

iCarPS: a computational tool for identifying protein carbonylation sites by novel encoded features

Dan Zhang^{1&}, Zhao-Chun Xu^{2&}, Wei Su¹, Yu-He Yang¹, Hao Lv¹, Hui Yang¹, Hao Lin^{1,*}

¹ School of Life Science and Technology and Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu 610054, China.

² Computer Department, Jingdezhen Ceramic Institute, Jingdezhen 333403, China.

& The authors wish it to be known that, in their opinion, this article represents joint First Authors.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Protein carbonylation is one of the most important oxidative stress-induced post-translational modifications (PTMs), which is generally characterized as stability, irreversibility and relative early formation. It plays significant role in orchestrating various biological processes and has been already demonstrated to be related to many diseases. However, the experimental technologies for carbonylation sites identification are not only costly and time-consuming, but also unable of processing a large number of proteins at a time. Thus, rapidly and effectively identifying carbonylation sites by computational methods will provide key clues for the analysis of occurrence and development of diseases.

Results: In this study, we developed a predictor called iCarPS to identify carbonylation sites based on sequence information. A novel feature encoding scheme called residues conical coordinates combined with their physicochemical properties was proposed to formulate carbonylated protein and non-carbonylated protein samples. To remove potential redundant features and improve the prediction performance, a feature selection technique was utilized. The accuracy and robustness of iCarPS were proved by experiments on training and independent datasets. Comparison with other published methods demonstrated that the proposed method is powerful and could provide powerful performance for carbonylation sites identification.

Availability: Based on the proposed model, a user-friendly webserver and a software package were constructed, which can be freely accessed at <http://lin-group.cn/server/iCarPS>.

Contact: hlin@uestc.edu.cn (H. L.)

Các biến đổi sau dịch mã do ứng suất oxy hóa gây ra (PTM), đặc trưng bởi:
- Tính bền vững
- Không thể đảo ngược
- Cấu thành sớm tương đối
Nó liên quan đến nhiều loại bệnh khác nhau.
Các công nghệ hiện nay: tốn kém, mất thời gian, xử lý nhiều proteon 1 lúc
=> Việc xác định nhanh, hiệu quả các vị trí cacbonyl hóa bằng pp tính toán cung cấp dấu hiệu của nguồn bệnh
=> Phương pháp dự báo iCarPS: để xác định carbonylation sites dự trên sequence information

robustness: độ tin cậy, độ chắc chắn.
accuracy: độ chính xác

Một schema encoding feature mới được gọi là tọa độ hình nón chất cận + đặc tính của chúng -> tạo ra các mẫu carbonylated protein + non-carbonylated protein.

1 Introduction

Reactive oxygen species (ROS) can be continuously produced within cells by a variety of different endogenous and exogenous sources (Halliwell, 2007). When the production of ROS is seriously unbalanced over antioxidant defenses, oxidative stress will arise (Halliwell, 2007; Reddy, et al., 2009). It has been found that oxidative stress can induce a variety of post-translational modifications (PTMs) on proteins, such as nitration, carbonylation, sulfhydrylation, hydroxylation, and S-glutathionylation (Gianazza, et al., 2007). Among these oxidative stress-

induced PTMs in all macromolecule, protein carbonylation, an irreversible and non-enzymatic modification, is the most common and has attracted much attention (Dalle-Donne, et al., 2006).

Protein carbonylation (Figure 1) (Rauniyar, et al., 2010), generally characterized as stability, irreversibility and relative early formation, has long been well studied and considered as a biomarker to measure oxidative stress levels (Moller, et al., 2011). Some researchers have confirmed that protein carbonylation has impact on protein function, protein folding, proteolysis, and cellular dysfunction, often leading to proteasomal degradation of proteins (Dalle-Donne, et al., 2006).

Moreover, numerous studies have been shown that high levels of protein carbonylation have been observed in a great lot kind of major human diseases, such as chronic lung disease, Parkinson's disease, cataract genesis, chronic renal failure, sepsis and many other age-related diseases (Dalle-Donne, et al., 2003; Moller, et al., 2011). Consequently, it is of great significance for biomedical research and drug development to carry on the research of protein carbonylation. Especially, identification of protein carbonylation sites could provide important clues for understanding the process and consequences of cellular metabolism and the affected protein.

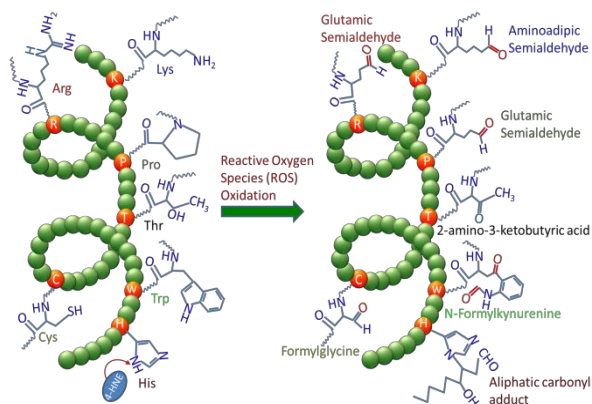


Fig. 1. Chemical structures of the oxidative carbonylation products, including Amino adipic Semialdehyde, Glutamic Semialdehyde, and 2-amino-3-ketobutyric acid, Formylglycine, N-Formylkynurenine, and Aliphatic carbonyl adduct, derived by carbonylation of K, R or P, T, C, W, and H, respectively, in which 4-HNE is subject to Michael addition reactions with the amino acid side chain of H.

Recent years, numerous experimental methods have been used to quantify protein carbonylation and identify their carbonylation sites (Bollineni, et al., 2011; Kuzmic, et al., 2016). Some interested outcomes have shown that certain amino acid residues, especially R, K, P and T, have the susceptibility to be carbonylated and have effect on adjacent residues (Bollineni, et al., 2014). Moreover, more carbonylation sites are found in RKPT-rich regions and carbonylation sites have a strong tendency to cluster (Rao and Moller, 2011).

Accurate identify the modification site could provide important clues for understanding ROS and multi diseases. However, it is time-consuming and expensive if only using purely biochemical experimental technique to identify the exact protein carbonylation sites, particularly for large-scale datasets. The accumulation of protein data and development of artificial intelligent techniques provide us an opportunity to generate a powerful model for carbonylation sites. To date, several computational classifiers have been successively constructed for identifying protein carbonylation sites. Lv et al. (Lv, et al., 2014) collected 250 carbonylated protein sequences verified by biochemical experiments, and constructed a benchmark dataset containing R, K, P and T modified residues in human and other mammals (mouse, rabbit and bovine). An online prediction tool named CarSPred was constructed by using four kinds of features and combining the minimum Redundancy Maximum Relevance (mRMR) feature selection technique. Based on Lv's dataset, Jia et al. (Jia, et al., 2016) developed a predictor named iCar-PseCp by using random forest (RF) algorithm. Later, Hasan et al. (Hasan, et al., 2017) constructed a SVM-based predictor called predCar-site. Moreover, Xu et al. (Xu, et al., 2014) developed an SVM-

based standalone software called PTMPred to predict all types of PTM sites including protein carbonylation sites. Besides, Lv et al. (Lv, et al., 2016) also developed a computational predictor named CarPred.Y based on SVM to predict carbonylation sites in yeast with several kinds of features. Combining profile hidden Markov model with SVM, Kao et al. (Kao, et al., 2017) developed an integrative model named MDD-carb by using multifarious features to identify protein carbonylation sites in mammalian proteins with substrate motifs. In the same year, Weng et al. (Weng, et al., 2017) built predictive models to identify carbonylation sites of proteins in human.

Although great results have been achieved in the previous works, unfortunately, some limitations and shortcoming still exist, shown as below. (i) Few webserver were established. Among above-mentioned predictors, there are only three webserver namely iCar-PseCp, predCar-site, and MDD-carb. However, the last two online predictors do not work now. The others such as CarSPred, PTMPred and CarPred.Y just provided stand-alone prediction tools which is difficult for most of biochemical scholars to use them. (ii) The performance of these predictors can be further improved from various evaluation indexes. Although some works reported high accuracies (AUC~1), it may be caused by an over-fitting of the training dataset because the proposed method was constructed based on high dimensional feature set. Moreover, no independent data was used to perform examination. (iii) Few independent testing datasets were constructed to validate the performance of predictors. (iv) Most of these methods did not select optimal features by using feature selection techniques, which is one of the most important steps of building an effective and robust prediction model.

In view of the above description, in this study, we devoted to improve the prediction capability in identifying protein carbonylation sites from the aforementioned four disadvantages. At first, high-quality training and testing datasets verified by experiment were obtained. Subsequently, a novel encoding scheme called residues conical coordinates based on mathematical construction were incorporated into nine physicochemical properties of amino acids to characterize carbonylated and non-carbonylated samples. Meanwhile, F-score was used to optimize features. And then, RF algorithm was employed to perform classification. We used cross-validation test and independent data test to evaluate our method. Based on the proposed model, a webserver named iCarPS was established.

2 Methods

2.1 Benchmark dataset

CarSPred's (Lv, et al., 2014) benchmark dataset was used in this study. It was derived from the 230 carbonylated protein sequences from human and 20 carbonylated protein sequences from other mammals. The modification sites contain four types of carbonylation sites, K, P, R and T. All of the carbonylated protein sequences were obtained from experimental data in the literatures by Lv et al. (Lv, et al., 2014). Considering the number of carbonylation sites on H, C, and W amino acids residues being extremely small and no reliable public data sources, we merely constructed the prediction model on K, P, R, T residues in this study.

The benchmark dataset consisted of four subsets termed \mathcal{S}_{\odot} (the symbol \odot denoting the single residue K, P, R or T), which can generally be formulated by

$$\mathcal{S}_{\odot} = \mathcal{S}_{\odot}^{+} \cup \mathcal{S}_{\odot}^{-} \quad (1)$$

where \mathbb{S}_{\odot}^{+} denotes the positive subset containing the samples of true carbonylation site for residue \odot , \mathbb{S}_{\odot}^{-} , the negative subset containing the samples of false carbonylation site for residue \odot , and the symbol \cup represents union in the set theory.

In order to construct the dataset \mathbb{S}_{\odot} formulated by Eq. (1), firstly, the $(2\xi + 1)$ -mer sliding window was used to extract the positive and negative samples with $\mathbb{U} = \odot$ at the center along each of protein sequence segments. Thus, a potential carbonylation site-contained protein sequence sample can be expressed as

$$P_{\xi}(\mathbb{U}) = P_{-\xi}P_{-(\xi-1)}\cdots P_{-2}P_{-1}\mathbb{U}P_{+1}P_{+2}\cdots P_{+(\xi-1)}P_{+\xi} \quad (2)$$

where the double-line character \mathbb{U} represents the residue \odot (one of residues K, P, R or T), the value of subscript ξ is an integer, $P_{-\xi}$ denotes the ξ -th upstream amino acid from the center, $P_{+\xi}$ represents the ξ -th downstream amino acid from the center, and so on. According to the location information of carbonylation sites, the protein sequence segments were considered as candidate positive samples and were put into the subset \mathbb{S}_{\odot}^{+} , if their centers were the experimentally confirmed carbonylation sites. Otherwise, the protein sequence segments were regarded as negative samples and were put into the negative subset \mathbb{S}_{\odot}^{-} . The final constructed benchmark datasets are summarized in Table 1. The sequence identity of benchmark dataset has been reduced to 30% by using CD-HIT program (Huang, et al., 2010).

Table 1. The sample number in training and testing datasets.

Group	Dataset	Carbonylation sites			
		K	P	R	T
Training	positive	266	114	119	116
	negative	1802	716	754	702
Testing	positive	34	12	17	5
	negative	147	76	93	30

2.2 Feature vector construction

2.2.1 Physicochemical properties (PCPs)

Each amino acid residue has many specific physicochemical and biologic properties, which could effect on protein properties and play an important role in determining the structures and functions of the protein (Zhao, et al., 2016). In this study, we adopted nine physicochemical properties used in previous reference (Tang, et al., 2016), including hydrophobicity, hydrophilicity, mass, pK1, pK2, pI, rigidity, flexibility, irreplaceability, of which the first six have been widely used in protein bioinformatics. Here we will briefly introduce the last three physicochemical properties (rigidity, flexibility, irreplaceability). The rigidity and flexibility of amino acid side chains have been estimated for polypeptides and local protein domains associated with protein property alterations (Gottfries and Eriksson, 2010). Furthermore, the two properties have also been used to predict conformation and protein fold changes (de Mol, et al., 2005). Besides, in the evolution, some residues are easy to be replaced, while others are difficult. And averaged mutational deteriorations (AMD) of amino acids can be used to describe their irreplaceability. Thus, the irreplaceability reflects mutational deterioration in the course of the evolution of life (Luo, 1988). Therefore, the three properties may play an important complementary role to other properties for describing protein or peptide features. Notably, the values of physicochemical properties of pseudo amino acid X was defined to 0 in our work. All the original values of these physicochemical properties can be found in <http://lin-group.cn/server/iCarPS/download.html>. The dimensionless processing of

all the original values should be made before the values of the nine physicochemical properties of amino acids are used, shown as below:

$$P_v(R_i) = \frac{P_v(R_i) - \langle P_v \rangle}{SD(P_v)} \quad (3)$$

in which, $P_v(R_i)$ is the numerical value of the v -th ($v=1, 2, \dots, L$, L denotes the length of the protein sequence) local amino acid physicochemical property for residue R_i at position i ; the symbol $\langle \rangle$ means the average value of amino acids; and SD denotes the corresponding standard deviation.

Accordingly, each carbonylation (or non-carbonylation) protein sequence sample $P_{\xi}(\mathbb{U})$ formulated by Eq. (2) can be denoted as a $n \times L$ dimensional vector, shown as below:

$$P_{\xi}(\mathbb{U}) = [x_1, x_2, \dots, x_n, \dots, x_n \times L]^T \quad (4)$$

in which, the value n is the number of the physicochemical properties; L is the length of the protein sequence, the symbol " T " is the transpose operator, and the element x denotes the values of physicochemical properties on the corresponding position of the amino acid residue along the protein sequence. Therefore, the value n in Eq. (3) equals to 9 and the length of each protein sequence L equals to 27, as used in CarSPred (Lv, et al., 2014). Then, each amino acid was constructed into 9 features. For a peptide fragment, a 243-dimension ($27 \times 9 = 243$) feature vector was obtained from this encoding scheme.

2.2.2 A novel encoding scheme based on mathematical construction

It is well known that some amino acids display similar characteristics. Thus, these amino acids can be clustered into several groups. In this work, 20 natural amino acids were divided into four groups, as listed in Table 2.

Table 2. The groups of 20 natural amino acids.

Groups	Description	Amino acids
Class I	non-polar residues	A, V, L, I, P, F, W, M
Class II	polar residues	G, S, T, C, Y, N, Q
Class III	basic residues	K, R, H
Class IV	acidic residues	D, E

We supposed that each amino acid could be mapped to a point $P(x, y, z)$ in 3-dimension space, in which we employed conical coordinates to display protein sequences. The transformation between Cartesian coordinates and conical coordinates is shown in Eq. (5):

$$\begin{cases} x = r \times \sin\varphi \times \cos\theta \\ y = r \times \sin\varphi \times \sin\theta \\ z = r \times \cos\theta \end{cases} \quad \varphi \in [0, \pi], \theta \in [0, 2\pi] \quad (5)$$

In order to capture key protein features in a simple and effective way, two basic hypotheses are put forward: (1) amino acids in the same group are distributed on a conical surface because they display similar characteristics. For example, class I stands for non-polar residues, i.e. A, V, L, I, P, F, W, and M ($P_{1j}, j=1, 2, \dots, 8$) are fixed on the conical surface $\varphi = \varphi_1$. Class II denotes polar residues, i.e. G, S, T, C, Y, N, and Q ($P_{2j}, j=1, 2, \dots, 7$) are located on the conical surface $\varphi = \varphi_2$, and so forth, shown in Figure 2. (2) To reflect the difference between amino acids, the length r of radius vector was set to molecular weight of the corresponding amino acid, which can be found in <http://lin-group.cn/server/iCarPS/download.html>.

Thus, combining the conical coordinates with the above-mentioned nine physicochemical properties and types of amino acids, each amino acid can be numerically converted into a three-dimensional vector by following equation:

$$\begin{cases} x_{ij} = r_{ij} \times \sin\varphi_i \times \cos\theta_{ij} \\ y_{ij} = r_{ij} \times \sin\varphi_i \times \sin\theta_{ij} \\ z_{ij} = r_{ij} \times \cos\theta_{ij} \end{cases} \quad \varphi_i \in [0, \pi], \theta_{ij} \in [0, 2\pi] \quad (6)$$

in which, r_{ij} represents the molecular weight of the j -th ($j = 1, 2, \dots, L_i$) amino acid in the i -th ($i = 1, 2, 3, 4$) group. L_i denotes the number of amino acids in the i -th group, and φ_i , θ_{ij} can be defined as below, respectively.

$$\varphi_i = \pi \times \left| \sin \frac{\bar{a}_i}{\left(\frac{1}{4} \sum_{j=1}^{L_i} \bar{a}_i \right) * \sqrt{\frac{1}{4} \sum_{j=1}^{L_i} (\bar{a}_i - \frac{1}{4} \sum_{j=1}^{L_i} \bar{a}_i)^2}} \right| \quad (7)$$

$$\theta_{ij} = \pi + 2 \times \arctan \frac{\sum_{m=1}^9 PC_{jm} - \bar{a}_i}{\sqrt{\frac{1}{L_i} \sum_{j=1}^{L_i} \left| \sum_{m=1}^9 PC_{jm} - \bar{a}_i \right|^2}} \quad (8)$$

in which, $\bar{a}_i = \frac{1}{L_i} \sum_{j=1}^{L_i} \sum_{m=1}^9 PC_{jm}$, and PC_{jm} represents the standard value of the m -th ($m = 1, 2, \dots, 9$) aforementioned physicochemical property of the j -th amino acid in the i -th group.

Therefore, in accordance with the above procedure, each amino acid is uniquely represented with a three-dimensional vector (x, y, z) . Thus, the sample sequence $P_\xi(\mathbb{U})$ formulated by Eq. (2) can be transformed into a $3 \times L = 3 \times (2\xi + 1)$ dimensional vector, shown as:

$$P_\xi(\mathbb{U}) = [x_1, y_1, z_1, x_2, y_2, z_2, \dots, x_L, y_L, z_L]^T \quad (9)$$

Then, the geometrical center $(\bar{x}, \bar{y}, \bar{z})$ of the sample $P_\xi(\mathbb{U})$ could be obtained as:

$$\bar{x} = \frac{1}{L} \sum_{n=1}^L x_n, \bar{y} = \frac{1}{L} \sum_{n=1}^L y_n, \bar{z} = \frac{1}{L} \sum_{n=1}^L z_n \quad (10)$$

In addition, the accumulated geometric center $(\bar{X}, \bar{Y}, \bar{Z})$ of the sample $P_\xi(\mathbb{U})$ could be obtained as sequence features, expressed as:

$$\bar{X} = \frac{1}{L} \sum_{h=1}^L X_h, \bar{Y} = \frac{1}{L} \sum_{h=1}^L Y_h, \bar{Z} = \frac{1}{L} \sum_{h=1}^L Z_h \quad (11)$$

in which, $X_h = \sum_{n=1}^h x_n$, $Y_h = \sum_{n=1}^h y_n$, $Z_h = \sum_{n=1}^h z_n$.

Moreover, the geometrical centers $(\bar{x}_i, \bar{y}_i, \bar{z}_i)$ of the i -th group ($i = 1, 2, 3, 4$) of amino acids in the sample $P_\xi(\mathbb{U})$ were calculated by using the following equation:

$$\bar{x}_i = \frac{1}{L_i} \sum_{n=1}^{L_i} x_n, \bar{y}_i = \frac{1}{L_i} \sum_{n=1}^{L_i} y_n, \bar{z}_i = \frac{1}{L_i} \sum_{n=1}^{L_i} z_n \quad (12)$$

Consequently, the sample $P_\xi(\mathbb{U})$ could be expressed as:

$$P_\xi(\mathbb{U}) = [\bar{x}_1, \bar{y}_1, \bar{z}_1, \bar{x}_2, \bar{y}_2, \bar{z}_2, \bar{x}_3, \bar{y}_3, \bar{z}_3, \bar{x}_4, \bar{y}_4, \bar{z}_4, \bar{x}, \bar{y}, \bar{z}, \bar{X}, \bar{Y}, \bar{Z}]^T \quad (13)$$

Finally, in accordance with Eq. (13), a 18-dimension vector, which was generated by integrating above-mentioned methods, was used as a feature vector to quantitatively describe intrinsic properties of the protein sample $P_\xi(\mathbb{U})$.

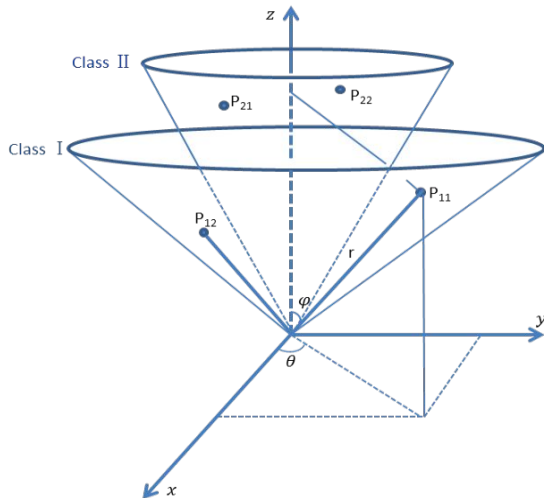


Fig. 2. Schematic illustration to show the 3-dimension conical representation for characterizing amino acid residues. Class I, II stand for non-polar residues group and polar residues group, respectively. P_{ij} represents each amino acid of corresponding group, where i denotes the i -th group, and j denotes j -th amino acid of corresponding group. φ represents the conical surface which is formed by projection of amino

acids of the corresponding group. In addition to, r represents the molecular weight of amino acids.

2.3 Random forest (RF)

Random forest (RF) is an ensemble learning method consisting of many individual decision trees, mainly used in regression and classification (L, 2001). It uses bootstrap resampling technique to generate a new training dataset which is randomly sampled from the original training dataset and used to evaluate at each node of the decision trees. Then, the final decision is made by decision fusion of all the trees by majority voting. Owing to its ability to supply an empirical approach to trail variable interactions, RF is considered as an appropriate classifier to handle large-scale dataset, especially for imbalanced dataset (Livingston, 2005; Zeng, et al., 2019). Therefore, RF has some unique advantages such as good anti-noise ability and easy parallelization so that it is widely used for constructing computational prediction models to solve bioinformatics problems (Manavalan, et al., 2019; Manavalan, et al., 2019). The detailed procedures of RF algorithm and its formulation have been very clearly elaborated in pervious study, and hence there is no need to repeat here.

In this study, the predictive model for identifying protein carbonylation sites based on RF algorithm was constructed by java application programming interface calling a library of RF program, which was integrated in WEKA data mining package (Frank, et al., 2004; Smith and Frank, 2016). The default parameters of random forest were utilized to construct our model in the Weka 3.8.

2.4 F-score and IFS feature selection

A suitable feature selection could not only overcome the curse of dimensionality and reduce overall training times, but also reduce the over-fitting risk and improve accuracy and generalization power of the proposed models. In view of these characteristics, feature selection strategies, including F -score, mRMR, analysis of variance (ANOVA) and binomial distribution (BD) and so on, have been successfully applied in the field of bioinformatics (Basith, et al., 2020; Basith, et al., 2019; Liu and Chen, 2020). In the current study, we committed to developing simpler and faster model for identifying protein carbonylation sites by applying F -score for optimizing the features. Finally, the above 261-dimensional features were ranked according to F -score values. And the increment feature selection (IFS) (Liu, et al., 2020; Tan, et al., 2019; Yang, et al., 2019) was used to determine the dimension of optimal feature sets by using cross-validation with RF.

2.5 Evaluation metrics

Ten-fold cross-validation test was employed to examine the performance of the proposed models. Moreover, the traditional metrics (Bao, et al., 2018; Song, et al., 2010; Tang, et al., 2018; Wang, et al., 2014) such as sensitivity (Sn), specificity (Sp), overall accuracy (ACC) and Matthews correlation coefficient (MCC) were adopted to evaluate the predictive performance of models, and were described as:

$$Sn = \frac{TP}{TP + FN} \quad (14)$$

$$Sp = \frac{TN}{TN + FP} \quad (15)$$

$$ACC = \frac{TP + FP + TN + FN}{TP + FP + TN + FN} \quad (16)$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TN + FN) \times (TP + FN) \times (TN + FP)}} \quad (17)$$

where TP , TN , FP , FN indicate the true positives (i.e. correctly predicted as carbonylated protein), true negatives (i.e. correctly predicted as non-

carbonylated protein), false positives (i.e. incorrectly predicted as carbonylated protein) and false negatives (i.e. incorrectly predicted as non-carbonylated protein), respectively. The higher the value of ACC , Sn and Sp are the more effective the predictor is. In addition, $-1 < MCC < 1$, a value of $MCC = 1$ indicates the best possible prediction while $MCC = -1$ indicates the worst possible prediction (or anti-correlation). $MCC = 0$ would be expected for a random prediction scheme.

Furthermore, the Receiver Operating Characteristic Curve (ROC) (Hanley and McNeil, 1982), which is drawn by the false positive rate ($FPR = 1 - Sp$) as X -axis and the true positive rate ($TPR = Sn$) as Y -axis, was also utilized to measure the performance of the predictor across the entire range of random forest decision values. The area under the ROC (AUC) can be calculated and usually used to quantitatively evaluate the quality of the predictors. A perfect predictor is proved to have the value of $AUC = 1$ and the random performance is proved to have the value of $AUC = 0.5$.

3 Results and discussion

3.1 Position-specific differences analysis

The position-specific sequence characteristics can be recognized as important conserved features. To reveal the position-specific differences between positive and negative samples, a web-based tool called Two Sample Logos (TSL) (Vacic, et al., 2006) was used in the current work to determine statistically difference of residues around true (or false) carbonylation sites. The distribution differences (t test, p value < 0.05) of four kinds of residues (Lysine, Proline, Arginine, Threonine) were graphically represented, in Figure 3, in which the residues enriched or depleted surrounding carbonylation sites are above or under the horizontal axis, respectively.

As illustrated in Fig. 3, residues K, P, R and T are enriched in the flanking sequences of the true carbonylation sites in positive samples. This observation that the carbonylation sites have a strong tendency to be in RKPT-enriched regions has also been observed by Rao and Moller (Rao and Moller, 2011). Second, in each sub-graph, the enrichment degree of residues in the upstream of carbonylation are significantly higher than those in the downstream of modification sites. For example, in Fig. 3(a), the occurrence frequency of residue K in the upstream of K carbonylation site is obviously higher than that in the downstream of the modification site. This indicates that the carbonylation sites are prone to cluster in carbonylation site. Third, most of graphical residues with different size and type at the same position show that there exists a big position-specific difference between the positive and negative sets. Only a few positions display insignificant differences. This suggests that the position-specific distribution of residues around residues K, P, R, T could influence their carbonylation.

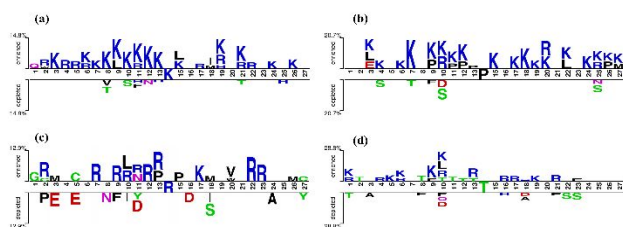


Fig. 3. Illustration to show significant position-specific differences of residues surrounding carbonylation and non-carbonylation sites. Sub-graphs from (a)-(d) are for K, P, R, T carbonylation sites, respectively.

In view of statistically significant position-specific differences between positive and negative samples of each type carbonylation site mentioned above, it is possible to develop a computational method to identify potential carbonylation sites only by using sequence information. In fact, conical coordinates of amino acids proposed in this paper could reflect the positional and composition information. Thus, the residues distribution can be used as features for identifying carbonylation sites.

3.2 Investigation of physicochemical properties around carbonylation sites

To further reveal the physicochemical properties surrounding carbonylation sites, we statistically analyzed the distribution of nine physicochemical properties of residues around the carbonylation and non-carbonylation site based on training datasets as shown in Figure S1. From the figure, we noticed that the carbonylation-containing segments are dramatically different from non-carbonylation-containing segments. At first, for four modified residues, positive samples displayed larger fluctuation than negative samples in term of the nine physicochemical properties. Secondly, we observed that the mean values of hydrophilicity, pI and flexibility of carbonylation-containing segments are larger than those of negative samples in most positions, especially in the positions near the modified sites ($-7 \sim +7$), which may be related to the change of protein structure caused by carbonylation modification. However, the distribution of hydrophobicity, pK_2 and irreplaceability are exhibited opposite phenome. It also showed that the distributions of nine physicochemical properties flanking the carbonylation sites are not symmetrical, suggesting that the residues in the upstream and downstream modified sites have different effect on residue carbonylation. Furthermore, it is observed that the difference of hydrophobicity, hydrophilicity, pI and flexibility at each position between positive and negative samples are larger when comparing with other physicochemical properties (pK_1 , rigidity and irreplaceability). We found an interesting phenomenon that, compared with modified T, R and P, the surrounding residues in residue K display less fluctuation. From Figure 3, we also found that the consensus sequences around modified residue K is more conserve than that around other modified residues. This might explain above phenomenon for modified residue K. These results further illustrated the above nine physicochemical properties encoding feature are of great significance for carbonylation and can be used as features to perform prediction.

3.3 Performance evaluation of various features

Above statistical results remind us that the carbonylation sites can be identified by computational method based on these characteristics. To assess the prediction performance of different features, the predictive models were trained and tested using 10-fold cross-validation based on RF. The reason why we used AUC values as standard is that the it could provide a more objective evaluation on imbalance benchmark dataset comparing with sensitivity, specificity and overall accuracy.

We used two kinds of features, namely nine physicochemical properties of amino acids (9_PCPs) and 3-dimension conical coordinates (CC) to formulate samples. Based on the 10-fold cross-validation test, the predictive performance of each sequence-based feature was drawn in Figure 4. It shows that the 9_PCPs encoded method could produce the AUC values of 0.741, 0.727, 0.580 and 0.626 respectively for K, P, R, and T carbonylation sites. Additionally, the CC encoded feature obtained

AUC values of 0.725, 0.786, 0.661, 0.735 for K, P, R, and T carbonylation site predictions respectively. The two features could describe carbonylation samples from sequence composition and physicochemical properties perspectives. Thus, we guessed that the combination between two features could improve the predictive performance. However, after performing the examination, we found that these combined features cannot improve the performance for all carbonylation sites. The combined features are effective for K, R and T carbonylation sites. They could increase the *AUC* values to 0.775, 0.662 and 0.745, respectively for K, R and T carbonylation sites. For R carbonylation site predictions, the *AUC* value of combined features is just slightly higher than that of CC encoded feature. What's more, for P carbonylation site prediction, the predictive performance of combined features even decreased to 0.765 compared with the performance of CC feature. Generally, noise or redundant information could reduce the model's performance, robust and efficiency. Thus, the phenomena about accuracy decrease was maybe derived from information redundancy. Therefore, it is necessary to pick out the best features for improving prediction accuracy. The following section will focus on feature optimization.

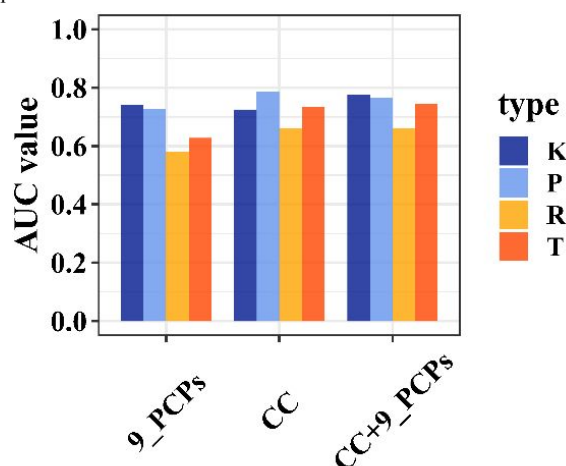


Fig. 4. Prediction performance of the RF models trained with various features based on 10-fold cross-validation.

3.4 Determination of optimal features

It is well known that the optimal feature subset can be found by investigating all possible combinations of features. However, it is impossible to do so because number of feature combinations is so large that we cannot examine them. For saving computing time and improving prediction accuracy, we used a feature optimization strategy to extract the most discriminative features. Firstly, the *F*-score was utilized to calculate the score of each feature of PCPs and CC feature set, then we sorted them in descending order by their scores. Each feature subset is added successively by the maximum value of *F*-score, and total of 261 feature subsets were obtained. We investigated the performance of these feature subsets by using RF algorithm with 10-fold cross-validation, and subsequently plotted four IFS curves for four modified residues in Figure 5 for determining the optimal feature subset. Figure 5 showed that the peak of *AUC* values for the K, P, R and T carbonylation sites were reached when the numbers of feature subsets equal to 193, 38, 77, and 60, respectively. The maximum *AUC* values are 0.789, 0.814, 0.726 and 0.790, respectively for K, P, R, and T carbonylation site predictions. For further evaluating the reliable and robust of the optimal models

constructed by the optimal features, independent testing datasets were used. The *AUC* values on independent data reach to 0.756, 0.752, 0.649 and 0.840, respectively for predicting K, P, R, and T carbonylation site. These results were recorded in Table 2. Comparison of results between training and test data showed that our proposed model is robust and can be used to identify carbonylation sites in proteins. Based on the proposed model, we constructed an online webserver called iCarPS which can be freely access from <http://lin-group.cn/server/iCarPS>. The tool will provide convenience to most of scholars.

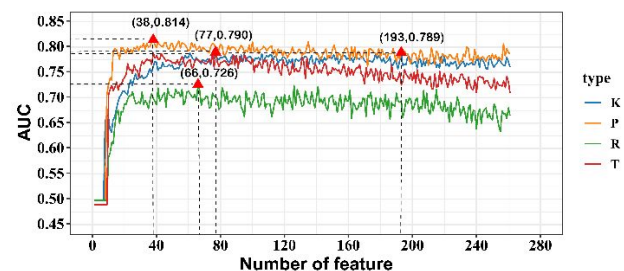


Fig. 5. A plot showing the IFS procedure optimized by *F*-score for identifying carbonylation sites. The top features where the peaks of *AUC* values of K, P, R and T marked by red triangle respectively in 10-fold cross-validation were used to perform prediction.

3.5 Compared with published method

To further demonstrate the performance of our method, we must compare our method with other published method. Here, the CarSPred was selected to perform comparison because it used same benchmark dataset. The results of CarSPred's method on the same training and independent testing dataset were directly obtained from their reports. The compared results were listed in the Table 3. As it can be seen from Table 3, the average *ACC* and *AUC* values of our method (iCarPS) were 1.20% and 7.95% higher than those of CarSPred's method on training dataset in 10-fold cross-validation. On independent data, although the *AUC* values of iCarPS was slightly lower than that of CarSPred for P-type of carbonylation site, the *AUC* values of iCarPS are significantly improved by 12% for other types (K, R, T) of carbonylation sites. Moreover, the average difference of *AUC* between training and independent data are 0.072 and 0.056, respectively for CarSPred and our model, suggesting that our proposed model is more stable. These comparisons indicated that our method is superior to CarSPred.

Table 3. Comparison between our model with the existing method

Predictor	AUC for Carbonylation Sites			
	K	P	R	T
CarSPred ¹	0.689	0.706	0.702	0.704
iCarPS ¹	0.789	0.814	0.726	0.790
CarSPred ²	0.670	0.783	0.535	0.680
iCarPS ²	0.756	0.752	0.649	0.840

¹ denotes the results from 10-fold cross-validation

² denotes the results from independent data

4 Conclusions

Carbonylation is an import post-translational modification which is generally considered as a biomarker of oxidative stress. In this study, we proposed a novel predictor for the identification of carbonylation sites by using sequence-derived features. A novel sequence composition descriptor called residues' conical coordinates was proposed to formulate residue sequence. Several experiments were made to prove the

robustness and effectiveness of our method. In addition, the algorithm in this study have taken full account of the sequence component information and physicochemical properties information. A feature selection method called *F-score* was utilized to successfully select the informative features and remove noisy. This process dramatically improved the prediction power of the model. Based on the optimal model, an online predictor iCarPS was established for identifying carbonylated proteins. It provided convenience to the experimental scientists who could obtain the desired results rapidly and accurately without repeating the mathematical details. In addition to, a software package for local computers was available at our website. Thus, we are sure that this predictor will become a useful tool for carbonylation analysis and further experimental researches. We also suggest that the method proposed in this study, especially for residues' conical coordinates, can be generalized to the prediction of other types of post-translational modification in the proteomics studies.

Funding

This work has been supported by the the National Nature Scientific Foundation of China (61772119), Sichuan Provincial Science Fund for Distinguished Young Scholars (2020JDQ0012), and the Science Strength Promotion Programme of UESTC.

Conflict of Interest: none declared.

References

- Bao, Y., et al. (2018) Toward more accurate prediction of caspase cleavage sites: a comprehensive review of current methods, tools and features, *Briefings in bioinformatics*.
- Basith, S., et al. (2020) Machine intelligence in peptide therapeutics: A next-generation tool for rapid disease screening, *Medicinal research reviews*.
- Basith, S., et al. (2019) SDM6A: A Web-Based Integrative Machine-Learning Framework for Predicting 6mA Sites in the Rice Genome, *Mol Ther Nucleic Acids*, **18**, 131-141.
- Bollineni, R., Hoffmann, R. and Fedorova, M. (2011) Identification of protein carbonylation sites by two-dimensional liquid chromatography in combination with MALDI- and ESI-MS, *Journal of proteomics*, **74**, 2338-2350.
- Bollineni, R.C., Hoffmann, R. and Fedorova, M. (2014) Proteome-wide profiling of carbonylated proteins and carbonylation sites in HeLa cells under mild oxidative stress conditions, *Free radical biology & medicine*, **68**, 186-195.
- Dalle-Donne, I., et al. (2006) Protein carbonylation, cellular dysfunction, and disease progression, *Journal of cellular and molecular medicine*, **10**, 389-406.
- Dalle-Donne, I., et al. (2003) Protein carbonylation in human diseases, *Trends in molecular medicine*, **9**, 169-176.
- de Mol, N.J., et al. (2005) Protein flexibility and ligand rigidity: a thermodynamic and kinetic study of ITAM-based ligand binding to Syk tandem SH2, *Chembiochem : a European journal of chemical biology*, **6**, 2261-2270.
- Frank, E., et al. (2004) Data mining in bioinformatics using Weka, *Bioinformatics*, **20**, 2479-2481.
- Gianazza, E., Crawford, J. and Miller, I. (2007) Detecting oxidative post-translational modifications in proteins, *Amino acids*, **33**, 51-56.
- Gottfries, J. and Eriksson, L. (2010) Extensions to amino acid description, *Molecular diversity*, **14**, 709-718.
- Halliwel, B. (2007) Biochemistry of oxidative stress, *Biochemical Society transactions*, **35**, 1147-1150.
- Hanley, J.A. and McNeil, B.J. (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve, *Radiology*, **143**, 29-36.
- Hasan, M.A., et al. (2017) predCar-site: Carbonylation sites prediction in proteins using support vector machine with resolving data imbalanced issue, *Analytical biochemistry*, **525**, 107-113.
- Huang, Y., et al. (2010) CD-HIT Suite: a web server for clustering and comparing biological sequences, *Bioinformatics*, **26**, 680-682.
- Jia, J., et al. (2016) iCar-PseCp: identify carbonylation sites in proteins by Monte Carlo sampling and incorporating sequence coupled effects into general PseAAC, *Oncotarget*, **7**, 34558-34570.
- Kao, H.J., et al. (2017) MDD-carb: a combinatorial model for the identification of protein carbonylation sites with substrate motifs, *BMC systems biology*, **11**, 137.
- Kuzmic, M., et al. (2016) In situ visualization of carbonylation and its co-localization with proteins, lipids, DNA and RNA in *Caenorhabditis elegans*, *Free radical biology & medicine*, **101**, 465-474.
- L., B. (2001) Random Forests., *Machine Learning* **45**(1):5-32.
- Liu, K. and Chen, W. (2020) iMRM: a platform for simultaneously identifying multiple kinds of RNA modifications, *Bioinformatics*.
- Liu, M.L., et al. (2020) An Overview on Predicting Protein Subchloroplast Localization by using Machine Learning Methods, *Current protein & peptide science*.
- Livingston, F. (2005) Implementation of Breiman's Random Forest Machine Learning Algorithm., *Mach Learn J Pap*, **2005**:ECE591Q.
- Luo, L.F. (1988) The degeneracy rule of genetic code, *Origins of life and evolution of the biosphere : the journal of the International Society for the Study of the Origin of Life*, **18**, 65-70.
- Lv, H., et al. (2014) CarSPred: a computational tool for predicting carbonylation sites of human proteins, *PloS one*, **9**, e111478.
- Lv, H.Q., et al. (2016) A computational method to predict carbonylation sites in yeast proteins, *Genetics and molecular research : GMR*, **15**.
- Manavalan, B., et al. (2019) mAHTPred: a sequence-based meta-predictor for improving the prediction of anti-hypertensive peptides using effective feature representation, *Bioinformatics*, **35**, 2757-2765.
- Manavalan, B., et al. (2019) Meta-4mCpred: A Sequence-Based Meta-Predictor for Accurate DNA 4mC Site Prediction Using Effective Feature Representation, *Mol Ther Nucleic Acids*, **16**, 733-744.
- Moller, I.M., Rogowska-Wrzesinska, A. and Rao, R.S. (2011) Protein carbonylation and metal-catalyzed protein oxidation in a cellular perspective, *Journal of proteomics*, **74**, 2228-2242.
- Rao, R.S. and Moller, I.M. (2011) Pattern of occurrence and occupancy of carbonylation sites in proteins, *Proteomics*, **11**, 4166-4173.
- Rauniar, N., Prokai-Tatrai, K. and Prokai, L. (2010) Identification of carbonylation sites in apomyoglobin after exposure to 4-hydroxy-2-nonenal by solid-phase enrichment and liquid chromatography-electrospray ionization tandem mass spectrometry, *Journal of mass spectrometry : JMS*, **45**, 398-410.
- Reddy, V.P., et al. (2009) Oxidative stress in diabetes and Alzheimer's disease, *Journal of Alzheimer's disease : JAD*, **16**, 763-774.
- Smith, T.C. and Frank, E. (2016) Introducing Machine Learning Concepts with WEKA, *Methods in molecular biology*, **1418**, 353-378.
- Song, J.N., et al. (2010) Cascleave: towards more accurate prediction of caspase substrate cleavage sites, *Bioinformatics*, **26**, 752-760.
- Tan, J.X., et al. (2019) Identification of hormone binding proteins based on machine learning methods, *Mathematical Biosciences and Engineering*, **16**, 2466-2480.
- Tang, H., Chen, W. and Lin, H. (2016) Identification of immunoglobulins using Chou's pseudo amino acid composition with feature selection technique, *Molecular bioSystems*, **12**, 1269-1275.
- Tang, H., et al. (2018) HBPred: a tool to identify growth hormone-binding proteins, *Int J Biol Sci*, **14**, 957-964.
- Vacic, V., Iakoucheva, L.M. and Radivojac, P. (2006) Two Sample Logo: a graphical representation of the differences between two sets of sequence alignments, *Bioinformatics*, **22**, 1536-1537.
- Wang, M.J., et al. (2014) Cascleave 2.0, a new approach for predicting caspase and granzyme cleavage targets, *Bioinformatics*, **30**, 71-80.
- Weng, S.L., et al. (2017) Investigation and identification of protein carbonylation sites based on position-specific amino acid composition and physicochemical features, *BMC bioinformatics*, **18**, 66.
- Xu, Y., et al. (2014) Prediction of posttranslational modification sites from amino acid sequences with kernel methods, *Journal of theoretical biology*, **344**, 78-87.
- Yang, W., et al. (2019) A brief survey of machine learning methods in protein sub-Golgi localization, *Curr Bioinform*, **14**, 234-240.
- Zeng, X., et al. (2019) Predicting Disease-associated Circular RNAs Using Deep Forests Combined with Positive-Unlabeled Learning Methods, *Briefings in Bioinformatics*, Doi: 10.1093/bib/bbz1080.
- Zhao, Y.W., et al. (2016) Prediction of phosphothreonine sites in human proteins by fusing different features, *Scientific reports*, **6**, 34817.