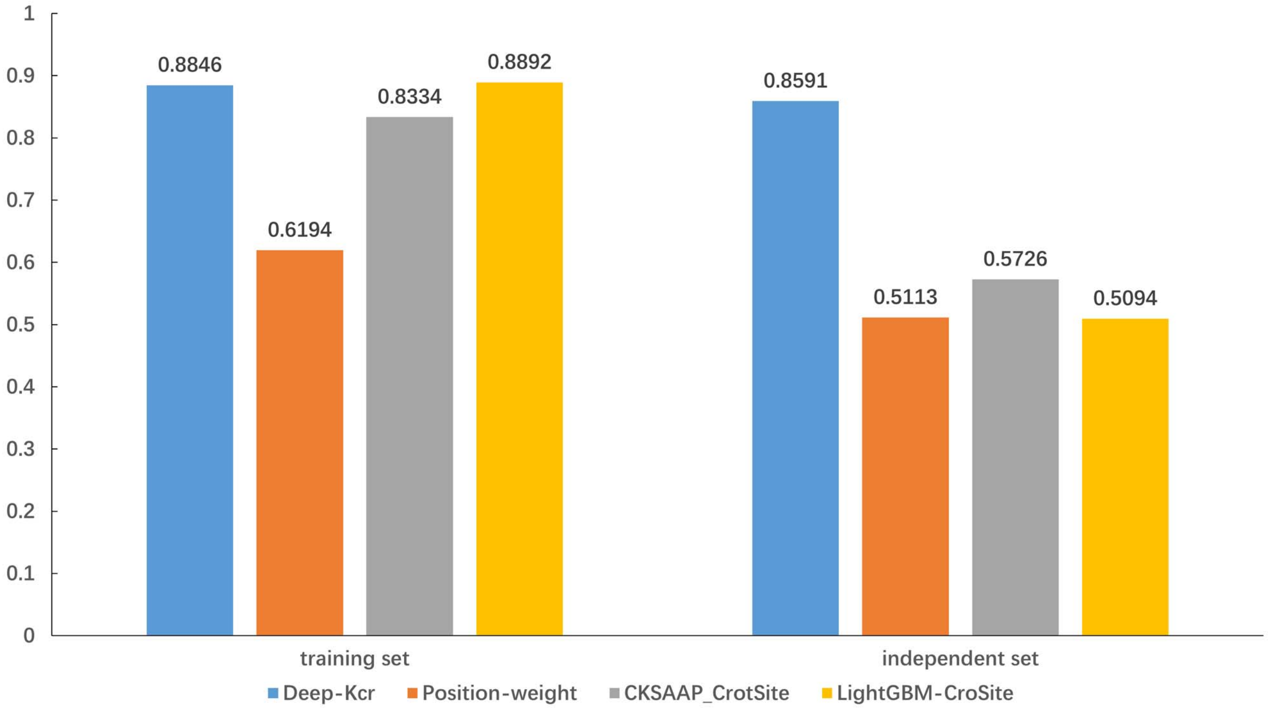

Figure 6. AUC values of our proposed method and other existing tools based on training set and independent set test.

generalization ability of Deep-Kcr. It is believed that Deep-Kcr could serve as a useful platform for the discovery of novel crotonylation.

Several directions are worth exploring in the future. First, there are other ways to attain cell-specific proteomic signals for this model, such as data for protein structure information and expression level. A comparison of model performance given different types of proteomics data as the input will shed light on the interpretation of the relationship between different proteomic events and PTMs. Second, NLP has widely proven its powerful ability to fit data in bioinformatics [55, 56]. Consistent with previous works, the Word2Vec-based NLP method used in this work can effectively extract the information in the sequence context. Therefore, the features learned by NLP could be further mined to map sequence data into three-dimensional (3D) space, thereby providing the possibility of exploring the information contained in the 3D structure of proteins. Third, in this work, a CNN-based model was applied to predict Kcr sites, achieving satisfactory performance and computational efficiency when used on the benchmark dataset of 10 000 levels. This success is attributable to its specific design advantage compared with traditional machine learning methods, making itself particularly suitable for high-throughput omics data and parallel computing. Therefore, it is expected that other types of deep learning-based methods—such as recurrent neural networks, deep belief networks, deep reinforcement learnings and generative adversarial networks—could be applied to general proteomics data in future works. Fourth, there is evidence suggesting that histone Kcr specifically labels enhancers and, most precisely, the transcription start site of active genes in both human somatic and mouse male germ cell genomes [2]. Based on this idea, it is speculated that there is a consistent phenomenon in human HeLa cells. This phenomenon can be further extended to histone Kcr to be used as a boundary signal for enhancer–promoter interactions in the 3D genome. Unfortunately, this assumption was implemented, but no significant results were found. However, this hypothesis is still worthy of being tested in other PTMs.

Key Points

- We provided a systematic comparison of six feature encoding schemes in identifying Kcr sites and obtained optimized feature set by using IG feature selection method.
- We proposed Deep-Kcr, a novel deep learning-based method for Kcr sites prediction.
- Experimental results on optimized feature set demonstrate the superior performance of Deep-Kcr compared with other four different machine learning classifiers.
- Comparative analysis of Deep-Kcr and other existing tools showed that Deep-Kcr exhibits an excellent predictive power.

Data Availability

We provide the Python source code of deep-Kcr model training, which is freely available at https://github.com/linDing-group/Deep-Kcr.

Supplementary Data

Supplementary data are available online at *Briefings in Bioinformatics*.

Author Contributions

Conceptualization: H.L., Y.-W.L.; investigation: H.L., F.-Y.D.; coding: H.L., Z.-X.G.; writing—original draft: H.L., F.-Y.D., H.Y.; writing—review and editing: H.L.; funding acquisition: H.L.

Conflict of Interest

The authors declare that they have no competing interests.

Acknowledgements

We would like to thank the anonymous reviewers for valuable suggestions.

Funding

National Nature Scientific Foundation of China (61772119); Sichuan Provincial Science Fund for Distinguished Young Scholars (2020JDJQ0012); Science Strength Promotion Programme of UESTC.

References

1. Wan J, Liu H, Chu J, *et al.* Functions and mechanisms of lysine crotonylation. *J Cell Mol Med* 2019;**23**:7163–9.
2. Tan M, Luo H, Lee S, *et al.* Identification of 67 histone marks and histone lysine crotonylation as a new type of histone modification. *Cell* 2011;**146**:1016–28.
3. Kouzarides T. Chromatin modifications and their function. *Cell* 2007;**128**:693–705.
4. Huang G, Li J. Feature extractions for computationally predicting protein post-translational modifications. *Current Bioinformatics* 2018;**13**:387–95.
5. Yu H, Bu C, Liu Y, *et al.* Global crotonylome reveals CDYL-regulated RPA1 crotonylation in homologous recombination-mediated DNA repair. *SciAdv* 2020;**6**: eaay4697.
6. Sabari BR, Zhang D, Allis CD, *et al.* Metabolic regulation of gene expression through histone acylations. *Nat Rev Mol Cell Biol* 2017;**18**:90–101.
7. Dai L, Peng C, Montellier E, *et al.* Lysine 2-hydroxyisobutyrylation is a widely distributed active histone mark. *Nat Chem Biol* 2014;**10**:365–70.
8. Chen Y, Sprung R, Tang Y, *et al.* Lysine propionylation and butyrylation are novel post-translational modifications in histones. *Mol Cell Proteomics* 2007;**6**:812–9.
9. Xie Z, Dai J, Dai L, *et al.* Lysine succinylation and lysine malonylation in histones. *Mol Cell Proteomics* 2012;**11**:100–7.
10. Tan M, Peng C, Anderson KA, *et al.* Lysine glutarylation is a protein posttranslational modification regulated by SIRT5. *Cell Metab* 2014;**19**:605–17.
11. Xie Z, Zhang D, Chung D, *et al.* Metabolic regulation of gene expression by histone lysine beta-Hydroxybutyrylation. *Mol Cell* 2016;**62**:194–206.
12. Bao W, Huang D-S, Chen Y-H. MSIT: Malonylation sites identification tree. *Current Bioinformatics* 2020;**15**:59–67.
13. Bao X, Wang Y, Li X, *et al.* Identification of 'erasers' for lysine crotonylated histone marks using a chemical proteomics approach. *Elife* 2014;**3**:e02999.
14. Sabari BR, Tang Z, Huang H, *et al.* Intracellular crotonyl-CoA stimulates transcription through p300-catalyzed histone crotonylation. *Mol Cell* 2015;**58**:203–15.
15. Wei W, Liu X, Chen J, *et al.* Class I histone deacetylases are major histone decrotonylases: evidence for critical and broad function of histone crotonylation in transcription. *Cell Res* 2017;**27**:898–915.
16. Huang G, WJJ Z. A discrete hidden Markov model for detecting histone crotonyllysine sites, MATCH Commun. *Math Comput Chem* 2016;**75**:717–30.
17. Qiu WR, Sun BQ, Tang H, *et al.* Identify and analysis crotonylation sites in histone by using support vector machines. *Artif Intell Med* 2017;**83**:75–81.

18. Ju Z, He JJ. Prediction of lysine crotonylation sites by incorporating the composition of k-spaced amino acid pairs into Chou's general PseAAC. *J Mol Graph Model* 2017;**77**:200–4.
19. Qiu WR, Sun BQ, Xiao X, *et al.* iKcr-PseEns: identify lysine crotonylation sites in histone proteins with pseudo components and ensemble classifier. *Genomics* 2018;**110**:239–46.
20. Malebary SJ, Rehman MSU, Khan YD. iCrotoK-PseAAC: identify lysine crotonylation sites by blending position relative statistical features according to the Chou's 5-step rule. *PLoS One* 2019;**14**:e0223993.
21. UniProt C. Ongoing and future developments at the universal protein resource. *Nucleic Acids Res* 2011;**39**:D214–9.
22. Huang Y, Niu B, Gao Y, *et al.* CD-HIT suite: a web server for clustering and comparing biological sequences. *Bioinformatics* 2010;**26**:680–2.
23. Wei L, Su R, Luan S, *et al.* Iterative feature representations improve N4-methylcytosine site prediction. *Bioinformatics* 2019;**35**:4930–7.
24. Chen W, Feng P, Song X, *et al.* iRNA-m7G: identifying N(7)-methylguanosine sites by fusing multiple features. *Mol Ther Nucleic Acids* 2019;**18**:269–74.
25. Li F, Li C, Wang M, *et al.* GlycoMine: a machine learning-based approach for predicting N-, C- and O-linked glycosylation in the human proteome. *Bioinformatics* 2015;**31**:1411–9.
26. Chen Z, Zhou Y, Zhang Z, *et al.* Towards more accurate prediction of ubiquitination sites: a comprehensive review of current methods, tools and features. *Brief Bioinform* 2015;**16**:640–57.
27. Chen J, Zhao J, Yang S, *et al.* Prediction of protein ubiquitination sites in Arabidopsis thaliana. *Current Bioinformatics* 2019;**14**:614–20.
28. Kawashima S, Pokarowski P, Pokarowska M, *et al.* AAindex: amino acid index database, progress report 2008. *Nucleic Acids Res* 2008;**36**:D202–5.
29. Dubchak I, Muchnik I, Holbrook SR, *et al.* Prediction of protein folding class using global description of amino acid sequence. *Proc Natl Acad Sci U S A* 1995;**92**:8700–4.
30. Zhang ZH, Wang ZH, Zhang ZR, *et al.* A novel method for apoptosis protein subcellular localization prediction combining encoding based on grouped weight and support vector machine. *FEBS Lett* 2006;**580**:6169–74.
31. Mikolov T, Chen K, Corrado G, *et al.* Efficient estimation of word representations in vector space. *arXiv* 2013;1301.3781.
32. Guo D, Wang Q, Liang M, *et al.* Molecular cavity topological representation for pattern analysis: a NLP analogy-based Word2Vec method. *Int J Mol Sci* 2019;**20**:6019.
33. Wang D, Liang Y, Xu D. Capsule network for protein post-translational modification site prediction. *Bioinformatics* 2019;**35**:2386–94.
34. Rao RSP, Zhang N, Xu D, *et al.* CarbonylDB: a curated data-resource of protein carbonylation sites. *Bioinformatics* 2018;**34**:2518–20.
35. Long H, Wang M, Fu H. Deep convolutional neural networks for predicting hydroxyproline in proteins. *Current Bioinformatics* 2017;**12**:233–8.
36. Xu H, Jia P, Zhao Z. Deep4mC: systematic assessment and computational prediction for DNA N4-methylcytosine sites by deep learning. *Brief Bioinform* 2020. doi: 10.1093/bib/bbaa099.
37. Hasan MAM, Ben Islam MK, Rahman J, *et al.* Citrullination site prediction by incorporating sequence coupled effects into PseAAC and resolving data imbalance issue. *Current Bioinformatics* 2020;**15**:235–45.
38. Basith S, Manavalan B, Shin TH, *et al.* SDM6A: a web-based integrative machine-learning framework for predicting 6mA sites in the Rice genome. *Mol Ther Nucleic Acids* 2019;**18**:131–41.
39. Manavalan B, Basith S, Shin TH, *et al.* Meta-4mCpred: a sequence-based meta-predictor for accurate DNA 4mC site prediction using effective feature representation. *Mol Ther Nucleic Acids* 2019;**16**:733–44.
40. Manavalan B, Basith S, Shin TH, *et al.* mAHTPred: a sequence-based meta-predictor for improving the prediction of anti-hypertensive peptides using effective feature representation. *Bioinformatics* 2019;**35**:2757–65.
41. Liu K, Chen W. iMRM: a platform for simultaneously identifying multiple kinds of RNA modifications. *Bioinformatics* 2020;**36**:3336–42.
42. Gao J, Thelen JJ, Dunker AK, *et al.* Musite, a tool for global prediction of general and kinase-specific phosphorylation sites. *Mol Cell Proteomics* 2010;**9**:2586–600.
43. Vacic V, Iakoucheva LM, Radivojac P. Two sample logo: a graphical representation of the differences between two sets of sequence alignments. *Bioinformatics* 2006;**22**:1536–7.
44. Li Y, Wang M, Wang H, *et al.* Accurate in silico identification of species-specific acetylation sites by integrating protein sequence-derived and functional features. *Sci Rep* 2014;**4**:5765.
45. Wei W, Mao A, Tang B, *et al.* Large-scale identification of protein crotonylation reveals its role in multiple cellular functions. *J Proteome Res* 2017;**16**:1743–52.
46. Petersen B, Petersen TN, Andersen P, *et al.* A generic method for assignment of reliability scores applied to solvent accessibility predictions. *BMC Struct Biol* 2009;**9**:51.
47. Xu H, Zhou J, Lin S, *et al.* PLMD: an updated data resource of protein lysine modifications. *J Genet Genomics* 2017;**44**:243–50.
48. Liu Z, Wang Y, Gao T, *et al.* CPLM: a database of protein lysine modifications. *Nucleic Acids Res* 2014;**42**:D531–6.
49. Liu Z, Cao J, Gao X, *et al.* CPLA 1.0: an integrated database of protein lysine acetylation. *Nucleic Acids Res* 2011;**39**:D1029–34.
50. Zhang Y, Xie R, Wang J, *et al.* Computational analysis and prediction of lysine malonylation sites by exploiting informative features in an integrative machine-learning framework. *Brief Bioinform* 2019;**20**:2185–99.
51. Shannon CE (ed). A mathematical theory of communication. *Bell Labs Tech J* 1948;**27**:379–423.
52. Yu J, Shi S, Zhang F, *et al.* PredGly: predicting lysine glycation sites for Homo sapiens based on XGboost feature optimization. *Bioinformatics* 2019;**35**:2749–56.
53. Shi SP, Qiu JD, Sun XY, *et al.* PLMLA: prediction of lysine methylation and lysine acetylation by combining multiple features. *Mol Biosyst* 2012;**8**:1520–7.
54. Liu Y, Yu Z, Chen C, *et al.* Prediction of protein crotonylation sites through LightGBM classifier based on SMOTE and elastic net. *Anal Biochem* 2020;**609**:113903.
55. Wang JH, Zhao LF, Wang HF, *et al.* GenCLiP 3: mining human genes' functions and regulatory networks from PubMed based on co-occurrences and natural language processing. *Bioinformatics* 2019. doi: 10.1093/bioinformatics/btz807.
56. Magge A, Weissenbacher D, O'Connor K, *et al.* GeoBoost2: anatural language processing pipeline for GenBankmetadata enrichment for virus Phylogeography. *Bioinformatics* 2020. doi: 10.1093/bioinformatics/btaa647.