J Comput Aided Mol Des (2011) 25:987–995
DOI 10.1007/s10822-011-9477-2

Carboxylator: incorporating solvent-accessible surface area for identifying protein carboxylation sites

Cheng-Tsung Lu · Shu-An Chen · Neil Arvin Bretaña · Tzu-Hsiu Cheng · Tzong-Yi Lee

Received: 30 May 2011 / Accepted: 29 September 2011 / Published online: 22 October 2011

Abstract In proteins, glutamate (Glu) residues are transformed into γ-carboxyglutamate (Gla) residues in a process called carboxylation. The process of protein carboxylation catalyzed by γ-glutamyl carboxylase is deemed to be important due to its involvement in biological processes such as blood clotting cascade and bone growth. There is an increasing interest within the scientific community to identify protein carboxylation sites. However, experimental identification of carboxylation sites via mass spectrometry-based methods is observed to be expensive, time-consuming, and labor-intensive. Thus, we were motivated to design a computational method for identifying protein carboxylation sites. This work aims to investigate the protein carboxylation by considering the composition of amino acids that surround modification sites. With the implication of a modified residue prefers to be accessible on the surface of a protein, the solvent-accessible surface area (ASA) around carboxylation sites is also investigated. Radial basis function network is then employed to build a predictive model using various features for identifying carboxylation sites. Based on a five-fold cross-validation evaluation, a predictive model trained using the combined features of amino acid sequence (AA20D), amino acid composition, and ASA, yields the highest accuracy at 0.874. Furthermore, an independent test done involving data not included in the cross-validation process indicates that *in silico* identification is a feasible means of preliminary analysis. Additionally, the predictive method presented in this work is implemented as Carboxylator (http://csb.cse.yzu.edu.tw/Carboxylator/), a web-based tool for identifying carboxylated proteins with modification sites in order to help users in investigating γ-glutamyl carboxylation.

Keywords Protein carboxylation · Amino acid composition · Solvent accessible surface area (ASA)

Availability: Carboxylator can be accessed via a web interface, and is freely available to all interested users at http://csb.cse.yzu.edu.tw/Carboxylator/. All of the data set that is used in this work is also available.

Electronic supplementary material The online version of this article (doi:10.1007/s10822-011-9477-2) contains supplementary material, which is available to authorized users.

C.-T. Lu · S.-A. Chen · N. A. Bretaña · T.-H. Cheng · T.-Y. Lee (✉)
Department of Computer Science and Engineering, Yuan Ze University, Room 3312, 135 Yuan-Tung Road, Chungli, Taoyuan 32003, Taiwan, ROC
e-mail: francis@saturn.yzu.edu.tw

C.-T. Lu
e-mail: s988306@mail.yzu.edu.tw

S.-A. Chen
e-mail: sachen.tw@gmail.com

N. A. Bretaña
e-mail: neil070490@yahoo.com

T.-H. Cheng
e-mail: czh0305@hotmail.com

Introduction

In proteins, glutamate (Glu) residues are transformed into γ-carboxyglutamate (Gla) residues in a process called carboxylation to aid in various cellular mechanisms. Carboxylation is a post-translational modification (PTM) of Glu residues in proteins wherein a carboxylic acid group is added into a substrate protein. The modification is catalyzed by a vitamin K-activated γ-glutamyl carboxylase which transforms a glutamate (Glu) residue to a γ-carboxyglutamate

(Gla) residue upon adding a carbon dioxide (CO_2) compound at the γ-position [1, 2]. Vitamin K-dependent gamma-glutamyl carboxylase plays a crucial role in the vitamin K cycle [3] and is associated the formation of calcium oxalate urolithiasis [4]. Carboxylated proteins can be activated when Gla domain binds Ca^{2+} [5, 6]. Since Glu is a weak Ca^{2+} chelator and Gla is a much stronger one, the vitamin K-dependent step greatly increases the Ca^{2+}-binding capacity of a protein [1]. Studies conducted over the last few years have revealed that the γ-glutamyl carboxylated proteins in vertebrates can be categorized into three main groups [7]. The first group comprises the carboxylated proteins with an amino terminal Gla domain, and includes vitamin K-dependent blood coagulation factors and co-regulators of blood coagulation [8]. The second group is composed of osteocalcin and matrix Gla protein (MGP) [6, 9], and includes three and five Gla residues, respectively, which are critical to the regulation of bone growth and extraosseous calcification [10, 11]. The third group is the γ-glutamyl carboxylase, itself, which includes Gla residues [12]. Carboxylation generally occurs in factors II, VII, IX, and X, protein C, protein S, as well in some bone proteins [9, 13].

The process of Carboxylation plays a significant role in a wide array of biological processes. It is primarily involved in the blood clotting cascade [14]. Moreover, it is also required for receptor-binding and initiating mitogenic activities in some proteins [8]. Owing to the importance of protein carboxylation in biological mechanisms, a great amount of effort is being put in order to continue identifying an increasing number of experimentally confirmed γ-glutamyl carboxylation sites. A previous work has utilized mass spectrometry to reveal that vitamin K-dependent carboxylation is a processive PTM in which multiple carboxylations occur during a single substrate binding event [14]. However, experimental identification of carboxylation sites via mass spectrometry-based methods is observed to be expensive, time-consuming, and labor-intensive. Potentially, *in silico* methods can be used to characterize carboxylated sites before experiments are conducted.

In this work, we present a novel method of identifying carboxylation sites. A side-chain of amino acid that undergoes PTM prefers to be accessible on the surface of a protein [15]. In addition to investigating the composition of amino acids that surround carboxylation sites, the structural characteristics such as solvent-accessible surface area (ASA) of carboxylated sites are also studied in detail. We then evaluate the capacity of various features in differentiating carboxylation sites from non-carboxylation sites by establishing predictive models using radial basis function network (RBFN). Lastly, we aim to further study carboxylation by investigating the linear distribution in carboxylation sites as well as the functional preference of carboxylated proteins. A web-based protein carboxylation sites prediction system utilizing the approach presented in this study is implemented for the scientific community (http://csb.cse.yzu.edu.tw/Carboxylator/).

Materials and methods

Figure 1 presents the analytical flowchart of this study which comprises of four major steps—data preparation, features investigation, model construction and assessment, and independent testing. Following the model construction and assessment, the selected models which contain the highest predictive accuracy are tested on an independent data set. The parameters and training features that provide the best predictive performance are used to implement the web-based system. Each process is described in detail as follows.

Data preparation

A total of 1,112 Gla residues from 177 protein entries from multiple organisms are available in dbPTM [16]—a comprehensive resource which contains PTM data from UniProtKB [17] and HPRD [18]. Non-experimental sites, annotated as "by similarity", "potential", and "probable" in the "MOD_RES" fields of UniProtKB are then removed yielding 454 experimentally-verified carboxylation sites from 132 carboxylated proteins. Basing on the observation that carboxylation sites mainly occur on Glu residues, this work regards all experimentally-verified carboxylated Glu residues as positive instances. On the other hand, Glu residues found in experimentally-verified carboxylated proteins but are not annotated as carboxylation sites are regarded as the negative instances. Consequently, a total of 844 non-carboxylated Glu sites are defined as the negative set. Next, amino acid composition (AAC) and ASA around the carboxylation sites are explored with reference to a previous work [19]. The said features are extracted and used in order to construct a predictive model. The resulting models are then evaluated in terms of its ability to differentiate carboxylation sites from non-carboxylation sites.

With reference to the reduction of the homology in the training set of N-Ace [20], two carboxylated protein sequences with more than 30% identity were defined as homologous sequences. Then, two homologous sequences are specified to re-align the fragment sequences using a window length of $2n + 1$, centered on the carboxylation sites using BL2SEQ [21]. For two fragment sequences having an identity higher than 50%, if the carboxylation sites from the two proteins are found in the same positions, only one site is kept while the other is discarded. The non-homologous negative data are generated using the same

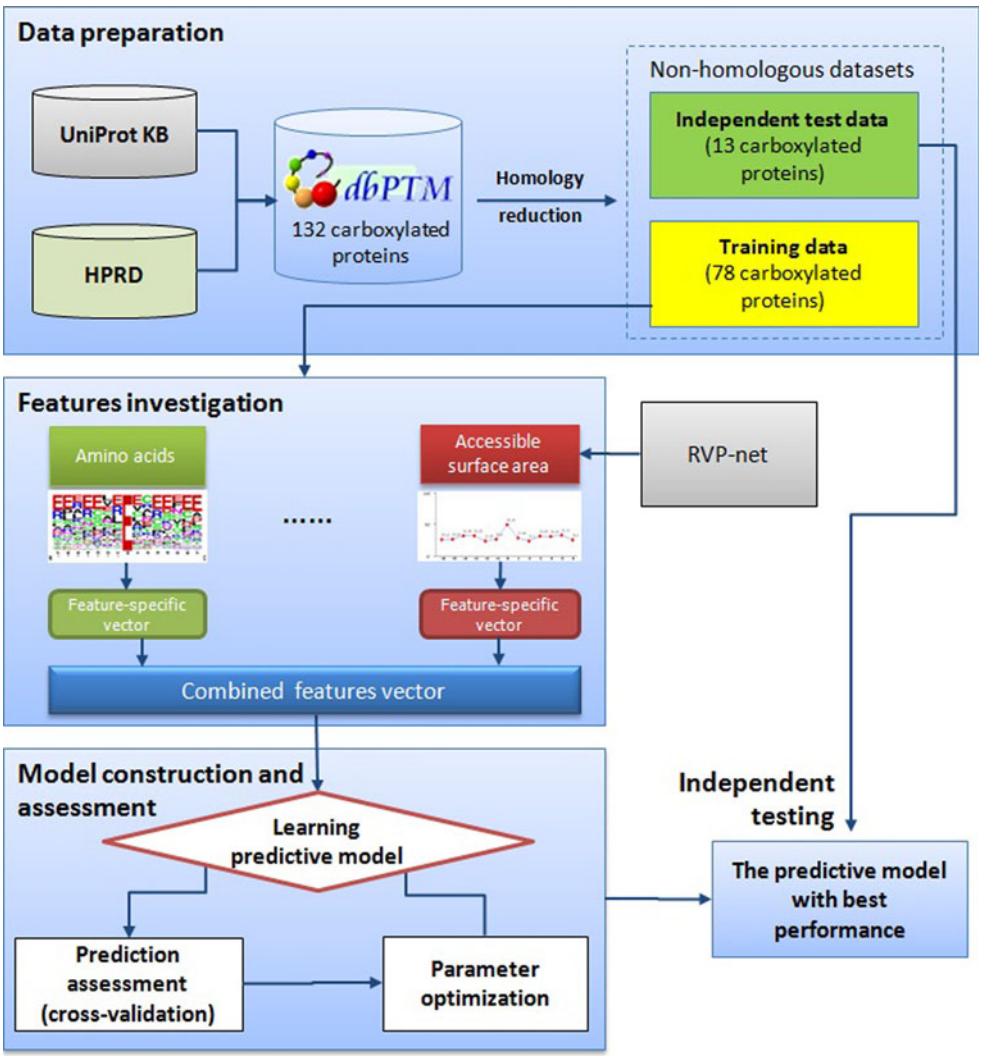

Fig. 1 Analytical flowchart. The four main steps in this study are data preparation, features investigation, model construction and assessment, and independent testing. In data preparation, the experimentally verified carboxylation sites are taken from UniProt and HPRD databases. The investigation of features explores the substrate site specificity of carboxylation sites based on sequence and structural characteristics. The explored features are then used in a predictive model to differentiate carboxylation sites from non-carboxylation sites. Following the model construction and prediction assessment, the model with the highest predictive accuracy is tested using an independent data set

approach as the positive. The removal of homologous sequences is done to in order to avoid an overestimation in the predictive performance.

Investigation of sequence feature

The composition of amino acids in carboxylation sites are investigated in this study. For the positive and negative sets, respectively, fragments of amino acids are extracted using a window size of $2n + 1$ centered on the Gla residue where different values of n varying from four to ten are used to determine the optimal window length. Next, an orthogonal binary coding scheme is adopted to transform amino acids into numeric vectors, in the so-called 20-dimensional vector coding (AA20D). For example, glycine (G) is encoded as "10000000000000000000;" alanine (A) is encoded as "01000000000000000000," and so on. The number of feature vectors that represent the flanking amino acids that surround the carboxylation site is $(2n + 1) \times 20$. With reference to a previous work [20], AAC is considered as the elementary feature in constructing the predictive model to determine the optimal window size.

Investigation of structural feature

In order to study the characteristics of carboxylation sites in a more in-depth manner, various structural features are investigated. It has been reported that amino acid side chains which undergo PTM tends to be accessible on the surface of a protein [15]. With this, the solvent-ASA preference surrounding carboxylation sites is measured. RVP-Net [22, 23] is used to compute the ASA values of a protein sequence due to the limitation that almost all experimental carboxylation sites do not have a corresponding protein tertiary structure in PDB [30]. A previous investigation of protein methylation [19] demonstrated that the RVP-Net-computed ASA value is very similar to the observed values in the protein tertiary structure. RVP-net applies neural network to predict the real value of residual

Table 1 Five-fold cross-validations of the 17-mer RBFN models trained with various features

Training features	Pre	Sn	Sp	Acc	MCC
Orthogonal binary coding of amino acid sequence (AA20D)	0.706	0.786	0.814	0.805	0.597
Amino acid composition (AAC)	0.686	0.790	0.802	0.792	0.587
Accessible surface area (ASA)	0.662	0.757	0.790	0.779	0.533
AA20D + AAC	0.736	0.815	0.842	0.833	0.641
AA20D + ASA	0.770	0.819	0.861	0.847	0.659
AAC + ASA	0.749	0.842	0.851	0.842	0.646
AA20D + AAC + ASA	**0.801**	**0.839**	**0.892**	**0.874**	**0.723**

The model containing best performance is marked with bold. *Pre* precision, *Sn* sensitivity, *Sp* specificity, *Acc* accuracy, *MCC* Matthews Correlation Coefficient

ASAs based on neighborhood information, with a mean absolute error of 18.0–19.5%, defined as the absolute difference between the predicted and experimental values of the relative ASA per residue [23]. The computed ASA value refers to the percentage of the solvent-accessible area of each amino acid on the protein sequence. The whole protein sequence containing an experimentally verified carboxylation site is entered into RVP-Net to compute for the ASA value of all residues.

Construction of predictive model

In this work, the QuickRBF package [24] has been employed to construct RBFN classifiers. The general architecture an RBFN consists of three layers, namely the input layer, the hidden layer, and the output layer. The input layer broadcasts the coordinates of the input vector to each of the nodes in the hidden layer. Each node in the hidden layer then produces an activation based on the associated radial basis kernel function. Finally, each node in the output layer computes a linear combination of the activations of the hidden nodes. The general mathematical form of the output nodes in RBFN is as follows: $c_j(x) = \sum_{i=1}^{k} w_{ji}\phi(\|x - \mu_i\|; \sigma_i)$; where $c_j(x)$ denotes the function corresponding to the j th output node and is a linear combination of k radial basis functions $\phi()$ with center μ_i and bandwidth σ_i; Also, w_{ji} denotes the weight associated with the correlation between the j th output node and the i th hidden node. In this work, we adopted a fixed bandwidth (σ) of 5, and used all input nodes as centers ($k = n$). With its several bioinformatics applications, classification based on RBFN has been extensively adopted to predict PTMs such as glycosylation sites [25] and ubiquitylation sites [26].

Assessment of predictive performance

Prior to the construction of a final model, the predictive performance of the models with varying parameters are evaluated by performing k-fold cross validation. It is reported that cross-validation evaluation is important for the application of a predictor [27]. In doing k-fold cross validation, rhe training data is divided into k groups by splitting each dataset into k approximately equal sized subgroups. In one round of cross-validation, a subgroup is regarded as the test set, and the remaining $k - 1$ subgroups are regarded as the training set. The cross-validation process is repeated k rounds, with each of the k subgroups used as the test set in turn. Then, the k results are combined to produce a single estimation. The advantage of k-fold cross-validation is that all original data are regarded as both training set and test set, and each data is used for testing exactly once. In this study, k is set to five. The impact of using the following features: amino acid sequence, AAC, and ASA, is evaluated by five-fold cross-validation to determine which features are best utilized to establish models that can effectively differentiate between carboxylation sites and non-carboxylation sites. The following measures of predictive performance of the trained models are defined:

$$\text{Precision (Pre)} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{1}$$

$$\text{Sensitivity (Sn)} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

$$\text{Specificity (Sp)} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{3}$$

$$\text{Accuracy (Acc)} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{TN} + \text{FP}} \tag{4}$$

$$\text{Matthews Correlation Coefficient (MCC)} = \frac{(\text{TP} \times \text{TN}) - (\text{FN} \times \text{FP})}{\sqrt{(\text{TP} + \text{FN}) \times (\text{TN} + \text{FP}) \times (\text{TP} + \text{FP}) \times (\text{TN} + \text{FN})}} \tag{5}$$

where TP, TN, FP and FN represent the numbers of true positives, true negatives, false positives and false

negatives, respectively. Additionally, the parameters of the predictive models, window length, cost, and gamma value of the RBFN models are optimized to maximize predictive accuracy. The optimized parameters which yield the highest accuracy are then used to construct predictive models for independent testing.

Independent testing

Subsequent to the construction of the predictive model, an independent test is carried out to ensure that the model if not over-fit to the training set. We randomly select around 15% of the experimental carboxylated proteins as the independent test set. The non-homologous training data comprises of 292 carboxylated glutamate residues (positive set of training data) and 566 non-carboxylated glutamate residues (negative set of training data) from 78 carboxylated proteins. Additionally, a total of 50 carboxylated glutamate residues and 93 non-carboxylated glutamate residues from 13 carboxylated proteins, none of which are included in the training data set, are regarded as the positive set and negative set for independent testing respectively. After the evaluation using *k*-fold cross-validation, the trained model with the highest accuracy is evaluated using the independent test data.

Results and discussion

Distribution of carboxylation sites and other PTMs on carboxylated proteins

Morris et al. [14] utilized mass spectrometry to show that multiple carboxylation processes occur during a single substrate binding event. The growing interest in mass spectrometric proteomic studies of γ-glutamyl carboxylation demands a detailed characterization of protein carboxylation sites. To examine the linear distribution of carboxylation sites on proteins, the distance between two neighboring carboxylation sites is measured. Figure 2 reveals that two immediately adjacent carboxylation sites are separated by one amino acid or by three amino acids. Overall, almost all (97.5%) carboxylation sites are located within a distance of seven amino acids, and so are very close together. Previous works have suggested that efficient carboxylation of native substrates requires the binding of a conserved region to carboxylase [2, 28]. Therefore, further analysis of PTMs on carboxylated proteins—especially those in the Gla domain—is needed. All annotations of PTMs in carboxylated proteins are taken from UniProtKB [17]. Table S1 (Supplementary Materials) reveals that hydroxylation and glycosylation are the most abundant PTM types on carboxylated proteins. Twenty-seven and 24 carboxylated proteins were found to have 41 hydroxylation sites and 84 glycosylation sites, respectively. As shown in Fig. S1 (Supplementary Materials), most of the 41 hydroxylation sites are located close to carboxylation sites whereas most of the 84 glycosylation sites are far from carboxylation sites. The 27 proteins (about 30% of all carboxylated proteins) that contain co-occurring hydroxylation and carboxylation sites are aligned using a multiple sequence alignment tool ClustalW [29]. Figure S2 (Supplementary Materials) shows that these 27 proteins are clustered into three groups of homologous proteins, which are secreted proteins [30], calcium binding proteins [31], and osteocalcin [32]. The occurrence of other PTMs may be considered to infer the function of carboxylated proteins.

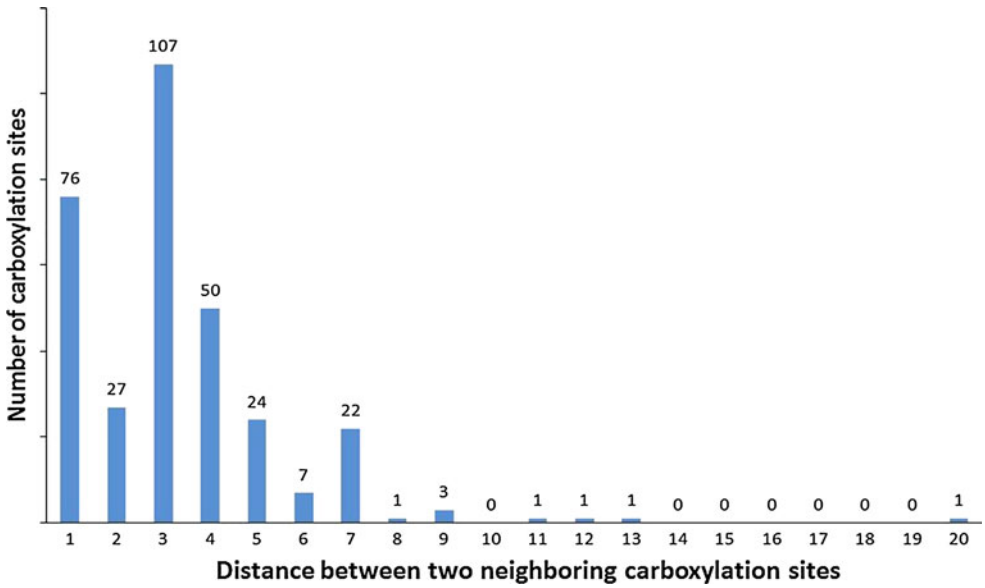

Fig. 2 Statistical analysis of distance between neighboring carboxylation sites. Almost all carboxylation sites (97.5%) were separated by seven or fewer amino acids

Amino acid composition and accessible surface area at carboxylation sites

The AAC of 21-mer carboxylated fragments is graphically visualized using WebLogo [33, 34] to reveal the relative frequency of the corresponding amino acid at each position around the carboxylation sites. Based on the frequency plot as shown in Fig. 3a, carboxylation sites are observed to contain highly concentrated Glu residues around the carboxylation sites. It is observed that the high conservation of negatively charged glutamate residue coincides with previous findings [5] that the γ-carboxylation recognition site suffices to direct vitamin K-dependent carboxylation on an adjacent glutamate-rich region of thrombin in a propeptide-thrombin chimera. Aside from the AAC feature, the solvent-ASA of amino acids was considered to explore the structural characteristics of carboxylation sites. Figure 3b compares the mean percentage of ASA obtained using a 21-mer window (from -10 to $+10$) between carboxylation sites and non-carboxylation sites. The analysis reveals that the flanking region of carboxylation sites has a high preference for the solvent-ASA, especially at carboxylation sites. The mean percentage of ASA on carboxylated glutamate residues is 37.6%, resulting in a great exposure to the solvent. In the investigation of ASA curves, the notable difference between carboxylation sites and non-carboxylation sites is found in the region from -7 to -4.

Optimizing window length based on amino acid sequence

The optimal window length that best predicts carboxylation sites on Glu residues is determined by doing a five-fold cross-validation on models that are trained using the AA20D features of carboxylation sites with various window sizes $2n + 1$, where n varies from four to ten. As the window size varies from 9 to 21, it is observed that the predictive accuracy improves slightly from 0.728 to 0.808. As the window size increases, the predictive specificity improves while the sensitivity declines. Overall, the models trained using a window size of 15 and 17 performs best. Given the computational efficiency and overall performance of the trained models, 17-mer is selected as the length of the window in the following implementation. Based on the training feature of amino acid sequence, the precision, sensitivity, specificity, accuracy, and MCC resulting from a model with a 17-mer window size are 0.706, 0.786, 0.814, 0.805, and 0.597, respectively.

Predictive performance of models using various features

The investigated features are evaluated to create a predictive model for identifying carboxylation sites. From the training data set, amino acid sequences are encoded into

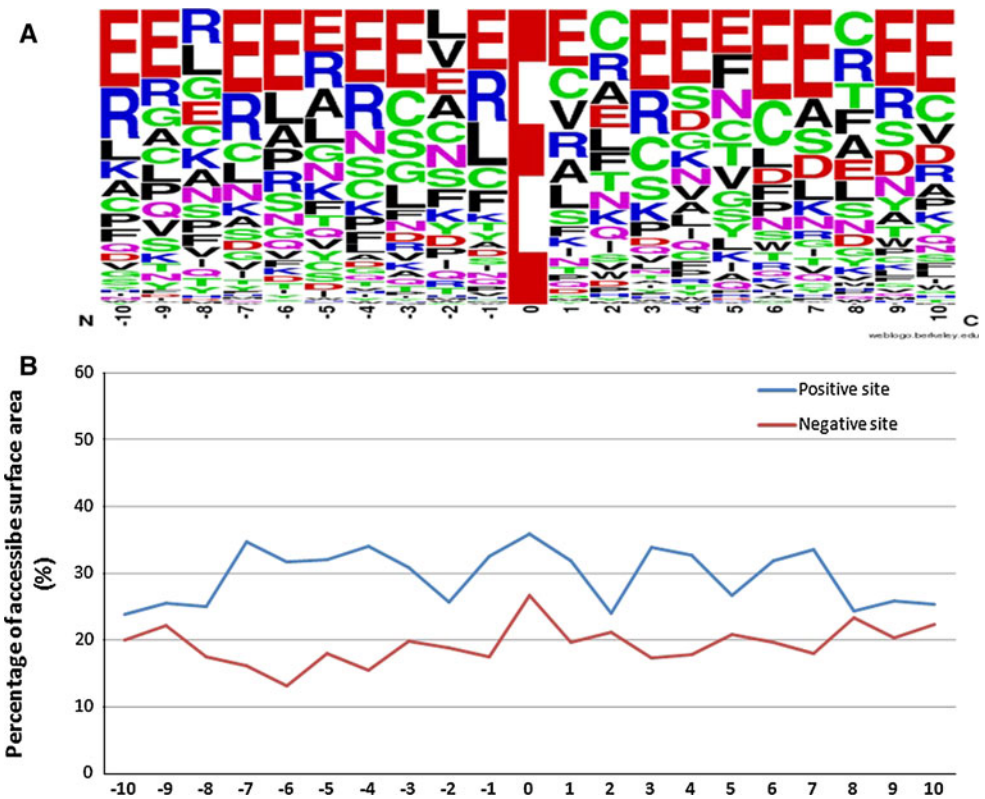

Fig. 3 Sequence and structural features around carboxylation sites with 21-mer window length (from -10 to $+10$). **a** Frequency plot of amino acid composition around carboxylation sites. **b** Average ASA percentage at carboxylation sites (*blue line*) and non-carboxylation sites (*red line*)

20-dimensional vector and AAC, denoted "AA20D" and "AAC", respectively. Moreover, the ASA using the RVP-net ASA values. Table 1 shows that the precision, sensitivity, specificity, accuracy, and MCC of the model trained with AAC could reach 0.686, 0.790, 0.802, 0.792, and 0.587, respectively. The model trained with ASA could reach an accuracy of 0.779. It is observed that the model trained using a orthogonal binary coding of amino acids (AA20D) slightly outperforms those trained with AAC alone or ASA alone.

With interest to a possible improvement in predictive performance by training the models using a hybrid combination of features, a model is trained and evaluated using a hybrid combination of AA20D, AAC, and ASA. In comparison to the performance achieved using individual features, AA20D is crucial for training a model with other individual features. The model trained using a combination of AA20D and ASA substantially outperforms those trained with single feature but slightly outperforms those that are trained with a hybrid combination of two features. The model trained using a combination of AA20D, AAC and ASA has the best overall accuracy. The predictive precision, sensitivity, specificity, accuracy, and MCC of the best model are 0.801, 0.839, 0.892, 0.874, and 0.723, respectively. It would be noticed that the model trained with the feature of ASA could improve the prediction performance. In conclusion, five-fold cross-validation indicates that the model that is trained using a combination of AA20D, AAC and ASA performs best, and is therefore adopted in further independent testing.

Evaluation of the predictive model using an independent test set

The effectiveness of the studied features that yield the highest accuracy in cross-validation is evaluated using an independent test. Based on the performance evaluation using five-fold cross-validation, the model trained using a 17-mer window length and the combined features of amino acid sequence (AA20D), AAC, and ASA are selected. Table 2 shows that the predictive sensitivity falls slightly during independent testing and specificity falls by around 10%. Overall, independent testing reveals that the model has an accuracy of 0.825, which approximates to that of cross-validation. The precision, sensitivity, specificity, and MCC in independent testing are 0.705, 0.860, 0.806, and 0.642, respectively. Accordingly, independent testing demonstrates that the amino acid sequence (AA20D), AAC, and ASA can distinguish between carboxylation and non-carboxylation sites when data are truly blind to the cross-validation process.

Implementation of a web-based system for identifying carboxylation sites

Since experimental identification is time-consuming and labor-intensive, precisely identifying the sites of

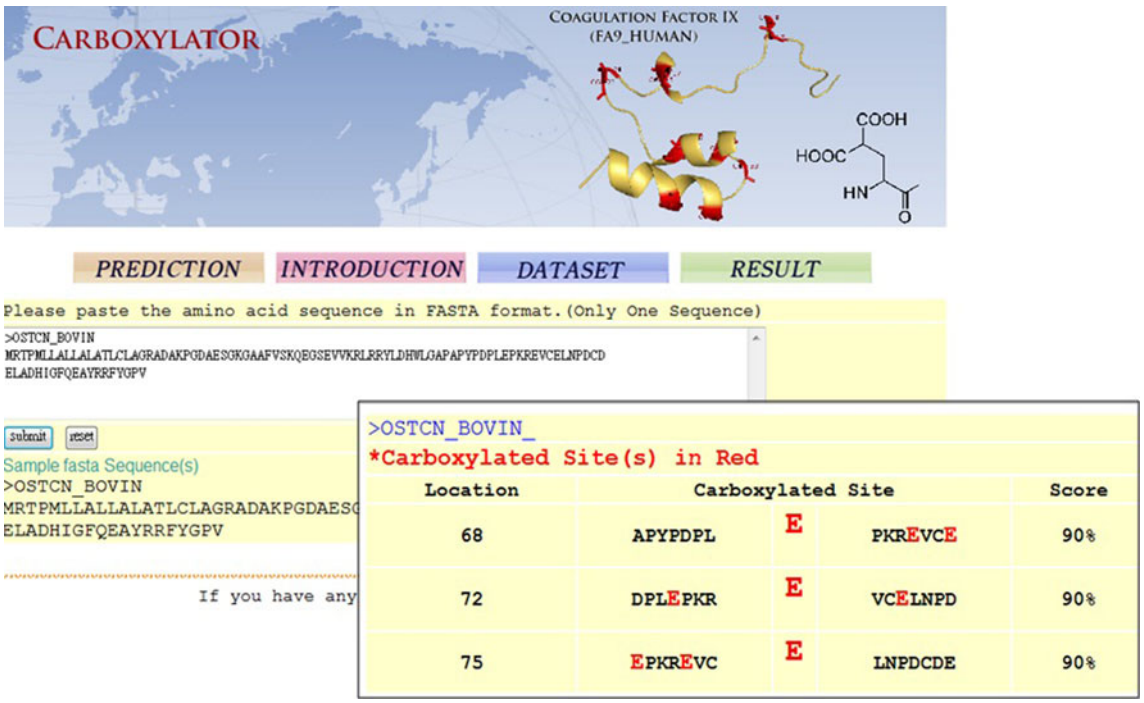

Fig. 4 Web interface (Carboxylator) for proposed prediction scheme

Table 2 Comparison of predictive performances between cross-validation and independent testing

	Number of data		Pre	Sn	Sp	Acc	MCC
	Positive set	Negative set					
Cross-validation	292	566	0.801	0.839	0.892	0.874	0.723
Independent testing	50	93	0.705	0.860	0.806	0.825	0.642

γ-glutamyl carboxylation on the substrate is difficult even in a carboxylated protein. Hence, an effective prediction tool is required to identify potential carboxylation sites efficiently. Following the cross-validation evaluation and independent testing, the amino acid sequence (AA20D), amino acids composition (AAC), and ASA are utilized to construct an RBFN model to predict the glutamate residues that are involved in carboxylation. As presented in Fig. 4, users can submit their uncharacterized protein sequences, then the prediction system, called Carboxylator, efficiently returns the predictions, including the carboxylated position, flanking amino acids, and the probability of carboxylation. With regard to the functional analysis of carboxylated proteins, the implementation of the proposed prediction scheme effectively helps users to elucidate the biological function of an uncharacterized protein.

Conclusions

Although the high-throughput mass spectrometry has been wildly used in proteomics, studies on substrate site specificity of γ-glutamyl carboxylation are subject to technical limitations. With the collection of experimentally verified carboxylation sites from UniProtKB and HPRD, 454 experimentally verified carboxylation sites have been identified in 132 carboxylated proteins. After the construction of a training data and an independent test data, the composition of the flanking amino acids among the training data is studied. This study also investigated the structural characteristics of carboxylated proteins such as solvent-ASA. Based on ASA curves, the region composed of amino acids from position -7 to -4 exhibits notable differences between carboxylation sites and non-carboxylation sites. The mean percentage of ASA values on the carboxylated glutamate residues is 37.6% which shows that it is greatly exposed to the solvent. A five-fold cross-validation evaluation demonstrates that incorporating the structural feature of ASA could improve the prediction of protein carboxylation sites. Furthermore, an independent test concurs that the proposed model can differentiate carboxylation sites from non-carboxylation sites. To enable efficient analysis of γ-glutamyl carboxylation, the predictive approach was established as Carboxylator, a web-based tool for identifying carboxylated proteins with modification sites.

Although the independent tests verify that the proposed method performs accurately and robustly, some issues warrant further investigation. The structural affinities of carboxylation sites should be studied in detail, particularly when the flanking amino acids are not conserved. The solvent-ASA and secondary structure, the B-factor, intrinsic disordered region, protein linker region, and other factors should also be examined at experimental carboxylation sites in the protein regions with PDB entries. More importantly, it should be noted that clues may be found from analysis of gene ontology [35], the occurrence of other PTMs, and the network context of protein–protein interactions regarding further functions of carboxylated proteins. These may be used to further investigate the biological functions of carboxylated proteins.

Acknowledgments The authors sincerely appreciate the National Science Council of the Republic of China for financially supporting this research under Contract Numbers of NSC 100-2221-E-155-079.

References

1. Vermeer C (1990) Biochem J 266:625
2. Knobloch JE, Suttie JW (1987) J Biol Chem 262:15334
3. King CR, Deych E, Milligan P, Eby C, Lenzini P, Grice G, Porche-Sorbet RM, Ridker PM, Gage BF (2010) Thromb Haemost 104:750
4. Wang T, Yang J, Qiao J, Liu J, Guo X, Ye Z (2010) Urol Int 85:94
5. Furie BC, Ratcliffe JV, Tward J, Jorgensen MJ, Blaszkowsky LS, DiMichele D, Furie B (1997) J Biol Chem 272:28258
6. Price PA, Urist MR, Otawara Y (1983) Biochem Biophys Res Commun 117:765
7. Bandyopadhyay PK, Garrett JE, Shetty RP, Keate T, Walker CS, Olivera BM (2002) Proc Natl Acad Sci USA 99:1264
8. Kulman JD, Harris JE, Xie L, Davie EW (2001) Proc Natl Acad Sci USA 98:1370
9. Price PA, Poser JW, Raman N (1976) Proc Natl Acad Sci USA 73:3374
10. Luo G, Ducy P, McKee MD, Pinero GJ, Loyer E, Behringer RR, Karsenty G (1997) Nature 386:78
11. Ducy P, Desbois C, Boyce B, Pinero G, Story B, Dunstan C, Smith E, Bonadio J, Goldstein S, Gundberg C, Bradley A, Karsenty G (1996) Nature 382:448
12. Berkner KL, Pudota BN (1998) Proc Natl Acad Sci USA 95:466
13. Olson RE, Suttie JW (1977) Vitam Horm 35:59
14. Morris DP, Stevens RD, Wright DJ, Stafford DW (1995) J Biol Chem 270:30491
15. Pang CN, Hayen A, Wilkins MR (2007) J Proteome Res 6:1833

16. Lee TY, Huang HD, Hung JH, Huang HY, Yang YS, Wang TH (2006) Nucleic Acids Res 34:D622
17. Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LS (2004) Nucleic Acids Res 32:D115
18. Peri S, Navarro JD, Kristiansen TZ, Amanchy R, Surendranath V, Muthusamy B, Gandhi TK, Chandrika KN, Deshpande N, Suresh S, Rashmi BP, Shanker K, Padma N, Niranjan V, Harsha HC, Talreja N, Vrushabendra BM, Ramya MA, Yatish AJ, Joy M, Shivashankar HN, Kavitha MP, Menezes M, Choudhury DR, Ghosh N, Saravana R, Chandran S, Mohan S, Jonnalagadda CK, Prasad CK, Kumar-Sinha C, Deshpande KS, Pandey A (2004) Nucleic Acids Res 32:D497
19. Shien DM, Lee TY, Chang WC, Hsu JB, Horng JT, Hsu PC, Wang TY, Huang HD (2009) J Comput Chem 30:1532
20. Lee TY, Hsu JB, Lin FM, Chang WC, Hsu PC, Huang HD (2010) J Comput Chem 31:2759
21. Tatusova TA, Madden TL (1999) FEMS Microbiol Lett 174:247
22. Ahmad S, Gromiha MM, Sarai A (2003) Bioinformatics 19:1849
23. Ahmad S, Gromiha MM, Sarai A (2003) Proteins 50:629
24. Yang ZR, Thomson R (2005) IEEE Trans Neural Netw 16:263
25. Chen SA, Lee TY, Ou YY (2010) BMC Bioinformatics 11:536
26. Lee TY, Chen SA, Hung HY, Ou YY (2011) PLoS One 6:e17331
27. Chou KC, Shen HB (2007) Anal Biochem 370:1
28. Pan LC, Price PA (1985) Proc Natl Acad Sci USA 82:6109
29. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG (2007) Bioinformatics 23:2947
30. Kauferstein S, Kendel Y, Nicke A, Coronas FI, Possani LD, Favreau P, Krizaj I, Wunder C, Kauert G, Mebs D (2009) Toxicon 54:295
31. Virdi AS, Willis AC, Hauschka PV, Triffitt JT (1991) Biochem Soc Trans 19:373S
32. Nielsen-Marsh CM, Richards MP, Hauschka PV, Thomas-Oates JE, Trinkaus E, Pettitt PB, Karavanic I, Poinar H, Collins MJ (2005) Proc Natl Acad Sci USA 102:4409
33. Crooks GE, Hon G, Chandonia JM, Brenner SE (2004) Genome Res 14:1188
34. Schneider TD, Stephens RM (1990) Nucleic Acids Res 18:6097
35. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Nat Genet 25:25