Computational and Structural Biotechnology Journal 21 (2023) 120–127

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

CapsNh-Kcr: Capsule network-based prediction of lysine crotonylation sites in human non-histone proteins

Jhabindra Khanal [a], Jeevan Kandel [b], Hilal Tayara [c,*], Kil To Chong [a,d,*]

[a] Department of Electronics and Information Engineering, Jeonbuk National University, Jeonju 54896, South Korea
[b] Graduate School of Integrated Energy-AI, Jeonbuk National University, Jeonju 54896, South Korea
[c] School of International Engineering and Science, Jeonbuk National University, Jeonju 54896, South Korea
[d] Advanced Electronics and Information Research Center, Jeonbuk National University, Jeonju 54896, South Korea

ARTICLE INFO

Article history:
Received 20 July 2022
Received in revised form 10 November 2022
Accepted 26 November 2022
Available online 1 December 2022

Keywords:
Lysine crotonylation (Kcr)
Deep learning
Capsule network
Motifs
Web-server

ABSTRACT

Lysine crotonylation (Kcr) is one of the most important post-translational modifications (PTMs) that is widely detected in both histone and non-histone proteins. In fact, Kcr is reported to be involved in various biological processes, such as metabolism and cell differentiation. However, the available experimental methods for Kcr site identification are laborious and costly. To effectively replace existing experimental approaches, some computational methods have been developed in the last few years. The available computational methods still lack some important aspects, as they can only identify Kcr sites on either histone-only or combined histone and nonhistone proteins. Although a tool was developed to identify Kcr sites on non-histone proteins only, its performance is inadequate and the exploration of hidden Kcr patterns (motifs) has been completely ignored, which might be significant for detailed Kcr studies. Therefore, algorithms that can more effectively predict Kcr sites on non-histone proteins with their biological meaning need to be designed. Accordingly, we developed a novel deep learning (capsule network)-based model, named CapsNh-Kcr, for Kcr site prediction, particularly focusing on non-histone proteins. Based on the independent results, the proposed model achieves an AUC of 0.9120, which is approximately 6% higher than that of previous nhKcr model in the prediction of Kcr sites on non-histone proteins. Further, we revealed, for the first time, that the proposed model can represent obvious motif distribution across Kcr sites in non-histone proteins. The source code (in Python) is publicly available at https://github.com/Jhabindra-bioinfo/CapsNh-Kcr.

1. Introduction

Lysine crotonylation (Kcr) is an important type of PTM widely detected in histone and non-histone proteins, and was initially identified in histone proteins through a comprehensive analysis of proteins in human somatic and mouse germ cells [1]. According to the study, crotonylation is enriched in the enhancer and promoter regions in both germinal and human somatic cells, indicating that histone Kcr may play a key role as an indicator of gene expression [1]. Kcr plays a crucial role in diverse diseases and biological processes, such as gene transcription regulation, spermatogenesis, tissue injury, inflammation, carcinogenesis, neuropsychiatric disease, telomere maintenance, cancer, and HIV latency [2–11]. Similarly, crotonylated non-histone proteins have been detected and described in recent studies [12–15], and were found to be involved in cellular organization, cell cycling, and cellular organization [15]. Therefore, accurate prediction of Kcr sites and detection of their patterns across Kcr sites are important for understanding the regulation of proteins in human biology. Advanced experimental technologies, such as high-performance liquid chromatography-tandem mass spectrometry, stable isotope labeling by amino acids in cell culture labeling, affinity enrichment, and specific antibodies, are popular experimental methods for predicting and detecting Kcr sites [16]. Although such advanced proteomics technologies can detect Kcr sites directly and effectively, they are labor-intensive and expensive. Therefore, modern artificial intelligence (AI)-based algorithms have been considered as alternative approaches for Kcr site identification. Over the past

* Corresponding authors at: School of International Engineering and Science, Jeonbuk National University, Jeonju 54896, South Korea (H. Tayara); Department of Electronics and Information Engineering, Jeonbuk National University, Jeonju 54896, South Korea (K.T. Chong).

E-mail addresses: hilaltayara@jbnu.ac.kr (H. Tayara), kitchong@jbnu.ac.kr (K.T. Chong).

https://doi.org/10.1016/j.csbj.2022.11.056

few years, several computational methods have been developed to identify Kcr sites based on amino acid sequences. Most of the tools use conventional machine learning (CML) with a small number of samples for prediction either on histone or mixed histone and non-histone proteins, whereas very few tools are available that utilize a large number of datasets either on mixed histone and non-histone or non-histone only, using deep learning (DL) methods. Detailed information on these models is presented in Table 1.

As mentioned in Table 1, many of the computational models are based on conventional machine learning (ML), which utilizes a small number of training samples for histone proteins. Some deep learning methods are used with a large number of training samples for histone and non-histone proteins. Very recently, in 2021, a bioinformatics tool named nhKcr [28] used a larger number of non-histone protein samples to train the DL method in which CNN was used as a classifier. Although the performance is acceptable, the visualization of deeper layers of the CNN is difficult because the pooling operation loses the information for several locations in a given dataset. Such complexities can be replaced by a CapsNet [29] strategy by replacing scalar-output features with vector-output and max-pooling operations with a routing process [29].

Although the ML and DL approaches showed good performance in the prediction of Kcr sites on histone, mixed histone, and non-histone proteins, improvements can be made from a certain perspective. (i) The aforementioned ML methods rely on manual feature extraction approaches; however, such difficulties can be solved using DL methods [30]. (ii) With improvements in several experimental technologies, data production for peptide mapping in histones and non-histones has significantly improved. However, ML models have not accepted this critical issue and lack systematic investigation and assessment of available features [25,27]. (iii) In contrast, DL-based models such as Deep-Kcr [25], BERT-Kcr [26], and DeepCap-Kcr [27] adopt an automatic feature selection approach; however, these tools are designed particularly for predicting Kcr sites in mixed histone and non-histone proteins. Predictive tools for Kcr sites on mixed histone and non-histone proteins might not provide accurate information of Kcr sites if the data cover only histone or non-histone proteins. To address this issue, the nhKcr tool [28] was recently developed using a large number of datasets derived from non-histone proteins only. However, the tool still disremembers significant issues, such as prediction performance and extraction of biological meaning (motif detection) from a large amount of data, which is significant for several biological and research tasks. The tool, nhKcr [28], utilized a CNN for feature engineering and final prediction of Kcr sites; however, the pooling operation used in CNN forces the loss of spatial information in a given peptide sequence, which causes the problem of finding accurate motifs across the Kcr sites [31].

To address the above-mentioned drawbacks of the existing computational models for identifying Kcr sites, we proposed a novel DL model based on Capsule Network (CapsNets), which was motivated by previous CapsNets-based bioinformatics studies [32–35]. Our proposed model utilizes the CapsNet strategy, which consists of two key concepts. The first concept is the initial feature extraction using CNN layers. The second major concept involves the addition of capsules to a CNN that serve as hierarchical relationships of features to enable the use of the model to increase learning efficiency. In this step, we apply dynamic routing between capsules (number of neurons) instead of the max pooling operation, as in the traditional CNN. This concept not only provides accurate biological (motif) information across Kcr sites, but also a strong discriminant power in distinguishing between classes (Kcr and non-Kcr).

2. Material and methods

2.1. Dataset

For training and independent testing, the proposed CapsNh-Kcr model, an experimentally verified dataset, was adopted from a recent previous study, nhKcr [28]. A total of 19,287 Kcr sites were identified in 4,230 human nonhistone proteins across HeLa, lung, A549, and HCT116 cells [12–16]. The CD HIT [36] software was used to remove the sequence redundancy with a threshold of 30 %. Finally, 15,603 positive (Kcr site-containing sequences) sequences and 164,709 negative (non-Kcr site-containing sequences) samples were obtained. The 15,603 negative samples were randomly selected from 164,709 negative dataset. To train and test our proposed model, 12,262 positive and 12,262 negative

Table 1
Characteristics of the existing computational methods and tools for Kcr site prediction.

Methods/Tools	ML category	Models/Classifiers	Year	Protein type	Number of training samples
CrotPred [17]	CML	Discrete Hidden Markov Model (DHMM)	2015	Histone	169 Kcr 847 non-Kcr
Position-weight [18]	CML	Support Vector Machine (SVM)	2017	Histone	169 Kcr 847 non-Kcr
CKSAAP CrotSite [19]	CML	SVM	2017	Histone	169 Kcr 847 non-Kcr
iKcr-PseEns [20]	CML	Random Forest (RF)	2018	Histone	169 Kcr 847 non-Kcr
iCrotoK-PseAAC [21]	CML	Artificial Neural Network (ANN)	2019	Histone	169 Kcr 847 non-Kcr
LightGBM-CroSite [22]	CML	LightGBM	2020	Histone	159 Kcr 847 non-Kcr
Rulan et.al [23]	CML	SVM, RF	2020	Histone	167 Kcr 388 non-Kcr
predML-Site [24]	CML	SVM	2020	Histone	115 Kcr 6,279 non-Kcr
Deep-Kcr [25]	DL	Convolutional Neural Network (CNN)	2020	Mixed histone and non-histone	6,975 Kcr 6,975 non-Kcr
BERT-Kcr [26]	DL	BiLSTM	2021	Mixed histone and non-histone	6,975 Kcr 6,975 non-Kcr
DeepCap-Kcr [27]	DL	Capsule Network (CapsNets)	2021	Mixed histone and non-histone	6,975 Kcr 6,975 non-Kcr
nhKcr [28]	DL	CNN	2021	Non-histone	12,262 Kcr 60,101 non-Kcr

samples were used, while for the independent test, 3,341 positive and 3,341 negative sequence were used. Training and testing dataset are not overlaped. All samples used in this study share the window size of 29.

2.2. Overall framework of the proposed model

Deep learning-based strategies, such as CNNs, have made breakthroughs in many fields, including computer vision [37] and bioinformatics [38–40], and significantly outperformed many conventional curated feature extraction ML models. However, these strategies have some limitations, such as the invariance caused by pooling processes and the inability to identify spatial relationships between features [41]. To solve these problems, Sabour et al. proposed a novel deep learning theory widely known as the capsule network (CapsNet) [29,42]. The main idea of the CapsNet model is a capsule (a group of neurons) whose activity vector represents the instantiation parameters of a specific type of entity, such as an object or an object part [29]; this means that the length of the activity vector represents the probability that the entity exists, and the instantiation parameters are represented by its (activity vector's) orientation. When a lower level of capsules makes predictions and agrees multiple times, a higher level of capsules becomes active. Thus, a lower-level capsule prefers to send its prediction (output) to a higher-level capsule. A detailed theoretical explanation and the working principle of a CNN-based feed-forward CapsNet were described by [27].

The simplified architecture of the proposed CapsNh-Kcr model is shown in Fig. 1, and is similar to the nature of the initial CapsNet proposed by Sabour et al. [29]. Our model consisted of three main layers: 1D convolutional (conv1D), PrimaryCaps (PrimaryCaps_Conv1D), and a fully connected layer (KcrCaps). As shown in Fig. 1, in the initial step, all amino acid sequences are encoded with widely used binary or one-hot encoding schemes to be fed into a CNN layer for initial feature extraction from a given raw sequence. In the next step, core layers named PrimaryCaps and KcrCaps are used for further feature abstraction. In the first step, the CNN layer is designed to increase the prediction power of the proposed CapsNet.

The hyperparameters of the model were tuned using a grid-search method. The tuned hyperparameters of the model included the number of layers, number of filters, filter size, and dropout rate. From the hyperparameter tuning method, one convolution layer was determined to be suitable before being fed into the PrimaryCaps, which also maintains 32 filters of size 7 (kernels) with a stride of one, and a ReLU as an activation function to update the weights. This initial layer was followed by a dropout layer at a rate of 0.7 to control the overfit. Another layer, PrimaryCaps, is based on a convolutional layer, which has 16 channels of convolutional capsules, and each capsule (272 capsules in total) consists of 8 convolutional units, each of which is the result of a size 7 1D convolutional kernel. Accordingly, PrimaryCaps has [17,16] 8D vector capsules, and each capsule in the [17,1] grid shares its weight with other capsules. These capsules were represented by probabilities. To scale the length (probability) of each capsule to [0 1], a nonlinear squash activation function was used [29]. Of note, the process of dynamic routing between capsules is used at this stage, or between PrimaryCaps and KcrCaps. The last layer, KcrCaps, had an 8D capsule in each of the two classes (positive and negative). In positive capsules, the Kcr sites were contained, while the negative capsules indicated the probability of non-Kcr sites. Finally, L2 norms were used to rescale the positive and negative output vectors of capsules resulting from KcrCaps.

2.3. Software and model training

To build the model, an open-source Python library Keras (https://keras.io/) using TensorFlow backend was used. Python version 3.7.4 and Keras version 2.2.4 were used to train the proposed model. To train the model, we used the widely applied K-fold cross-validation (CV) method [43], where k was tuned to 5. The final results were obtained from an averaged 5-fold. To guide overfitting during model training, an early stopping method was applied [44]. In our case, this method was applied when the generalization loss increased over 10 successive epochs. The learning rate and batch size were set to 0.001 and 128, respectively, and the optimizer was Adam [45]. The binary cross-entropy was utilized as a loss function [46]. The total number of trainable parameters of the model was 68,128.

2.4. Evaluation parameters

Four metrics can be employed to measure the performance of the proposed model: the Matthew correlation coefficient (MCC), accuracy (ACC), specificity (SP), and sensitivity (SN). The numerical expression of these metrics are given in Eqs. (1)–(4) are widely

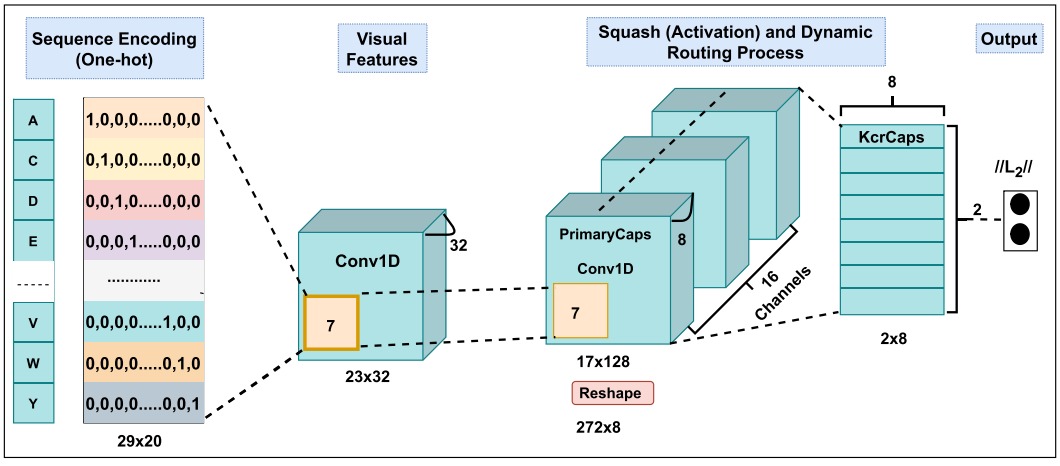

Fig. 1. Overview of the proposed model, CapsNh-Kcr.

used in binary classification problems in many bioinformatics tasks [47,48].

$$Sn = \frac{TP}{TP + FN} \quad (1)$$

$$Sp = \frac{TN}{TN + FP} \quad (2)$$

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (4)$$

where TP, FP, TN, and FN are the true positive, false positive, true negative, and false negative values, respectively. In this study, the area under the receiver operating characteristics curve (AUC-ROC) matrix was used to evaluate the overall performance quality of the model. A higher AUC value indicates better model performance.

3. Results

3.1. Performance of the model using the training data

The proposed model was evaluated using 12,262 balanced training samples and the same number of negative samples, indicating that the model learned from balanced data. Our primary focus was not only to distinguish the Kcr sites from the given large number of non-Kcr samples, as in the previous model Nh-Kcr ([28]), but also to identify a possible biological significance from the balanced data; this is because the capsule network does not require a large number of samples to learn the information from the given sequence, as required by other DL models, such as CNN [27,29]. Therefore, we used balanced data for the prediction and investigation of the Kcr sites. For the training data, the model achieved Sn = 0.8794, Sp = 0.8817, Acc = 0.8806, and MCC = 0.7611. These results were obtained from the averaged 5-fold CV. The ROC curve and its corresponding AUC values in five-fold CV are shown in Fig. 2 (a). The obtained average AUC was 0.90, with standard deviation of 0.01, which is approximately 2% greater than the existing model Nh-Kcr [28]. The main reason for the model's higher performance is due to the accurate feature vectors for Kcr sites being captured by the CapsNets strategy; evidently, these final 8D capsule vectors were captured in KcrCaps layer. To depict this visually, the features captured by the first layer (Conv1D) and the last layer (KcrCaps) were computed using t-distributed stochastic neighbor embedding (t-SNE) [49] in the Scikit-Python (https://scikit-learn.org) library, as displayed in Fig. 2 (b) and (c), respectively, where the red and blue circles symbolize Kcrs and non-Kcrs, respectively. This demonstration shows that KcrCaps learns robust features compared to the CNN used in the first layer.

3.2. Model performance using unseen data and comparison with a previous model

To confirm whether CapsNh-Kcr could distinguish Kcr and non-Kcr from unseen (independent) data, the model was run with a dataset that contained 3,341 positive and 3,341 negative sequences; the model obtained Sn = 0.8834, Sp = 0.8764, Acc = 0.8799, and MCC = 0.7597. To demonstrate the superiority of our model on the balanced dataset, a comparison with the previous model, nhKcr [28], was performed. This model and our model utilize the same window size as the data for non-histone proteins. Although the developers of nhKcr provided the final results, the results were based on an unbalanced dataset. Therefore, in this study, we recomputed the previous model based on the information (source code and web server) provided by the authors and ran their model on a balanced independent dataset. A comparison of our model with nhKcr is presented in Table 2 and Fig. 3. As

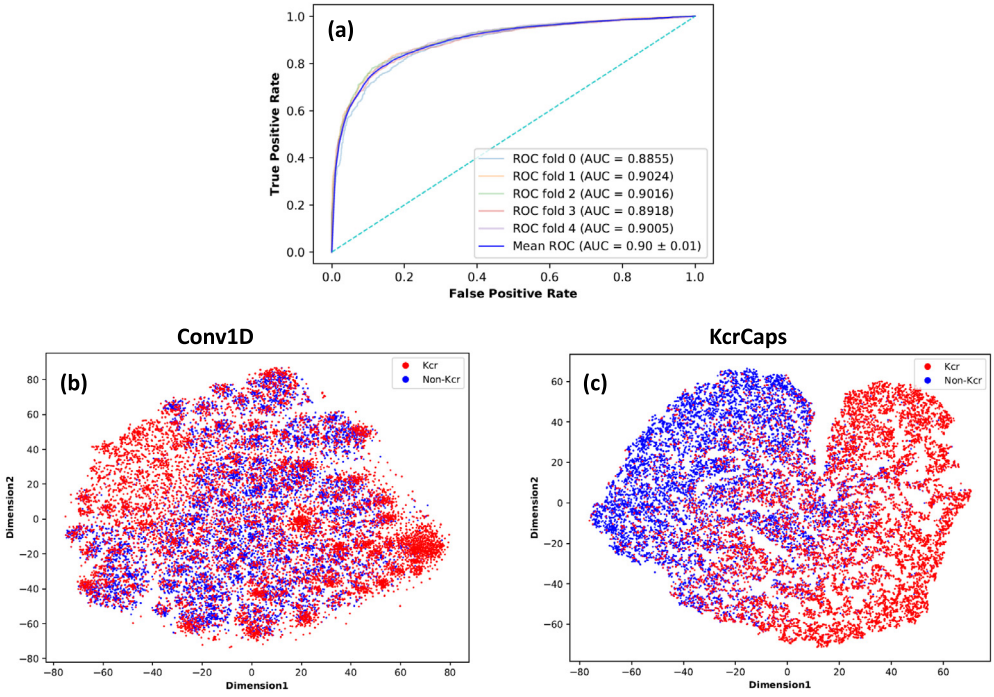

Fig. 2. Model performance in training; a) ROC curves and their corresponding AUC values in 5-fold CV. b) and c), t-SNE visualization of the Conv1D and KcrCaps layer, respectively.

Table 2
Comparison of nhKcr and CapsNh-Kcr in terms of five major metrices on an independent dataset.

Models	Acc	Sn	Sp	MCC	AUC
nhkcr	0.7606	0.7848	0.7365	0.7366	0.8501
CapsNh-Kcr	**0.8799**	**0.8834**	**0.8764**	**0.7597**	**0.9120**

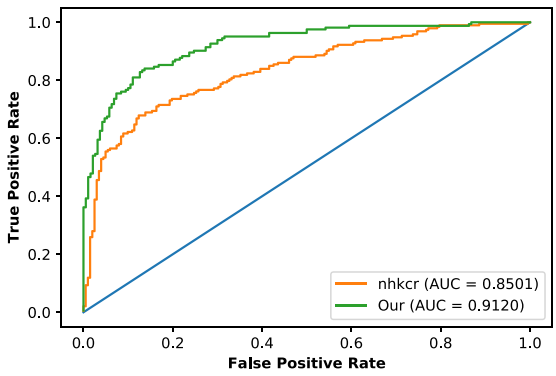

Fig. 3. Comparison of the ROC curves and AUC values between nhKcr and our model on independent dataset; the image demonstrates that our model is accurate in the prediction of Kcr-sites in non-histone proteins.

shown in Table 2, CapsNh-Kcr outperformed their model in terms of all metrics, that is, Acc, Sn, Sp, MCC, and AUC. Particularly, the proposed model obtained an AUC of 0.9120, which was approximately 6 % higher than that of nhKcr model in the prediction of Kcr sites on non-histone proteins.

Of note, for fair comparison and practical use of our model, we did not compare our model with the aforementioned (in Table 1) existing computational models, except nhKcr, because other models are particularly designed to identify Kcr sites on either histone proteins or mixed histone and non-histone proteins..

We computed the t-SNE to visualize the layers to ensure that KcrCaps captures the robust feature to distinguish between Kcrs and non-Kcrs on an unseen (independent) dataset. A visual representation of the t-SNE computation in the 2D plot is shown in Fig. 4. The images demonstrate that the Kcr sites are more clearly visible using features learned by KcrCaps, even though few sequences overlap compared to the features captured in the first layer (Conv1D).

3.3. Learnt motifs

Another main goal of this computational task is to interpret the trained model that might have various hidden or learned information that can explore Kcr information more clearly in non-histone proteins. To the best of our knowledge, this computational model is the first to be designed to extract clear biological significance (motifs) across Kcr sites in non-histone proteins. First, we generated the sequence pattern captured by the first-layer (Conv1D) filters. The layer comprises a set of filters, each of which can be assumed to be a position weight meter. Each filter (in total, 32 filters) across the input sequences encoded by one-hot encoding outputs a non-linear similarity score at each position, known as a feature map. These feature patterns can be effectively interpreted. Because we used a filter size of seven, the motifs (amino acid sequence patterns) were captured with a length of seven. We converted the generated PWM to sequence logos (motifs) [50]. A visual representation of the motifs captured by filters 1, 2, 7, 8, 13, 16, 17, 18, 19, 21, 22, and 31 in Conv1D is shown in Fig. 5.

Second, to investigate the features captured by individual capsules in PrimaryCaps, sequence logos were generated according to each capsule's response to the input amino acid sequences. We fed all the peptides through all the capsules in PrimaryCaps (272 capsules in total), and aligned the peptides in response to a positive capsule in KcrCaps with a capsule length greater than a threshold of 0.5. Of note, the peptides aligned only positive capsules in KcrCaps responsible for Kcr sites information's. Thus, we generated PWMs for these aligned peptides to transform them into sequence motifs (logos); the visual representation of these motifs is depicted in Fig. 5 (b). Not all capsules had values greater than 0.5. Of the 272 capsules, only 187 provided greater than or equal to 0.5. We then compared the learned motifs of each capsule in KcrCaps with the ground-truth motifs KExxxK, EKxxxxxK, and KxxxEK. Several studies have claimed that these ground-truth motifs are identified as significantly overrepresented hotspots for Kcr sites [13,15,27,28]. As shown in Fig. 6, the motifs learned by capsules 45, 101, 109, 170, 245, and 254 agree with the ground-

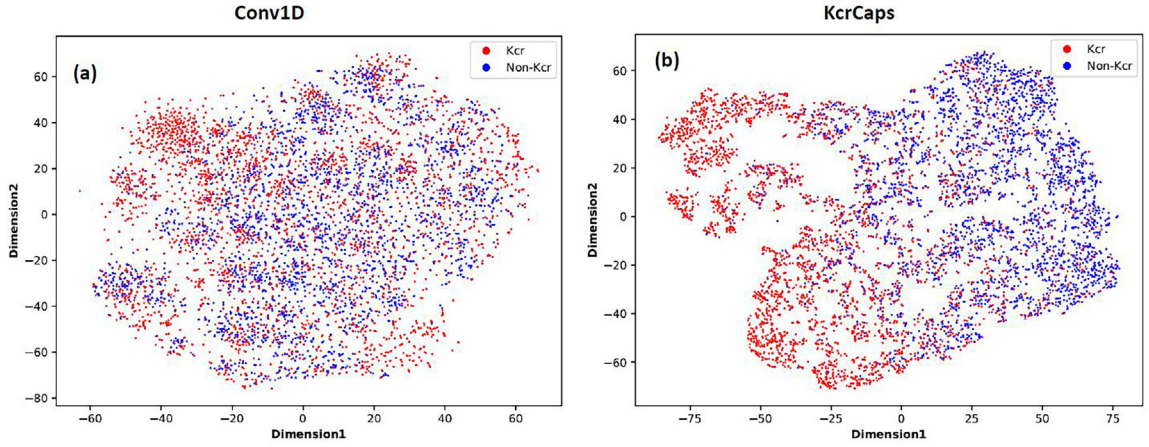

Fig. 4. t-SNE visualization of layers of the proposed model on the independent dataset: a) Conv1D and b) KcrCaps.

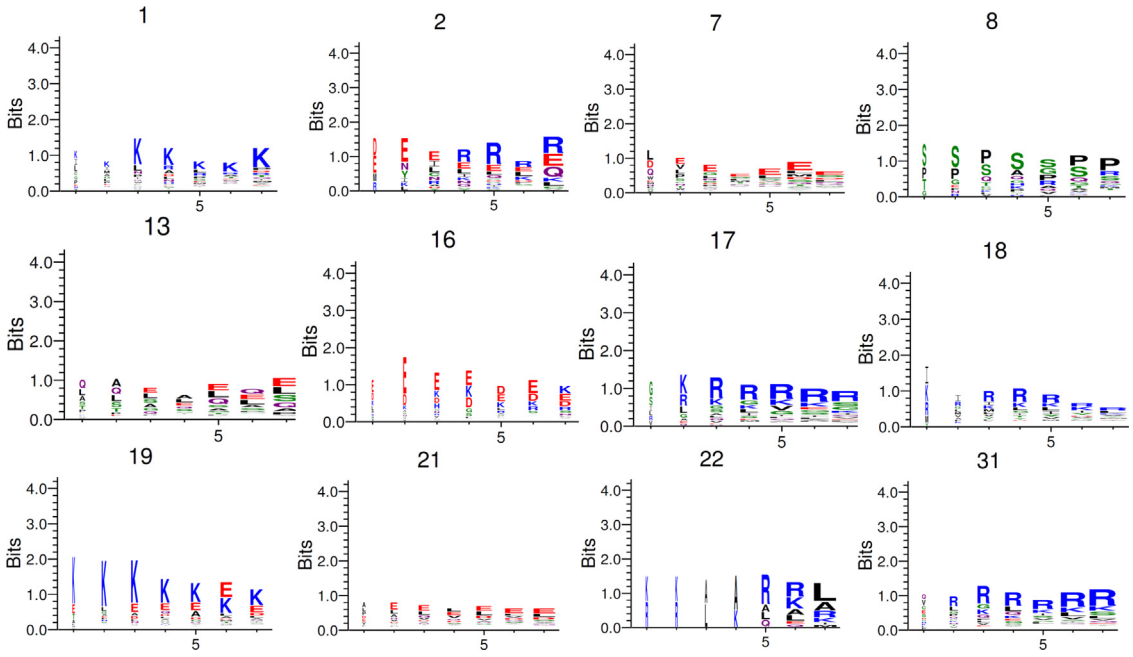

Fig. 5. Sequence motif captured by some filters in the first layer (Conv1D) of the proposed model, shows that the amino acids K, E, and R are frequently overrepresented. Note the underrepresented (lower case are ignored when plotting the sequence logo).

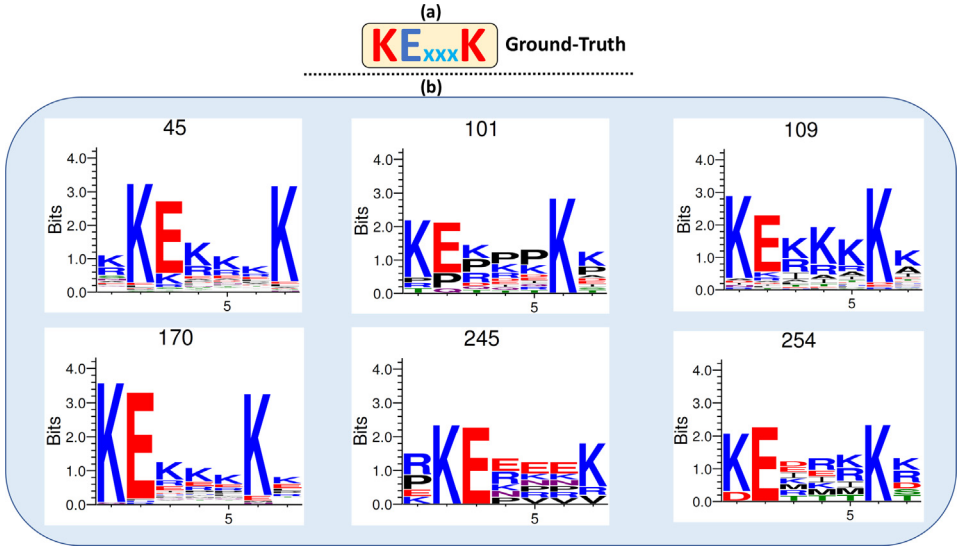

Fig. 6. Comparison of the motifs learnt by our models with ground-truth motifs of Kcr sites in proteins. (a) ground-truth motifs, (b) motifs captures by capsules (45, 101, 109, 170, 245, and 254) in KcrCaps, compared to ground-truth, showing that KcrCaps captured obvious motifs for the representation of Kcr sites.

truth (KExxxK in Fig. 6 (a)) motifs. Further, the proposed model can learn the sequence patterns (motifs) effectively.

4. Discussion

Although capsule networks are still in the development phase, this strategy solves various bioinformatics tasks [32–35]. In this study, we identified some important benefits of using a capsule network. The capsule strategy is better suited for feature learning for Kcr site prediction in nonhistone proteins compared to the CNN used in the previous study with nhKcr. Second, the motifs converted by the internal capsules were very obvious when comparing ground-truth motifs, revealing its effectiveness for learning motifs across Kcr sites. However, it is not easy to represent clear motifs in other DL models, such as CNN. In a CNN, the max-pooling operation influences the ability to build hierarchical motif representations [31]; this is due to the max pooling loss of the spatial information of the features and the inability to interact one feature with another feature. In such cases, capturing motifs from deeper layers in the CNN is complicated.

The trained model is effective for interpretation, analogous to *in silico* statistical mutagenesis analysis. Data contain many features, and mutations can occur in a specific feature without varying the remaining features. Mutation is observed at the output of the networks by taking the absolute difference between the predictions of the mutated and reference sequences. For this analysis, the impact of the mutation is shown in the heat map in Fig. 7; this shows the

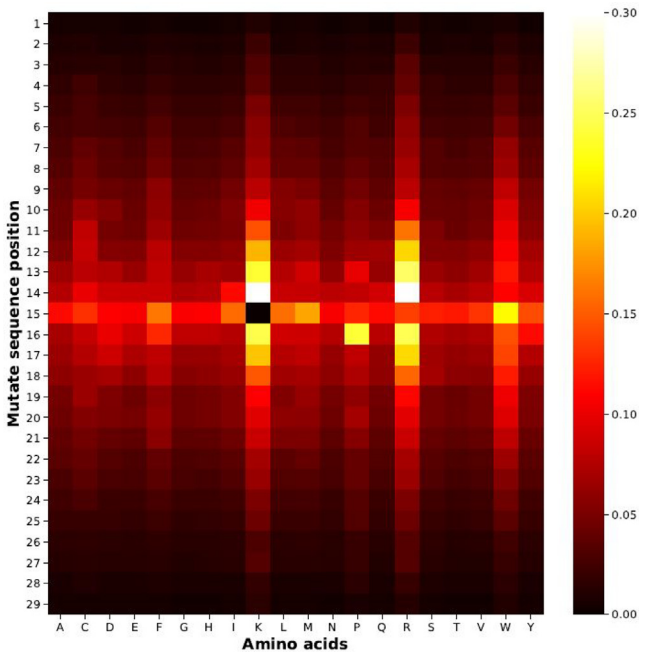

Fig. 7. Heatmap Visualization of *in silico* mutation.

mutation visualization in the independent dataset. The mutation at positions 11 to 18 changed the prediction probability by approximately 10 %, and position 15 had the highest effect. Moreover, the mutations to K, F, R, and W almost changed the prediction probability by approximately 10 %.

5. Conclusion

In this study, we proposed a DL model that can identify Kcr sites in non-histone proteins. The proposed model adopts a CNN-based CapsNet strategy. The results revealed that the strategy used in our model outperformed the previous model, nhKcr, which was based on the well-known DL model CNN. From a biological perspective, the model showed effective properties that can explore the internal data distribution across given amino acid sequences, and could capture the obvious motif across Kcr sites in a given dataset. Specially, this study proved that the biological significance such as sequence motifs in protein sequences can be sought effectively through our applied strategy. Further research is warranted, although the binary cross-entropy as a loss function used in this study helped in improving the performance, different loss functions can be tested. Even though our study claims the Kcr patterns in non-histone proteins computationally, a comprehensive study through a wet lab experiment is also recommended to map these newly discovered Kcr sites. As a final point, we believe that the proposed model and strategy would be beneficial for other bioinformatics tasks and research.

Data and code availability

The source code and data are available at https://github.com/Jhabindra-bioinfo/CapsNh-Kcr.

Funding

The authors thank the anonymous reviewers for their valuable suggestions. This work was supported in part by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2020R1A2C2005612) and in part by Universities leading lab-specific start-ups through the Commercializations Promotion Agency for R&D Outcomes (COMPA) grant funded by the Korea government(MSIT) (No.startuplab22-016).

CRediT authorship contribution statement

Jhabindra Khanal: Conceptualization, Data curation, Formal analysis, Methodology, Software, Validation, Visualization, Investigation, Writing – original draft, Writing – review & editing. **Jeevan Kandel:** Conceptualization, Formal analysis, Software, Validation, Investigation, Writing – review & editing. **Hilal Tayara:** Conceptualization, Writing – original draft, Writing – review & editing. **Kil To Chong:** Writing – original draft, Writing – review & editing, Project administration, Supervision, Resources, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] M. T, H. L, S. L, et al. Identification of 67 histone marks and histone lysine crotonylation as a new type of histone modification. Cell 2011;146(6):1016–1028.
[2] Jiang G, Li C, Lu M, Lu K, Li H. Protein lysine crotonylation: past, present, perspective. Cell Death Disease 2021;12(7):1–11.
[3] Wei W, Liu X, Chen J, Gao S, Lu L, Zhang H, Ding G, Wang Z, Chen Z, Shi T, 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(7):898–915.
[4] Li K, Wang Z. Histone crotonylation-centric gene regulation. Epigenetics Chromatin 2021;14(1):1–6.
[5] Liu X, Wei W, Liu Y, Yang X, Wu J, Zhang Y, Zhang Q, Shi T, Du JX, Zhao Y, et al. Mof as an evolutionarily conserved histone crotonyltransferase and transcriptional activation by histone acetyltransferase-deficient and crotonyltransferase-competent cbp/p300. Cell Discov 2017;3(1):1–17.
[6] Sabari BR, Tang Z, Huang H, Yong-Gonzalez V, Molina H, Kong HE, Dai L, Shimada M, Cross JR, Zhao Y, et al. Intracellular crotonyl-coa stimulates transcription through p300-catalyzed histone crotonylation. Mol Cell 2015;58(2):203–15.
[7] Berger K, Moeller MJ. Mechanisms of epithelial repair and regeneration after acute kidney injury. Seminars in nephrology, vol. 34. Elsevier; 2014. p. 394–403.
[8] Liu S, Yu H, Liu Y, Liu X, Zhang Y, Bu C, Yuan S, Chen Z, Xie G, Li W, et al. Chromodomain protein cdyl acts as a crotonyl-coa hydratase to regulate histone crotonylation and spermatogenesis. Mol Cell 2017;67(5):853–66.
[9] Fu H, Tian C-L, Ye X, Sheng X, Wang H, Liu Y, Liu L. Dynamics of telomere rejuvenation during chemical induction to pluripotent stem cells. Stem Cell Rep 2018;11(1):70–87.
[10] Jiang G, Nguyen D, Archin NM, Yukl SA, Méndez-Lagares G, Tang Y, Elsheikh MM, Thompson GR, Hartigan-O'Connor DJ, Margolis DM, et al. Hiv latency is reversed by acss2-driven histone crotonylation. J Clin Invest 2018;128(3):1190–8.
[11] Wan J, Liu H, Ming L. Lysine crotonylation is involved in hepatocellular carcinoma progression. Biomed Pharmacother 2019;111:976–82.
[12] Wei W, Mao A, Tang B, Zeng Q, Gao S, Liu X, Lu L, Li W, Du JX, Li J, et al. Large-scale identification of protein crotonylation reveals its role in multiple cellular functions. J Proteome Res 2017;16(4):1743–52.
[13] H. H, D.-L. W, Y. Z. Quantitative crotonylome analysis expands the roles of p300 in the regulation of lysine crotonylation pathway. Proteomics 2018;18(15):1700230.
[14] Q. W, W. Li, C. W, et al. Ultradeep lysine crotonylome reveals the crotonylation enhancement on both histones and nonhistone proteins by saha treatment. J Proteome Res 2017;16(10):3664–3671.
[15] W. X, J. W, J. Z, et al. Global profiling of crotonylation on non-histone proteins. Cell Res 2017;27(7):946–949.
[16] H. Y, C. B, Y. L, et al. Global crotonylome reveals cdyl-regulated rpa1 crotonylation in homologous recombination–mediated dna repair. Sci Adv 2020;6(11):eaay4697.
[17] G. H, W. Z. A discrete hidden markov model for detecting histone crotonyllysine sites. Match Commun Math Comput Chem 2016;75:717–730.
[18] W.-R. Q, B.-Q. S, H. T, et al. Identify and analysis crotonylation sites in histone by using support vector machines. Artif Intell Med 2017;83:75–81.

[19] Z. J, J.-J. H. Prediction of lysine crotonylation sites by incorporating the composition of k-spaced amino acid pairs into chou's general pseaac. J Mol Graphics Model 2017;77:200–204.
[20] W.-R. Q, B.-Q. S, X. X, et al. ikcr-pseens: Identify lysine crotonylation sites in histone proteins with pseudo components and ensemble classifier. Genomics 2018;110(5):239–246.
[21] S. JM, M. RSU, Y.D. K. icrotok-pseaac: Identify lysine crotonylation sites by blending position relative statistical features according to the chou's 5-step rule. PloS one 2019;14(11):e0223993.
[22] Y. L, Z. Y, C. C, et al. Prediction of protein crotonylation sites through lightgbm classifier based on smote and elastic net. Anal Biochem 2020;609:113903.
[23] R. W, Z. W, H. W, et al. Characterization and identification of lysine crotonylation sites based on machine learning method on both plant and mammalian. Scientific Rep 202;10(1):1–12.
[24] S. A, A. R, M.A.M. H, et al. predml-site: Predicting multiple lysine ptm sites with optimal feature representation and data imbalance minimization. IEEE/ACM Trans Comput Biol Bioinf.
[25] H. L, F.-Y. D, Z.-X. G, et al. Deep-kcr: accurate detection of lysine crotonylation sites using deep learning method. Brief Bioinf.
[26] Qiao Y, Zhu X, Gong H. Bert-kcr: prediction of lysine crotonylation sites by a transfer learning method with pre-trained bert models. Bioinformatics.
[27] Khanal J, Tayara H, Zou Q, To Chong K. Deepcap-kcr: accurate identification and investigation of protein lysine crotonylation sites based on capsule network. Briefings Bioinf.
[28] Y.-Z. C, Z.-Z. W, Y. W, et al. nhkcr: a new bioinformatics tool for predicting crotonylation sites on human nonhistone proteins based on deep learning. Briefings Bioinf.
[29] S. S, N. F, G.E. H. Dynamic routing between capsules. arXiv preprint arXiv:1710.09829.
[30] J. Z, M. H, A. A, P. M, et al. A primer on deep learning in genomics. Nat Genet 2019;51(1):12–18.
[31] Koo PK, Eddy SR. Representation learning of genomic sequence motifs with convolutional neural networks. PLoS Comput Biol 2019;15(12):e1007560.
[32] D. W, Y. L, D. X. Capsule network for protein post-translational modification site prediction. Bioinformatics 2019;35(14):2386–2394.
[33] Y. Z, F. L, D. Xi, et al. Computational identification of eukaryotic promoters based on cascaded deep capsule neural networks. Briefings Bioinf 2021;22(4): bbaa299.
[34] Y. L, B. L, B. J. A deep learning method for motor fault diagnosis based on a capsule network with gate-structure dilated convolutions. Neural Comput Appl 2021;33:1401–1418.
[35] W. D, Y. S, G. L, et al. Capsnet-ssp: multilane capsule network for predicting human saliva-secretory proteins. BMC Bioinf 2020;21(1):1–17.
[36] Huang Y, Niu B, Gao Y, Fu L, Li W. Cd-hit suite: a web server for clustering and comparing biological sequences. Bioinformatics 2010;26(5):680–2.
[37] L. L, Y. Y, F. H, et al. Integrating local cnn and global cnn for script identification in natural scene images. IEEE Access 2019;7:52669–52679.
[38] X. C, W. H, Z. C, et al. Pssp-mvirt: peptide secondary structure prediction based on a multi-view deep learning architecture. Briefings Bioinf.
[39] J. K, I. N, H. T, et al. 4mccnn: Identification of n4-methylcytosine sites in prokaryotes using convolutional neural network. IEEE Access 2019;7:145455–145461.
[40] J. K, H. T, K. TC. Identifying enhancers and their strength by the integration of word embedding and convolution neural network. IEEE Access 2020;8:58369–58376.
[41] Z. D, S. L. Research on image classification based on capsnet. In 2019 IEEE 4th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), vol. 1, IEEE; 2019. pp. 1023–1026.
[42] G. HE, S. S, N. F. Matrix capsules with em routing, in: International conference on learning representations; 2018.
[43] J. K, H. T, Q. Z, et al. Identifying dna n4-methylcytosine sites in the rosaceae genome with a deep learning model relying on distributed feature representation. Comput Struct Biotechnol J 2021;19:1612–1619.
[44] L. P. Early stopping-but when?. In Neural Networks: Tricks of the trade, Springer; 1998. pp. 55–69.
[45] Kingma DP, Ba J. Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980.
[46] De Boer P-T, Kroese DP, Mannor S, Rubinstein RY. A tutorial on the cross-entropy method. Ann Oper Res 2005;134(1):19–67.
[47] J. K, D. L, H. T, et al. i6ma-stack: a stacking ensemble-based computational prediction of dna n6-methyladenine (6ma) sites in the rosaceae genome. Genomics 2021;113(1):582–592.
[48] D. L, J. K, H. T, K. TC. ienhancer-rf: Identifying enhancers and their strength by enhanced feature representation using random forest. Chemometrics Intell Lab Syst 2021;212:104284.
[49] Vdm L. Accelerating t-sne using tree-based algorithms. J Mach Learn Res 2014;15(1):3221–45.
[50] M.C. T, M. N. Seq2logo: a method for construction and visualization of amino acid binding motifs and sequence profiles including sequence weighting, pseudo counts and two-sided representation of amino acid enrichment and depletion. Nucl Acids Res 2012;40(W1):W281–W287.