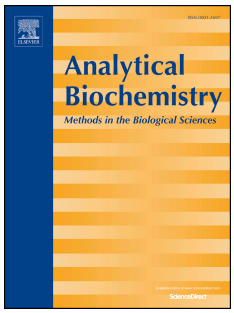
Accepted Manuscript



predCar-site: Carbonylation sites prediction in proteins using support vector machine with resolving data imbalanced issue

Md Al Mehedi Hasan, Jinyan Li, Shamim Ahmad, Md Khademul Islam Molla

PII: S0003-2697(17)30120-3

DOI: 10.1016/j.ab.2017.03.008

Reference: YABIO 12649

To appear in: *Analytical Biochemistry*

Received Date: 9 December 2016

Revised Date: 26 February 2017

Accepted Date: 7 March 2017

Please cite this article as: M.A.M. Hasan, J. Li, S. Ahmad, M.K.I. Molla, predCar-site: Carbonylation sites prediction in proteins using support vector machine with resolving data imbalanced issue, *Analytical Biochemistry* (2017), doi: 10.1016/j.ab.2017.03.008.

predCar-Site: Carbonylation Sites Prediction in Proteins Using Support Vector Machine with Resolving Data Imbalanced Issue

Md. Al Mehedi Hasan[1], Jinyan Li[2], Shamim Ahmad[1], Md. Khademul Islam Molla[1]

[1]Department of Computer Science & Engineering, University of Rajshahi, Bangladesh
[2]Advanced Analytics Institute and Centre for Health Technologies, University of Technology Sydney, Australia

E-mail: mehedi_ru@yahoo.com, Jinyan.Li@uts.edu.au, shamim_cst@yahoo.com, khademul.cse@ru.ac.bd

Abstract

The carbonylation is found as an irreversible post-translational modification and considered a biomarker of oxidative stress. It plays major role not only in orchestrating various biological processes but also associated with some diseases such as Alzheimer's disease, diabetes, and Parkinson's disease. However, since the experimental technologies are costly and time-consuming to detect the carbonylation sites in proteins, an accurate computational method for predicting carbonylation sites is an urgent issue which can be useful for drug development. In this study, a novel computational tool termed predCar-Site has been developed to predict protein carbonylation sites by (1) incorporating the sequence-coupled information into the general pseudo amino acid composition, (2) balancing the effect of skewed training dataset by Different Error Costs method, and (3) constructing a predictor using support vector machine as classifier. This predCar-Site predictor achieves an average AUC (area under curve) score of 0.9959, 0.9999, 1, and 0.9997 in predicting the carbonylation sites of K, P, R, and T, respectively. All of the experimental results along with AUC are found from the average of 5 complete runs of the 5-fold cross-validation and those results indicate significantly better performance than existing predictors. A user-friendly web server of predCar-Site is available at http://research.ru.ac.bd/predCar-Site/

Kewords

Carbonylation Sites Prediction, Sequence-coupling model, General PseAAC, Data Imbalance Issue, Different Error Costs , Support Vector Machine, predCar-Site Web-server

1. Introduction

The structural and functional diversities of proteins as well as plasticity and dynamics of living cells are significantly dominated by the post-translational modifications (PTMs) [1]. Not only that, PTMs are also responsible for expanding the genetic code and for regulating cellular physiology as well. [2, 3]. A variety of PTMs such as hydroxylation, nitration, sulfhydrylation, carbonylation and glutathionylation have been induced from Oxidative stress [4] which is the direct result of imbalance in the production and degradation of reactive oxygen species (ROS) and reactive nitrogen species (RNS) [5]. Oxidative stress may occur when an excess production of reactive oxygen species (ROS) has surpassed the detoxification ability of cells and weakened the damage-repairing ability [5, 6, 7, 8].

Among a variety of oxidative stress-induced PTMs, the protein carbonylation has been considered as a biomarker for oxidative stress due to its some unique characteristics such as relatively early formation, stability, and irreversibility [9, 10]. However, the density of protein carbonylation increases with increase of external oxidative stress, aging and obesity which provides an indication of early stage of diseases [11, 12]. Various types of major human diseases including Alzheimer's disease, diabetes, Parkinson's disease, chronic renal failure, chronic lung disease, sepsis are associated with protein carbolnylation [9, 13].

As a result, the identification of carbonylation sites in proteins has become a vital question in cellular physiology and pathology, which in turns, helps in providing some valuable evidence for both biomedical research and drug development [5, 6].

Mass spectrometry and liquid chromatography are the most common techniques to analyze protein susceptibility of the oxidative PTMs and determine its exact carbonylation sites recently [8, 14, 15]. It should be mentioned that among all the residues of protein molecules, four types of amino acid residues, namely lysine (K), proline (P), arginine (R), and threonine (T), have been found susceptible to carbonylation [16–18]. However, the purely experimental technique to determine the exact modified sites of carbonylated substrates is expensive as well as time-consuming, especially for large-scale datasets [8, 14].

In this context, it is highly demanded to use computational approaches to identify the carbonylated sites effectively and accurately [5, 6]. Recently various types of computational classifiers have been developed to identify carbonylation sites through different types of machine learning algorithms [5, 6, 19, 20]. However, in order to meet the current demand to produce efficient high-throughput tools, additional effort are required to further improve the prediction quality [5, 6].

In the development of computational classifier, one of the major challenges is to handle imbalance dataset problem [6, 21], as it is found in most of the dataset for this kind of prediction, the number negative subset is much larger than the corresponding positive subset [6, 21]. As the real world picture is that the non-carbonylation sites are always the majority compared with the carbonylation ones, so naturally the predictor should be biased to the non-carbonylation sites. Here the problem is that, for this type of predictors may interpret many carbonylation sites as non-carbonylation sites [22, 23, 24]. But, the information about the carbonylation sites is mostly desired than non-carbonylation sites. As a result, it is crucial to find an effective solution to balance this kind of bias consequence.

The current study has begun with an attempt to address the problems mentioned above and then tried to develop a more powerful predictor using support vector machine. In our predictor, the Different Error Costs (DEC) method [25, 26, 27] has been used to resolve the data imbalance issue. It should be noted here that the features used in this predictor are extracted by using vctorized sequence-coupling model [28]. In the recent works, the performance of PTMPred [19], CarSpred [5], and iCar-PseCp [6] on a large set of proteins has been studied in [6]. Therefore, in order to compare the performance of predCar-Site with those systems (PTMPred [19], CarSpred [5], and iCar-PseCp [6]), we use the exactly same dataset employing the commonly used stratified 10-fold cross-validation [6]. Since the information about the exact 10-way splits used in previous studies [6] is not available, so we have performed five complete runs of the 10-fold-crossvalidation, where each complete run of the 10-fold cross-validation uses a different 10-way splits. The use of multiple runs with different splits helps to validate the stability and the statistical significance of the results. Finally, the average results of all metrics found from this study has been reported. Our experimental results indicate that predCar-Site achieves significantly better results than those found from other top systems (PTMPred [19], CarSpred [5], and iCar-PseCp [6]).

In order to launch a useful sequence-based statistical predictor for a biological system as demonstrated in a series of recent publications [6, 21, 30-37] , the Chou's five-step rules [29] should be followed: (i) construct or select a valid benchmark dataset to train and test the predictor, (ii) formulate the biological sequence samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the target to be predicted, (iii) introduce or develop a powerful algorithm (or engine) to operate the prediction, (iv) properly perform cross-validation tests to objectively evaluate its anticipated accuracy, and (v) establish a user-friendly webserver or software package that is accessible to the public.

2. Material and Methods

2.1 Benchmark Dataset

iCar-PseCp's [6] benchmark dataset set has been used in this study. iCar-PseCp's dataset was derived from the 230 carbonylated protein sequences from human [15, 38-41] and 20 carbonylated protein sequences from Photobacterium and Escherichia coli [17, 39, 42, 43].

In iCar-PseCp [6], according to Chou's scheme, a peptide sample was generally expressed as

$$P_{\xi}(\circledast) = R_{-\xi}R_{-(\xi-1)} \dots \dots R_{-2}R_{-1}\circledast R_1R_2 \dots \dots R_{+(\xi-1)}R_{+\xi} \quad (1)$$

where the symbol ⊛ represents a sinlge amino acid code K, P, R, or T, the subscript ξ is an integer, $R_{-\xi}$ represents the ξ-th up stream amino acid residue from the center, the $R_{+\xi}$ represents the ξ-th downstream amino acid residue, and so forth.

The $(2\xi + 1)$-tuple peptide sample $P_\xi(\circledast)$ was further classified into the following two categories [6]

$$P_\xi(\circledast) \in \begin{cases} P_\xi^+(\circledast), & \text{if its center is a carbonylation site} \\ P_\xi^-(\circledast), & \text{otherwise} \end{cases} \quad (2)$$

where $P_\xi^+(\circledast)$ denotes a true carbonylation segment with K, P, R, or T at its center, $P_\xi^-(\mathbb{K})$ a false carbonylation segment with K, P, R, or T at its center, and the symbol ∈ means "a member of" in the set theory.

The benchmark dataset $S_\xi(\circledast)$ in iCar-PseCp's study was formulated as

$$\begin{cases} S_\xi(K) = S_\xi^+(K) \cup S_\xi^-(K), & when \circledast = K \\ S_\xi(P) = S_\xi^+(P) \cup S_\xi^-(P), & when \circledast = P \\ S_\xi(R) = S_\xi^+(R) \cup S_\xi^-(R), & when \circledast = R \\ S_\xi(T) = S_\xi^+(T) \cup S_\xi^-(T), & when \circledast = T \end{cases} \quad (3)$$

where the positive subset of $S_\xi^+(\circledast)$ only contains the samples of true carbonylation segments $P_\xi^+(\circledast)$, and negative subset $S_\xi^-(\circledast)$ only contains the samples of false carbonylation segments $P_\xi^-(\circledast)$ and the symbol ∪ means "union" in the set theory.

In iCar-PseCp's work, $(2\xi + 1)$-tuple peptide window was used to collect peptide segment that had K, P, R, or T at the center. It should be mentioned here that if the upstream or downstream in a protein sequence is less than ξ or greater than L-ξ (L is the length of the protein sequence concerned) then the lacking amino acid has been filled with a dummy residue X in iCar-PseCp [6].

After applying some screening procedure based on some constraints on that collected peptide samples, for example, considering window size, <= 30% pairwise sequence identity to any other peptides, iCar-PseCp finally constructed a benchmark dataset [6]. The detail procedure about the construction of iCar-PseCp's benchmark dataset is explained in [6].

Note that, depending on some preliminary test, window size was selected as 15 (2*ξ+1) in iCar-PseCp's study, where ξ=7. Thus, the benchmark dataset obtained by iCar-PseCp for $S_{\xi=7}(K)$, $S_{\xi=7}(P)$, $S_{\xi=7}(R)$, and $S_{\xi=7}(T)$ are available at online supplementary materials (http://research.ru.ac.bd/predCar-Site/) as Supporting Information S1, S2, S3, and S4, respectively. It should be mention that our published online supplementary materials are taken from iCar-PseCp's work [6]. A summary of this benchmark dataset is given in Table 1.

Table 1. Summary of Carbonylation Site Samples in the Benchmark Dataset

Subset	Carbonylation Type and Number of Samples			
	⊛=K	⊛=P	⊛=R	⊛=T
Positive	300	126	136	121
Negative	1949	792	847	732

2.2 Feature Extraction

The appropriate features of protein sequences or samples plays very important roles for the prediction of carbonylation site, as a result it draws the much attention of scientist that how to select the core and essential features of protein samples. As most existing machine learning algorithm can handle only vector but not sequence sample, one of the critical problem in bioinformatics is how to extract vector from biological sequence with keeping

considerable sequence characteristics [37]. In this paper, to avoid complete losing the sequence pattern information for protein, the Chou's general PseAAC [29] has been adopted to extract feature from peptide segment using sequence-coupling model [28, 44] which has been described briefly below.

Now, based on the concept of sequence-coupled information [28, 44] into the general PseAAC, the peptide sequence of Eq. (1) can be formulated as

$$P_{\xi}(\circledast) = P_{\xi}^{+}(\circledast) - P_{\xi}^{-}(\circledast) \tag{4}$$

where

$$P_{\xi}^{+}(\circledast) = \begin{bmatrix} p_{-\xi}^{+}\left(R_{-\xi}\middle|R_{-(\xi-1)}\right) \\ p_{-(\xi-1)}^{+}\left(R_{-(\xi-1)}\middle|R_{-(\xi-2)}\right) \\ \cdots \\ \cdots \\ p_{\xi}^{+}(R_{-2}|R_{-1}) \\ p_{\xi}^{+}(R_{-1}) \\ p_{\xi}^{+}(R_{+1}) \\ p_{\xi}^{+}(R_{+2}|R_{+1}) \\ \cdots \\ \cdots \\ p_{+(\xi-1)}^{+}\left(R_{+(\xi-1)}\middle|R_{+(\xi-2)}\right) \\ p_{+\xi}^{+}\left(R_{+\xi}\middle|R_{+(\xi-1)}\right) \end{bmatrix} \tag{5}$$

and

$$P_{\xi}^{-}(\circledast) = \begin{bmatrix} p_{-\xi}^{-}\left(R_{-\xi}\middle|R_{-(\xi-1)}\right) \\ p_{-(\xi-1)}^{-}\left(R_{-(\xi-1)}\middle|R_{-(\xi-2)}\right) \\ \cdots \\ \cdots \\ p_{\xi}^{-}(R_{-2}|R_{-1}) \\ p_{\xi}^{-}(R_{-1}) \\ p_{\xi}^{-}(R_{+1}) \\ p_{\xi}^{-}(R_{+2}|R_{+1}) \\ \cdots \\ \cdots \\ p_{+(\xi-1)}^{-}\left(R_{+(\xi-1)}\middle|R_{+(\xi-2)}\right) \\ p_{+\xi}^{-}\left(R_{+\xi}\middle|R_{+(\xi-1)}\right) \end{bmatrix} \tag{6}$$

In Eq. (5) $p_{-\xi}^{+}\left(R_{-\xi}\middle|R_{-(\xi-1)}\right)$ is the conditional probability of amino acid $R_{-\xi}$ occurring at the left 1st position (see Eq. (1)) given that its closest right neighbor is $R_{-(\xi-1)}$, $p_{-(\xi-1)}^{+}\left(R_{-(\xi-1)}\middle|R_{-(\xi-2)}\right)$ is the conditional probability of amino acid $R_{-(\xi-1)}$ occurring at the left 2nd position given that its closest right neighbor is $R_{-(\xi-2)}$, and so forth. It should be mentioned here that in Eq. (5), only $p_{\xi}^{+}(R_{-1})$ and $p_{\xi}^{+}(R_{+1})$ are of non-conditional probability since the right neighbor of R_{-1} and the left neighbor of R_{+1} are always ⊛ (K, P, R, or T). All these probability values can be easily derived from the positive benchmark dataset ($S_{\xi=7}^{+}(\circledast)$) given in Supporting Information S1, S2, S3, and S4, respectively as done in [44]. Likewise, the components in Eq. (6) are the same as those in Eq. (5) except for that they are derived from the negative benchmark dataset given in Supporting Information S1, S2, S3, and S4, respectively.

2.3 SVM Classification

Consider the problem of separating the set of training vectors belong to two separate classes, $(x_1, y_1), (x_2, y_2), \ldots\ldots, (x_n, y_n)$ where $x_i \in R^p$ and $y_i \in \{-1, +1\}$ is the corresponding class label, $1 \leq i \leq n$. The main task of this problem is to find a classifier with a decision function f(x, θ) such that y = f(x, θ), where y is the class label for x and θ is a vector of unknown parameters of the decision function. The support vector machine is a

well-known classifier and it has been applied broadly in many classifications problems. Geometrically, the SVM modeling algorithm finds an optimal hyperplane with the maximal margin to separate two classes [45], which requires solving the following constraint problem:

$$\text{minimize}_{w,b}\ \frac{1}{2}\|w\|^2$$

Subject to:

$$y_i(w^T x_i + b) \geq 1, i = 1,2,3, \ldots\ldots, n \quad (7)$$

To allow errors, the optimization problem now becomes:

$$\min_{w,b}\ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i$$

Subject to:

$$y_i(w^T x_i + b) \geq 1 - \xi_i, i = 1,2,3, \ldots\ldots, n \quad (8)$$
$$\xi_i \geq 0\ , i = 1,2,3, \ldots\ldots, n$$

Using the method of Lagrange multipliers, we can obtain the dual formulation which is expressed in terms of variables α_i [45, 46]:

$$\text{maximize}_{\alpha}\sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j x_i^T x_j$$

Subject to: $\sum_{i=1}^{n} y_i\alpha_i = 0,\ 0 \leq \alpha_i \leq C$

for all $i = 1,2,3, \ldots\ldots, n$ (9)

Finally, the linear discriminant function takes the following form

$$f(x) = \sum_i^n \alpha_i x_i^T x + b \quad (10)$$

In many applications a non-linear classifier provides better accuracy. In SVM, the naive way of making a non-linear classifier out of a linear classifier is to map our data from the input space X to a feature space F using a non-linear function $\emptyset: X \rightarrow F$. In the space F, the optimization takes the following form using kernel function [45, 46, 47]:

$$\text{maximize}_{\alpha}\sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j k(x_i, x_j)$$

Subject to:

$$\sum_{i=1}^{n} y_i\alpha_i = 0\ , 0 \leq \alpha_i \leq C$$
for all $i = 1,2,3, \ldots\ldots, n$ (11)

Now, in terms of the kernel function the discriminant function takes the following form:

$$f(x) = \sum_i^n \alpha_i k(x, x_i) + b \quad (12)$$

It noted here that a kernel function and its parameter have to be chosen to build a SVM classifier [45, 46, 47]. In this work, radial basis function kernel has been used to build SVM classifier which is defined below:

$K(x_i, x_j) = \exp(-\frac{\|x_i - x_j\|^2}{\sigma})$, σ is the width of the function

2.3.1 Imbalance Dataset Problem Management

Any data set that shows an unequal distribution between its classes can be considered as imbalanced dataset problem. The main challenge in imbalance problem is that the small classes are often more useful, but standard classifiers tend to be weighed down by the huge classes and ignore the tiny ones. Although SVMs work effectively with balanced datasets, they provide sub-optimal models with imbalanced datasets [25, 26]. The main reason for the

SVM algorithm to be sensitive to class imbalance would be that the soft margin objective function given in Eq. (11) assigns the same cost (i.e., C) for both positive and negative misclassifications in the penalty term [27].

In this paper, we have used a Different Error Costs (DEC) method to handle imbalance dataset problem of carbonylation sites prediction. The Different Error Costs (DEC) method is a cost-sensitive learning solution proposed in [25] to overcome the imbalance dataset problem for SVM. In DEC method, the soft margin objective function of SVM is modified to assign two misclassification costs, such that C^+ is the misclassification cost for positive class examples, while C^- is the misclassification cost for negative class examples. In our work, the following equations give the cost for the positive and negative classes

$$C^+ = \frac{N}{2*N_1}, \qquad C^- = \frac{N}{2*N_2} \tag{13}$$

where N is the total number of instances, N_1 is the number of instances for positive class, and N_2 is the number of negative class.

It should be noted that we have used Matlab 2014b version to implement our system where the *trainsvm* function of Matlab by default uses DEC with the same cost defined in Eq. (13) to handle imbalance situation.

2.4 Experimental Setting

In statistical prediction, there are three commonly used methods to derive the metric values for a predictor: the independent dataset test, subsampling (e.g., k-fold cross-validation) test, and jackknife test [6, 37, 48]. These methods are often used for testing the accuracy of a statistical prediction algorithm. However, among those three methods, the jackknife test is deemed the most objective because it can always yield a unique result for a given benchmark data set, as reported in a comprehensive review [29]. Although the jackknife test has been increasingly and widely adopted by investigators to examine the power of various prediction methods, it takes huge computational time for a larger dataset.

In this study, we have used k-fold cross-validation (subsampling) method to save the computational time. As the information about the exact 10-way splits of dataset used in previous studies is not published [6], therefore, in order to validate the stability and the statistical significance of our results, we have repeated the 10-fold cross-validation for 5 times. It can be mentioned here that in each 10-fold cross-validation the given training samples are randomly partitioned into 10 mutually exclusive sets of approximately equal size and approximately equal class distribution. Finally, we have reported the average results of all metrics in this study.

It can be mentioned here that all the trains and tests have been conducted on a standard machine of DELL Optiplex 390 with 8 GB RAM and Core-i3 processor running at 3.30 GHz.

2.5 Measuring Metrics

For measuring the predictive capability and reliability for this kind of classification, a set of four metrics is usually used in the literature: (i) overall accuracy or Acc, (ii) Mathew's correlation coefficient or MCC, (iii) sensitivity or Sn, and (iv) specificity or Sp [5, 49, 50, 51].

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

$$Sp = \frac{TN}{TN+FP}$$

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

$$MCC = \frac{(TP\times TN)-(FP\times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{14}$$

where TP (true positive) denotes the number of carbonylated peptides correctly predicted, TN (true negative) the numbers non-carbonylated peptides correctly predicted, FP (false positive) the non-carbonylated incorrectly predicted as the carbonylated peptides, and FN (false negative) the carbonylated peptides incorrectly predicted as the non-carbonylated peptides.

However, in addition to these four metrics, we have used the measure of precision since it is one of the most important measurements to evaluate the degree of credibility of a prediction system. The precision is defined as

$$precision = \frac{TP}{TP+FP} \qquad (15)$$

where the meaning of TP and FP is defined in Eq. (14).

At last, we have included the metric AUC (area under the curve) in order to evaluate our system. The AUC can be calculated from ROC curve (receiver operating characteristic curve) as well.

3. Results and Discussion

3.1 Model Selection for SVM

In order to generate highly performing SVM classifiers capable of dealing with real data an efficient model selection is required. In our experiment, grid-search technique has been used to find the best model for SVM. In our experiments, this method selects the values of parameters considering highest performance which will be measured using a specific metric (AUC, in this case) and then time if more than one position in search space has the same performance. We have performed 5 complete runs of the 10 fold cross-validation and each time we have selected the best parameter of the classifier on basis of the value of AUC (area under the curve).

Table 2. Selected C and σ of 5 times run of 5 folds cross-validations for RBF kernel

No. of Completes Run	**Type of Carbonylation**							
	K		P		R		T	
	C	σ	**C**	σ	**C**	σ	**C**	σ
1st	2^5	2^4	2^6	2^3	2^{-1}	2^3	2^{-2}	2^3
2nd	2^6	2^4	2^6	2^3	2^2	2^3	2^1	2^3
3rd	2^5	2^4	2^5	2^3	2^2	2^3	2^1	2^3
4th	2^6	2^4	2^6	2^3	2^2	2^3	2^{-2}	2^3
5th	2^5	2^4	2^4	2^3	2^{-2}	2^2	2^{-2}	2^3

It noted here that depending on the four types of residues (K, P, R, or T) which are susceptive to carbonylation, four times model selection has been considered. If the center residue of a query peptide is $\circledast = K$ then the corresponding training data must be taken from $S_{\xi=7}(K)$ if the center residue of a query peptide is $\circledast = P$, then the training data must be taken from $S_{\xi=7}(P)$ and so forth.

Table 3. Final Selection of C and σ to Train the System for Web Server

Type of Carbonylation	**C**	$\boldsymbol{\sigma}$
K	2^5	2^4
P	2^6	2^3
R	2^2	2^3
T	2^{-2}	2^3

For radial basis function (RBF) kernel, to find the parameter value *C* (penalty term for soft margin) and σ (sigma), we have considered the value from 2^{-8} to 2^{8} for *C* and from 2^{-8} to 2^{8} for sigma as our searching space. The selected C and sigma of 5 complete runs of the 10-fold cross-validation on each types (dataset depending on K, P, R, or T) of training dataset is shown in Table 2. Finally, we have averaged our results in order to ensure unbiased model selection.

It should be mentioned that we have used that value of C and sigma which appears most of the times as best model in 5 complete runs of the 10-fold cross-validation to train the system for the web server. Considering the mentioned criteria, the selected C and sigma for each type of residue (K, R, P, or T) is given in Table 3.

3.2 Comparison with the Existing Methods

The values of the four metrics (cf. Eq. (14)) and the value of AUC obtained by the current predCar-Site predictor for K-, P-, R-, and T-type carbonylation are given in the Table 4. These values are the average result of 5 complete runs of the 10 fold cross-validation on the benchmark dataset given in Supporting Information S1, S2, S3 and S4 respectively. In addition to it, standard deviations of each metrics of 5 complete runs of the 10-fold cross-validation are shown in parentheses.

The Table 4 also includes the corresponding rates achieved by PTMPred [19], CarSpred [5], and iCar-PseCp[6], the three existing predictors for identifying the carbonylation sites in the aforesaid benchmark dataset. It should be mentioned here that the performance of PTMPred [19], CarSpred [5], and iCar-PseCp[6] as shown in Table 4 are noted from [6].

Table 4: A Comparison of the Proposed Predictor with the Existing Methods on the Same 250 Carbonylated Proteins

Predictor	Metrics	Type of Carbonylation			
		K	P	R	T
PTMPred	**Acc (%)**	88.59	82.93	86.64	88.39
CarSpred		87.22	82.93	86.22	86.61
iCar-PseCp		84.43	86.79	84.23	86.17
predCar-Site		96.95 (±0.10)	99.61(±0.09)	99.10(±0.26)	99.11(±0.16)
PTMPred	**MCC**	0.1892	0.2573	0.1878	0.2186
CarSpred		0.2268	0.2331	0.2245	0.2040
iCar-PseCp		0.5906	0.6006	0.6076	0.6185
predCar-Site		0.8799(±0.0034)	0.9837(±0.0039)	0.9642(±0.0101)	0.9646(±0.0059)
PTMPred	**Sn (%)**	23.45	21.43	20.02	22.38
CarSpred		23.17	25.34	25.47	21.39
iCar-PseCp		45.18	48.20	46.67	50.68
predCar-Site		96.67(±0.33)	99.68(±0.43)	1	99.34(±0.69)
PTMPred	**Sp (%)**	92.99	93.20	90.99	91.36
CarSpred		92.43	93.28	93.39	93.42
iCar-PseCp		99.25	98.54	99.57	98.58
predCar-Site		96.99(±0.14)	99.60(±0.14)	98.96(±0.31)	99.07(±0.22)
PTMPred	**AUC**	0.6858	0.6903	0.5981	0.6563
CarSpred		0.6849	0.7163	0.7158	0.7134
iCar-PseCp		0.8728	0.8484	0.8668	0.8603
predCar-Site		0.9959(±0.00002)	0.9999(±0.000005)	1	0.9997(±0.000005)

It is obvious from the Table 4, predCar-Site has performed remarkably better over PTMPred [19], CarSpred [5], and iCar-PseCp[6] while considering Acc, MCC, and Sn. It indicates that, the proposed new predictor has produced over all better accuracy, sensitivity, and stability. Although the achieved Sp by iCar-PseCp is higher than that by our predictor in the case of center residues K and R, the gap between its Sn and Sp is very large (54% for K, 53% for R). Which implies that the results achieved by iCar-PseCp contain many false negative events [52] and hence its higher achieved Sp rate is problematic.

The area under the ROC curve is called AUC (area under the curve). The greater the AUC value is, the better the predictor will be [53, 54]. As we can see from Table 4, the value of AUC clearly indicates that the proposed predictor is better than PTMPred [19], CarSpred [5], and iCar-PseCp [6]. Therefore, it is projected that predCar-Site may become a useful and higher throughput tool in carbonylation sites predictions.

Apart from the above mentioned metrics, we have calculated precision too for our system and got the average (± standard deviation) values of 83.19(±0.62)%, 97.52(±0.82)%, 93.95(±1.67)%, and 94.66(±1.19)% in predicting the carbonylation sites for K, P, R, and T, respectively. Since the values of precision for the other systems (PTMPred [19], CarSpred [5], and iCar-PseCp [6]) are not publicly available, as a result, we could not show those findings. The achieved values of precision of our system is very promising and encouraging. Note that precision measures how much believable the system is when it says a peptides sample is carbonylated.

Why can the proposed method enhance the prediction quality so significantly? First, the coupling effects among the amino acids around the target sites have been taken into account via the conditional probability. Second, the predictor used Different Error Costs (DEC) method to balance the effect of skewed training dataset.

3.3 Protocol Guide

To attract more users especially for the convenience of experimental scientists and enhance the value of practical application, a user-friendly web-server for predCar-Site has been established at http://research.ru.ac.bd/predCar-Site/. A step-by-step guide on how to use the web server is given below:

Step 1. Open the web server at http://research.ru.ac.bd/predCar-Site/ and you will find the home page of the predictor on your display as shown in Fig. 1. Once you click on the Read Me button, you will get a brief introduction about predCar-Site predictor.

Step 2. You will have to either type or copy and paste the query protein sequence into the input text box at the center of Fig. 1. The input sequence should follow the FASTA format. The example of a sequence of FASTA format is available by clicking at example button located right above the input text box.

Step 3. In order to get the predicted result, at first you have to check one of the four options (K, P, R, or T) and then click on the Submit button. For example, if you use the Sequence_K query protein sequence given under Example button as input and check on the K button, it will take 20s or more from the time of your submission to get desired output. All the predicted results of each lysine (K) are presented in each row of a table.

Step 4. In order to get batch prediction, you will have to enter desired batch input file (in FASTA format of course) via Browse button located on the lower panel, as shown in Fig. 1.

Step 5. To download the benchmark datasets used to train and test the predCar-Site predictor, click on the Supporting Information button as shown in Fig. 1.

Step 6. You can get the relevant papers that document the detailed development and algorithm of predCar-Site by clicking the Citation button as shown in Fig. 1.

4. Conclusion

In this article, we have designed a simple and efficient predictor predCar-Site for predicting carbonylation sites. Experimental results show that our method is very promising and can be a useful tool for prediction of carbonylation sites. The predCar-Site has achieved remarkably higher success rates in comparison with the existing predictors in this area. In addition to it, we have established a user-friendly web server and provided step by step guide for convenience of the experimental scientists. It provides as easier way to obtain the desired results without knowing the mathematical details. We have projected that the predCar-Site will become a very useful and higher throughput tool for predicting of protein carbonylation sites.

References

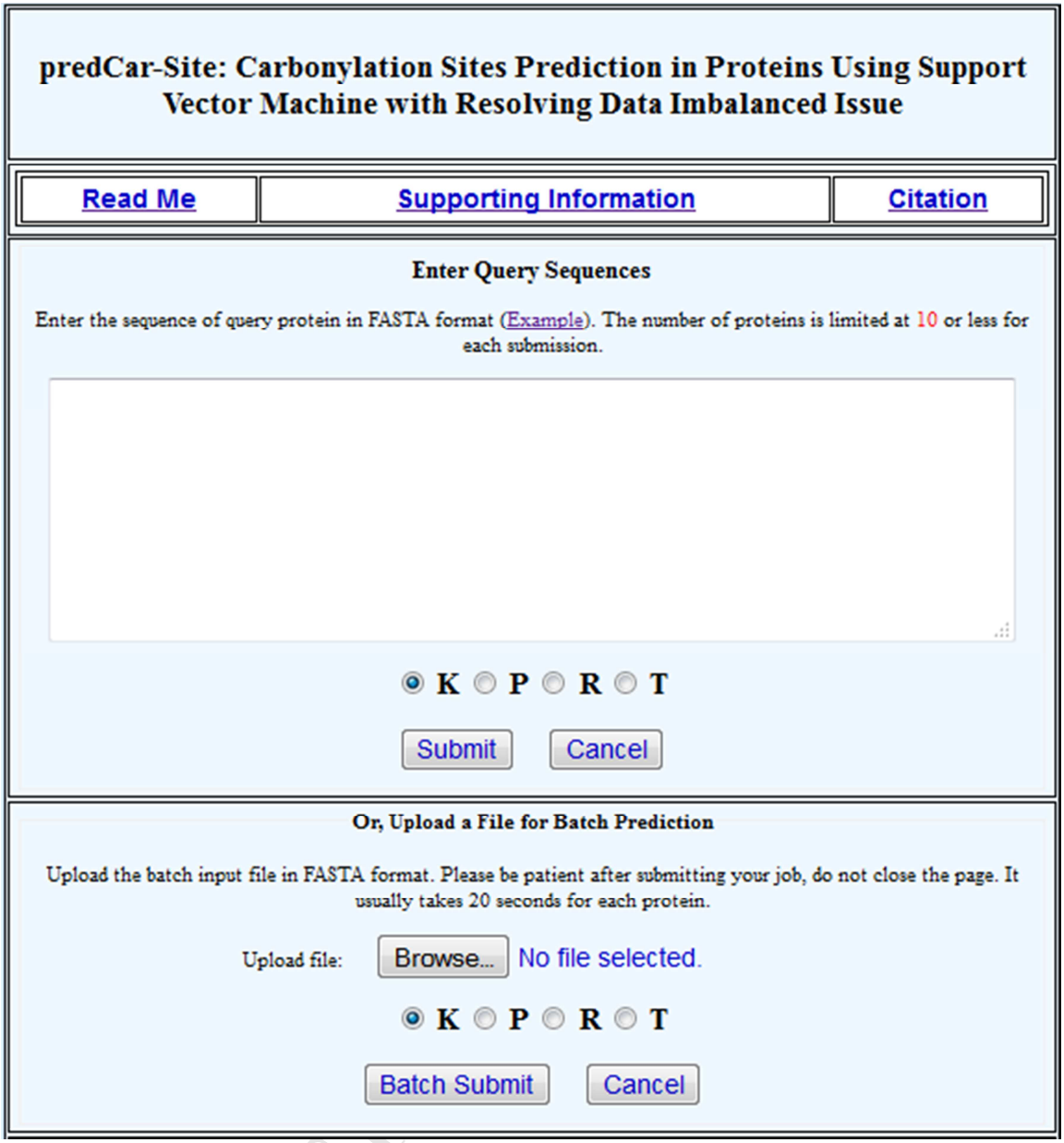
1. Xu, Y., Ding, J., Wu, L. Y., Chou, K. C., 2013. iSNO-PseAAC: predict cysteine S-nitrosylation sites in proteins by incorporating position specific amino acid propensity into pseudo amino acid composition. PLoS One, 8(2), e55844.
2. Walsh, C. T., Garneau-Tsodikova, S., Gatto, G. J., 2005. Protein posttranslational modifications: the chemistry of proteome diversifications. Angewandte Chemie International Edition, 44(45), 7342-7372.
3. Witze, E. S., Old, W. M., Resing, K. A., Ahn, N. G., 2007. Mapping protein post-translational modifications with mass spectrometry. Nature methods, 4(10), 798-806.
4. Gianazza, E., Crawford, J., Miller, I., 2007. Detecting oxidative post-translational modifications in proteins. Amino Acids, 33(1), 51-56.
5. Lv, H., Han, J., Liu, J., Zheng, J., Liu, R., Zhong, D., 2014. CarSPred: a computational tool for predicting carbonylation sites of human proteins. PloS one, 9(10), e111478.
6. Jia, J., Liu, Z., Xiao, X., Liu, B., Chou, K. C., 2016. iCar-PseCp: identify carbonylation sites in proteins by Monto Carlo sampling and incorporating sequence coupled effects into general PseAAC. Oncotarget, 7(23), 34558-34570.
7. Reddy, V. P., Zhu, X., Perry, G., & Smith, M. A. (2009). Oxidative stress in diabetes and Alzheimer's disease. Journal of Alzheimer's Disease, 16(4), 763-774.
8. Bollineni, R. C., Hoffmann, R., Fedorova, M., 2011. Identification of protein carbonylation sites by two-dimensional liquid chromatography in combination with MALDI-and ESI-MS. Journal of proteomics, 74(11), 2338-2350.
9. Dalle-Donne, I., Giustarini, D., Colombo, R., Rossi, R., Milzani, A., 2003. Protein carbonylation in human diseases. Trends in molecular medicine, 9(4), 169-176.
10. Møller, I. M., Rogowska-Wrzesinska, A., Rao, R. S. P., 2011. Protein carbonylation and metal-catalyzed protein oxidation in a cellular perspective. Journal of proteomics, 74(11), 2228-2242.
11. Bota, D. A., Van Remmen, H., Davies, K. J., 2002. Modulation of Lon protease activity and aconitase turnover during aging and oxidative stress. FEBS letters, 532(1-2), 103-106.
12. Frohnert, B. I., Sinaiko, A. R., Serrot, F. J., Foncea, R. E., Moran, A., Ikramuddin, S., ..., Bernlohr, D. A., 2011. Increased adipose protein carbonylation in human obesity. Obesity, 19(9), 1735-1741.
13. Dalle-Donne, I., Aldini, G., Carini, M., Colombo, R., Rossi, R., Milzani, A., 2006. Protein carbonylation, cellular dysfunction, and disease progression. Journal of cellular and molecular medicine, 10(2), 389-406.
14. Colzani, M., Aldini, G., Carini, M., 2013. Mass spectrometric approaches for the identification and quantification of reactive carbonyl species protein adducts. Journal of proteomics, 92, 28-50.
15. Bollineni, R. C., Hoffmann, R., Fedorova, M., 2014. Proteome-wide profiling of carbonylated proteins and carbonylation sites in HeLa cells under mild oxidative stress conditions. Free Radical Biology and Medicine, 68, 186-195.
16. Stadtman, E. R., Levine, R. L., 2003. Free radical-mediated oxidation of free amino acids and amino acid residues in proteins. Amino acids, 25(3-4), 207-218.
17. Maisonneuve, E., Ducret, A., Khoueiry, P., Lignon, S., Longhi, S., Talla, E., Dukan, S., 2009. Rules governing selective protein carbonylation. PloS one, 4(10), e7269.
18. Rao, R., Møller, I. M., 2011. Pattern of occurrence and occupancy of carbonylation sites in proteins. Proteomics, 11(21), 4166-4173.

19. Xu, Y., Wang, X., Wang, Y., Tian, Y., Shao, X., Wu, L. Y., Deng, N., 2014. Prediction of posttranslational modification sites from amino acid sequences with kernel methods. Journal of theoretical biology, 344, 78-87.
20. Lv, H. Q., Liu, J., Han, J. Q., Zheng, J. G., Liu, R. L., 2016. A computational method to predict carbonylation sites in yeast proteins. Genetics and molecular research: GMR, 15(2).
21. Jia, J., Liu, Z., Xiao, X., Liu, B., Chou, K. C., 2016. iSuc-PseOpt: identifying lysine succinylation sites in proteins by incorporating sequence-coupling effects into pseudo components and optimizing imbalanced training dataset. Analytical biochemistry, 497, 48-56.
22. Liu, Z., Xiao, X., Qiu, W. R., Chou, K. C., 2015. iDNA-Methyl: identifying DNA methylation sites via pseudo trinucleotide composition. Analytical biochemistry, 474, 69-77.
23. Sun, Y., Wong, A. K., Kamel, M. S., 2009. Classification of imbalanced data: a review. International Journal of Pattern Recognition and Artificial Intelligence, 23(04), 687-719.
24. Nath, A., Karthikeyan, S., 2016. Enhanced Prediction and Characterization of CDK Inhibitors Using Optimal Class Distribution. Interdisciplinary Sciences: Computational Life Sciences, 1-12.
25. Veropoulos, K., Campbell, C., Cristianini, N., 1999. Controlling the sensitivity of support vector machines. In Proceedings of the International Joint Conference on Artificial Intelligence, pp. 55-60.
26. Akbani, R., Kwek, S., Japkowicz, N., 2004. Applying support vector machines to imbalanced datasets. in Proceedings of 15th European Conference on Machine Learning, Pisa, Italy, pp.39-50
27. Batuwita, R., Palade, V., 2010. Efficient resampling methods for training support vector machines with imbalanced datasets. The 2010 International Joint Conference on Neural Networks (IJCNN), Barcelona, pp. 1-8.
28. Chou, K. C., 1993. A vectorized sequence-coupling model for predicting HIV protease cleavage sites in proteins. Journal of Biological Chemistry, 268(23), 16938-16948.
29. Chou, K. C., 2011. Some remarks on protein attribute prediction and pseudo amino acid composition. Journal of theoretical biology, 273(1), 236-247.
30. Chen, D., Tian, X., Zhou, B., Gao, J., 2016. ProFold: Protein Fold Classification with Additional Structural Features and a Novel Ensemble Classifier. BioMed Research International, 2016.
31. Ju, Z., Cao, J. Z., Gu, H., 2016. Predicting lysine phosphoglycerylation with fuzzy SVM by incorporating k-spaced amino acid pairs into Chou′ s general PseAAC. Journal of theoretical biology, 397, 145-150.
32. Chen, P., Hu, S., Zhang, J., Gao, X., Li, J., Xia, J. Wang, B. 2016, A sequence-based dynamic ensemble learning system for protein Ligand-binding site prediction, IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 13, no. 5, pp. 901-912.
33. Qiu, W. R., Zheng, Q. S., Sun, B. Q., Xiao, X., 2016. Multi-iPPseEvo: A Multi-label Classifier for Identifying Human Phosphorylated Proteins by Incorporating Evolutionary Information into Chou′ s General PseAAC via Grey System Theory. Molecular Informatics. In Press.
34. Wang, X., Yan, R., Li, J., Song, J., 2016. SOHPRED: a new bioinformatics tool for the characterization and prediction of human S-sulfenylation sites. Molecular BioSystems, 12(9), 2849-2858.
35. Hu, J., Li, Y., Yang, J. Y., Shen, H. B., Yu, D. J., 2016. GPCR–drug interactions prediction using random forest with drug-association-matrix-based post-processing procedure. Computational biology and chemistry, 60, 59-71.
36. Hu, J., Han, K., Li, Y., Yang, J. Y., Shen, H. B., Yu, D. J., 2016. TargetCrys: protein crystallization prediction by fusing multi-view features with two-layered SVM. Amino acids, 48(11), 2533-2547.

37. Jia, J., Liu, Z., Xiao, X., Liu, B., Chou, K. C., 2016. pSuc-Lys: predict lysine succinylation sites in proteins with PseAAC and ensemble random forest approach. Journal of theoretical biology, 394, 223-230.
38. Ishii, T., Ito, S., Kumazawa, S., Sakurai, T., Yamaguchi, S., Mori, T., ..., Uchida, K., 2008. Site-specific modification of positively-charged surfaces on human serum albumin by malondialdehyde. Biochemical and biophysical research communications, 371(1), 28-32.
39. Madian, A. G., Diaz-Maldonado, N., Gao, Q., Regnier, F. E., 2011. Oxidative stress induced carbonylation in human plasma. Journal of proteomics, 74(11), 2395-2416.
40. Mirzaei, H., Regnier, F., 2006. Identification and quantification of protein carbonylation using light and heavy isotope labeled Girard's P reagent. Journal of Chromatography A, 1134(1), 122-133.
41. Temple, A., Yen, T. Y., Gronert, S., 2006. Identification of specific protein carbonylation sites in model oxidations of human serum albumin. Journal of the American Society for Mass Spectrometry, 17(8), 1172-1180.
42. Chavez, J. D., Bisson, W. H., Maier, C. S., 2010. A targeted mass spectrometry-based approach for the identification and characterization of proteins containing α-aminoadipic and γ-glutamic semialdehyde residues. Analytical and bioanalytical chemistry, 398(7-8), 2905-2914.
43. Mirzaei, H., Regnier, F., 2005. Affinity chromatographic selection of carbonylated proteins followed by identification of oxidation sites using tandem mass spectrometry. Analytical chemistry, 77(8), 2386-2392.
44. Chou, K. C., 1996. Prediction of human immunodeficiency virus protease cleavage sites in proteins. Analytical biochemistry, 233(1), 1-14.
45. Vapnik V. N., 1999. The Nature of Statistical Learning Theory, Second Edition, Springer, New York, ISBN 0-387-98780-0.
46. Scholkopf, B., Smola, A. J., 2001. Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT press. [6,7,22,64]
47. Hasan, M. A. M., Nasser, M., Pal, B., Ahmad, S., 2014. Support vector machine and random forest modeling for intrusion detection system (IDS). Journal of Intelligent Learning Systems and Applications, 6(1), 45.
48. Ju, Z., Cao, J. Z., Gu, H., 2016. Predicting lysine phosphoglycerylation with fuzzy SVM by incorporating k-spaced amino acid pairs into Chou′ s general PseAAC. Journal of theoretical biology, 397, 145-150.
49. Xu, Y., Ding, Y. X., Deng, N. Y., Liu, L. M., 2016. Prediction of sumoylation sites in proteins using linear discriminant analysis. Gene, 576(1), 99-104.
50. Liu, B., Liu, Y., Jin, X., Wang, X., Liu, B., 2016. iRSpot-DACC: a computational predictor for recombination hot/cold spots identification based on dinucleotide-based auto-cross covariance. Scientific Reports, 6.
51. Liao, Z., Ju, Y., Zou, Q. 2016. Prediction of G protein-coupled receptors with SVM-prot features and random forest. Scientifica, 2016.
52. Chen, J., Liu, H., Yang, J., Chou, K. C., 2007. Prediction of linear B-cell epitopes using amino acid pair antigenicity scale. Amino acids, 33(3), 423-428.
53. Fawcett, T., 2006. An introduction to ROC analysis. Pattern recognition letters, 27(8), 861-874.
54. Davis, J., Goadrich, M., 2006. The relationship between Precision-Recall and ROC curves. In Proceedings of the 23rd international conference on Machine learning (pp. 233-240). ACM.

Figure Legends

Figure 1. A semi-screenshot for the home page of the webserver predCar-Site at http://research.ru.ac.bd/predCar-Site/


predCar-Site: Carbonylation Sites Prediction in Proteins Using Support Vector Machine with Resolving Data Imbalanced Issue
Read Me
Supporting Information
Citation
Enter Query Sequences
Enter the sequence of query protein in FASTA format (Example). The number of proteins is limited at 10 or less for each submission.
K P R T
Submit
Cancel
Or, Upload a File for Batch Prediction
Upload the batch input file in FASTA format. Please be patient after submitting your job, do not close the page. It usually takes 20 seconds for each protein.
Upload file:
Browse...
No file selected.
K P R T
Batch Submit
Cancel