The Identifications of Post Translational Modification Sites with Capsule Network

Baitong Chen[1], Yujian Gu[2], Bin Yang[3(✉)], and Wenzheng Bao[2]

[1] Xuzhou No. 1 People's Hospital, Xuzhou 221000, China
[2] School of Information Engineering, Xuzhou University of Technology, Xuzhou 221000, China
[3] School of Information Science and Engineering, Zaozhuang University, Zaozhuang 277160, China

Abstract. Post-translational modification (PTM) is considered a significant biological process with a tremendous impact on the function of proteins in both eukaryotes, and prokaryotes cells. Malonylation of lysine is a newly discovered post-translational modification, which is associated with many diseases, such as type 2 diabetes and different types of cancer. In addition, compared with the experimental identification of propionylation sites, the calculation method can save time and reduce cost. In this paper, we combine principal component analysis with support vector machine (SVM) to propose a new computational model - Mal-prec (malonylation prediction). Firstly, the one-hot encoding, physicochemical properties and the composition of k-spacer acid pairs were used to extract sequence features. Secondly, we preprocess the data, select the best feature subset by principal component analysis (PCA), and predict the malonylation sites by SVM. And then, we do a five-fold cross validation, and the results show that compared with other methods, Mal-prec can get better prediction performance. In the 10-fold cross validation of independent data sets, AUC (area under receiver operating characteristic curve) analysis has reached 96.39%. Mal-pred is used to identify the malonylation sites in the protein sequence, which is a computationally reliable method. It is superior to the existing prediction tools that found in the literature and can be used as a useful tool for identifying and discovering novel malonylation sites in human proteins.

Keywords: Post translational modification · Malonylation · One-hot encoding · Principal component analysis · Support vector machine

1 Introduction

Post translational modifications play vital roles not only during biological processes but in various cell functions. They also work in the regulation of cellular plasticity and dynamics [1]. What's more, lysine is one of the most heavily modified residues of the 20 kinds of natural amino acids in proteins. Lysine is one of the essential amino acids for human beings and mammals [2]. The body cannot synthesize it by itself, so it must be supplemented from food. Lysine mainly exists in animal food and beans, but

D. Jiang and H. Song (Eds.): SIMUtools 2021, LNICST 424, pp. 546–554, 2022.
https://doi.org/10.1007/978-3-030-97124-3_42

lysine content in cereal food is very low. Lysine has positive nutritional significance in promoting human growth and development, enhancing body immunity, anti-virus, promoting fat oxidation, relieving anxiety and so on [3, 4]. At the same time, it can also promote the absorption of some nutrients, and cooperate with some nutrients to better play the physiological functions of various nutrients [5]. Recent studies have found multiple types of new protein lysine acylations, which have greatly deepened our understanding of the post translational modification sites of lysine. The structurally similarity of the three types of acidic lysine modifications and the potential to regulate different types of proteins in different pathways are determined since malony1, succiny1 and glutary1 groups have a negatively charged carboxy1 group. It is also proved that malonylation, succinylation, and glutarylation of lysine residues are deeply concerned with evolution and dynamic under various biological and cellular conditions, including stress response, genetic mutations and more [4, 6–8].

Because of the great effects of post translational modifications, distinguishing them from various modifications is quite necessary [9–11]. However, the huge amounts of features as well as simples make it really hard to distinguish between the modifications which are post translation or not.

In this paper, we do some efforts to solve the recognition of post translational modifications. Since our dataset consisted of weak classification features, we did a series of processing works, and try to improve its effect. What's more, we used the capsule to complete the classification. Since the dataset contains a lot of invalid features, various arrangements and combinations of different features are tried to discover the features which contribute more. After all, we deleted Interference terms, and the results witness the improvement of the final effect. We had additionally constructed LSTM classifier for comparing the predictive performance. Experimental results demonstrate that the model based on capsule and a series of processing works, which is used by us performs relatively better.

2 Methods and Materials

We first did some processing works to divide each features apart, and then reassembled them by the method of combination. After all, we got various new datasets which contains different features, and whose quantities of the features are diverse too. Then we did the classification through every single of the reassembled datasets, which aims to find out which features are beneficial to classification and which hinder the progress. In addition, we utilized AUC, AC to evaluate the performance of our models and the LSTM model utilized as a comparison.

2.1 Dataset

We used the database of numerical indices named AAindex representing various physicochemical and biochemical properties of amino acids [12–14]. The data consists of three sections: (1) AAindex1 including 566 properties for the amino acid index of 20 numerical values; (2) AAindex2 containing amino acid mutation matrix, and (3) AAindex3 with

protein contact potentials. The data could be found at the following URL address https://www.genome.jp/aaindex/ [15–18]. In this paper, eight physical and chemical properties are used, which are hydrophilicity value, mean polarity, isoelectric point, refractivity, average flexibility indices, average volume of buried residue, transfer free energy to surface, and consensus normalized hydrophobicity. The length of each peptide is 17, so the physiochemical properties is 17 * 8-dimensional vector. The physical and chemical properties are shown in the table below (Table 1).

Table 1. Eight physicochemical properties.

Properties description	Reference
Hydrophilicity value	Hopp and woods
Mean polarity	Radzicka and wolfenden
Isoelectric point	Zimmerman et al.
Refractivity	Treece et al.
Average flexibility indices	Bhaskaran and Ponnuswamy
Average volume of buries residue	Chothia
Transfer free energy to surface	Bull and Breese
Consensus normalized hydrophobicity	Eisenberg

2.2 Reassemble Dataset

We divided each features apart from the dataset, and pair the post translational modifications with the none translational modifications. Since there are 1735 post translational modifications and 1735 none translational modifications, we get eight 1735 * 17 datasets for each single feature, and then reassemble them by the law of combination. At last, we get 28 different datasets consisted of 2 of the 8 features, 56 new datasets composed of 3 of the 8 features, 56 different datasets composed of 3 of the 8 features, 70 datasets made up of 4 of the 8 features, 56 datasets consisted of 5 of the 8 features, 28 datasets made up of 6 of the 8 features, 8 datasets made up of 7 of the 8 features and a dataset made up of 8 of the 8 features. In a word, we get various new datasets which contains different features, and whose quantities of the features are diverse too.

2.3 Classifier Construction

Convolutional neural network is very successful and popular. However, it is not suitable for all tasks. Due to some defects in the architecture, it cannot complete some tasks well. CNN extracts features from data and recognizes objects through feature learning. The bottom layer of the network learns general features. With the deepening of layers, the extracted features are more complex. Then, the network uses all the features it

learns to make the final prediction. There is a drawback here: there is no available spatial information in CNN, and the pooling layer used for connection is actually very inefficient.

So, we used the capsule to avoid the problem above. The solution of capsule to the problem is to encode the spatial information and calculate the existence probability of objects. This can be represented by vector, the modulus of vector represents the probability of feature existence, and the direction of vector represents the attitude information of feature.

The working principle of capsule can be summarized into a sentence, that is, all the important information of the state of the feature in capsule detection will be encapsulated in the form of vector.

Capsule's network structure consists of parts named input layer, convolution layer, main capsule layer and digital capsule layer.

Take one sample from the dataset made up of 8 of the 8 features as an example. After the 8 * 17 sample scanned by 128 2 * 5 convolution kernels with 2 steps, we obtains a 4 * 9 * 128 feature map. This layer is a common convolutional neural network, and the next layer uses 8 groups of 2 * 2 * 16 convolution kernel with 2 steps convolution 8 times. Then, each feature map is expanded into one dimension, and the corresponding positions are combined. A total of 128 8-dimensional vector neurons, namely capsules, are obtained. Finally, the dynamic routing algorithm is used to get the digital capsule layer, and the module length of the digital capsule layer vector is the prediction result.

During the dynamic routing algorithm, it has 128 * 2 weights, and every weight is a 16 * 1 vector W_{ij}. And the capsules from the previous layer is u_{ij}. The update formula is as following:

$$u_{j|i} = W_{ij} * u_i + B_j \tag{1}$$

Then, we used the next formula to complete the vector compression. It is designed to combine the information of all capsules. C_{ij} is calculated by B_{ij} with softmax function.

$$s_j = \sum_i c_{ij} * u_{j|i} \tag{2}$$

The module length is compressed to 0–1 by the squashing function, the formula is as following:

$$v_j = squash\left(s_j\right) \tag{3}$$

$$squash(s) = \frac{||s||^2}{1 + ||s||^2} * \frac{s}{||s||} \tag{4}$$

And, the dynamic routing forward propagation has completed.

When it comes to loss function, it is constructed by 2 parts: the first part is the interval loss, and the second part uses the original 8 * 17 data minus the 8 * 17 data of reconstruction, Then square the result, and ride 0.005 at the same time.

2.4 Evaluation of the Predictor

In order to verify the reliability and stability of our model, we used 10-fold cross-validation to get the result. In this paper, we employ two evaluation indicators to evaluate the predictive performance of our proposed method, including accuracy (AC) and area under curve (AUC). Among them, AC reflects the model's ability to classify positive samples correctly. AUC is an evaluation index to measure the pros and cons of the binary classification model, which indicates the probability that the positive cases of prediction are in front of the negative ones. Their definitions are as follows:

$$AC = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

Where TP is the number correctly divided into positive samples, FP is the number incorrectly divided into positive samples, FN is the number incorrectly divided into negative samples, and TN is correctly divided into negative samples.

$$AUC = \frac{\sum_{i \in positive\_class} rank_i - \frac{M*(M+1)}{2}}{M * N} \tag{6}$$

Where $\sum_{i \in positive\_class}$ means add up the serial numbers of the positive samples, $rank_i$ represents the serial number of the i sample(The probability score is from small to large, ranking in the rank position), M, N is the number of positive samples and the number of negative samples.

3 Results and Discussions

3.1 Model Stability Analysis

K-fold cross-validation is widely utilized to compare the performance of different machine learning models on a specific dataset. The principle of k-fold cross-validation is to divide the dataset into equal k shares for k trainings and finally take the average of the K results. So we use 10-fold cross-validation to get the result, which can guarantee the stabilization of the result.

3.2 Model Performances

To verify the reliability of our proposed method, we constructed LSTM classifiers for comparison. We utilized 28 different datasets consisted of 2 of the 8 features, 56 new datasets composed of 3 of the 8 features, 56 different datasets composed of 3 of the 8 features, 70 datasets made up of 4 of the 8 features, 56 datasets consisted of 5 of the 8 features, 28 datasets made up of 6 of the 8 features, 8 datasets made up of 7 of the 8 features and a dataset made up of 8 of the 8 features. What's more, we also chose 10-fold cross-validation to evaluate the classifiers we constructed. The results are shown in the table following (Fig. 1):

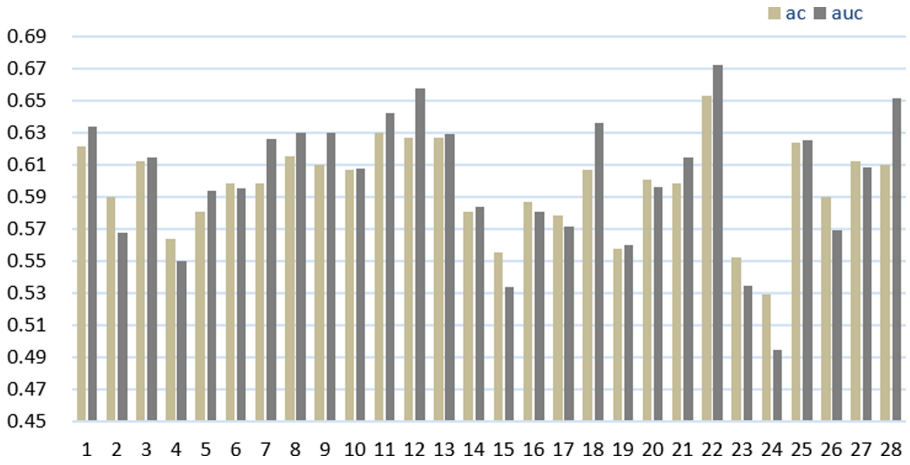

Fig. 1. The result gotten from 2 of the 8 features.

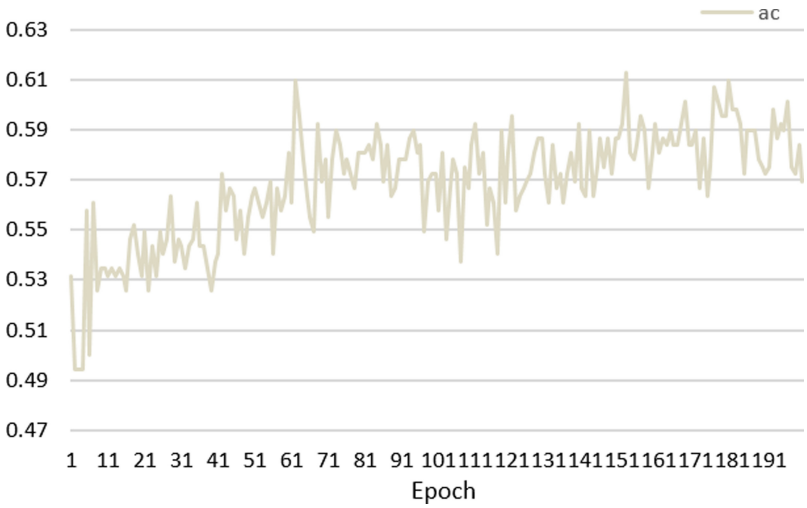

Fig. 2. The result gotten from the 8 features.

From the result gotten from 28 different datasets consisted of 2 of the 8 features, we noticed that the 22nd combination stands apart, which could mean that the mixture of 4th feature and the 7th feature contributes more to the classification, while the mixture of the 5th and 6th feature behaved relatively not well (Fig. 2).

From the result gotten from the datasets consisted of the 8 features, we can see that the result of not only AC but AUC rises in general, but fluctuated greatly. We thought the reason for it belongs to not only the weak classification features but the influence of interference term.

The comparation we chose is LSTM. The control flow of LSTM is similar to that of RNN. They process the data flowing by cells in the process of forward propagation. The difference is that the structure and operation of cells in LSTM change.

The core of LSTM is cell state and "gate" structure. Cell state is equivalent to the path of information transmission, so that information can be transmitted in the sequence. In theory, cell state can transmit the relevant information in the process of sequence processing all the time.

Therefore, even the information of earlier time step can be carried to the cells of later time step, which overcomes the influence of short-term memory. We can add and remove information through the "gate" structure, which will learn what information to save or forget in the training process.

We choose the best result from the experiment below, and together with the result from LSTM model, are shown in the table below (Table 2).

Table 2. The performances of each classification model.

Feature number	Acc	AUC
C_8^8 (lstm)	0.5364	0.5530
C_8^8 (capture)	0.61271	0.62117
C_8^6_14 (capture)	0.65318	0.69013

It can be seen that, even without the selections of features, the capture model performs better than the LSTM model. And when remove some features which obstacle classification, an improvement of the effect of the classification has been reflected clearly. Among the 8 features which are hydrophilicity value, mean polarity, isoelectric point, refractivity, average flexibility indices, average volume of buried residue, transfer free energy to surface, and consensus normalized hydrophobicity, the 6 most effective features for classification are hydrophilicity value, mean polarity, refractivity, average flexibility indices, transfer free energy to surface, and consensus normalized hydrophobicity.

4 Conclusions

In recent years, studies about post translational modifications have grown more and more popular. Because of the great effects of post translational modifications, distinguishing them from various modifications is quite necessary. However, the huge amounts of features as well as simples make it really hard to distinguish between the modifications which are post translation or not. What's more, the features for classification are quite weak. So we did a dozen of works to improve its precision. We picked up the capsule as our initial model. The model behaves way better than the LSTM model, which is already a better model compared to the basic models like RNN. Then, we reassemble the database by the law of combination. As a result, we get 28 new datasets consisted of 2 of the 8 features, 56 new datasets composed of 3 of the 8 features, 56 different

datasets composed of 3 of the 8 features, 70 datasets made up of 4 of the 8 features, 56 datasets consisted of 5 of the 8 features, 28 datasets made up of 6 of the 8 features, 8 datasets made up of 7 of the 8 features and a dataset made up of 8 of the 8 features. We put these datasets into the model for classification, and compared them among each other to found the feature combination which has better classification effect. The combination consists of features named hydrophilicity value, mean polarity, refractivity, average flexibility indices, transfer free energy to surface, and consensus normalized hydrophobicity behaved best. What's more, when it comes to the combinations consisted of different amount of features, we also picked some better ones, and pointed out the combinations which behaved bad. Apart from the features themselves, sometimes the right combination just helps a lot too.

Acknowledgement. This work is supported by the fundamental Research Funds for the Central Universities, 2020QN89, Xuzhou science and technology plan project (KC19142), the talent project of 'Qingtan scholar' of Zaozhuang University, Jiangsu Provincial Natural Science Foundation, China (SBK2019040953), Youth Innovation Teamof Scientific Research Foundation of the Higher Education Institutions of Shandong Province, China (2019KJM006), the Key Research Program of the Science Foundation of Shandong Province (ZR2020KE001), the PhD research startup foundation of Zaozhuang University (2014BS13) and Zaozhuang University Foundation (2015YY02), the Natural Science Foundation of China (61902337), Natural Science Fund for Colleges and Universities in Jiangsu Province (19KJB520016), Xuzhou Natural Science Foundation KC21047 and Young talents of science and technology in Jiangsu.

References

1. Molinie, B., Giallourakis, C.C.: Genome-Wide Location Analyses of N6-Methyladenosine Modifications (m6A-Seq), pp. 45–53. Humana Press (2017)
2. Nye, T.M., van Gijtenbeek, L.A., Stevens, A.G.: Methyltransferase DnmA is responsible for genome-wide N6-methyladenosine modifications at non-palindromic recognition sites in Bacillus subtilis. Nucleic Acids Res. **48**, 5332–5348 (2020)
3. O'Brown, Z.K., Greer, E.L.: N6-methyladenine: a conserved and dynamic DNA mark. In: Jeltsch, A., Jurkowska, R.Z. (eds.) DNA Methyltransferases - Role and Function. AEMB, vol. 945, pp. 213–246. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-43624-1_10
4. Zhang, G., Huang, H., Liu, D.: N6-methyladenine DNA modification in Drosophila. Cell **161**(4), 893–906 (2015)
5. Janulaitis, A., Klimašauskas, S., Petrušyte, M.: Cytosine modification in DNA by BCNI methylase yields N4-methylcytosine. FEBS Lett. **161**, 131–134 (1983)
6. Unger, G., Venner, H.: Remarks on minor bases in spermatic desoxyribonucleic acid. Hoppe-Seyler's Zeitschrift fur physiologische Chemie **344**, 280–283 (1966)
7. Fu, Y.: N6-methyldeoxyadenosine marks active transcription start sites in Chlamydomonas. Cell **161**, 879–892 (2015)
8. Greer, E.L., Blanco, M.A., Gu, L.: DNA methylation on N6-adenine in C. elegans. Cell **161**, 868–878 (2015)
9. Wu, T.P., Wang, T., Seetin, M.G.: DNA methylation on N6-adenine in mammalian embryonic stem cells. Nature **532**, 329–333 (2016)
10. Xiao, C.L., Zhu, S., He, M.: N-methyladenine DNA modification in the human genome. Mol. Cell **71**, 306–318 (2018)

11. Zhou, C., Wang, C., Liu, H.: Identification and analysis of adenine N6-methylation sites in the rice genome. Nat. Plants **4**, 554–563 (2018)
12. Chen, W., Lv, H., Nie, F.: i6mA-Pred: identifying DNA N6-methyladenine sites in the rice genome. Bioinformatics 2796–2800 (2019)
13. Almagor, H.: A Markov analysis of DNA sequences. J. Theor. Biol. **104**, 633–645 (1983)
14. Borodovsky, M., Mclninch, J.D., Koonin, E.V.: Detection of new genes in a bacterial genome using Markov models for three gene classes. Nucleic Acids Res. **17**, 3554–3562 (1995)
15. Durbin, R., Eddy, S.R., Krogh, A.: Biological Sequence Analysis Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press, Cambridge (1998)
16. Ohler, U., Harbeck, S., Niemann, H.: Interpolated Markov chains for Eukaryotic promoter recognition. Bioinformatics 362–369 (1999)
17. Reese, M.G., Eeckman, F.H., Kulp, D.: Improved splice site detection in Genie. J. Comput. Biol. 311–323 (1997)
18. Wren, J.D., Hildebrand, W.H., Chandrasekaran, S.: Markov model recognition and classification of DNA/protein sequences within large text databases. Bioinformatics 4046–4053 (2005)