scientific reports

Check for updates

OPEN

Identification of 4-carboxyglutamate residue sites based on position based statistical feature and multiple classification

Asghar Ali Shah[1]✉ & Yaser Daanial Khan[2]

Glutamic acid is an alpha-amino acid used by all living beings in protein biosynthesis. One of the important glutamic acid modifications is post-translationally modified 4-carboxyglutamate. It has a significant role in blood coagulation. 4-carboxyglumates are required for the binding of calcium ions. On the contrary, this modification can also cause different diseases such as bone resorption, osteoporosis, papilloma, and plaque atherosclerosis. Considering its importance, it is necessary to predict the occurrence of glutamic acid carboxylation in amino acid stretches. As there is no computational based prediction model available to identify 4-carboxyglutamate modification, this study is, therefore, designed to predict 4-carboxyglutamate sites with a less computational cost. A machine learning model is devised with a Multilayered Perceptron (MLP) classifier using Chou's 5-step rule. It may help in learning statistical moments and based on this learning, the prediction is to be made accurately either it is 4-carboxyglutamate residue site or detected residue site having no 4-carboxyglutamate. Prediction accuracy of the proposed model is 94% using an independent set test, while obtained prediction accuracy is 99% by self-consistency tests.

Proteins are a key element of every cell necessary to build and repair tissues. They are macromolecules constructed using a chain of amino acid residues. Proteins exhibit numerous properties, they may work as hormones, enzymes or may be a part of structural cellular component. Among 20 common proteins, glutamic acid is an important protein with a wide range of functions. Specifically, it has role in proper functioning of central and the peripheral nervous system[1].

Vitamin K-dependent carboxylase is a bifunctional enzyme. It catalyzes the oxygenation of vitamin K hydroquinone, helps in formation of vitamin K epoxide, resulting the formation of carboxyglutamate. 4-carboxyglutamate is a modification of glutamic acid formed due to post-translational modification (PTM). The structure of glutamic acid and 4-carboxyglutamate is explained in Figs. 1 and 2. These modified residues are then further exploited to bind calcium ions. These calcium ions provide positive charges to glutamic acids which further interact with the negatively charged phospholipid membrane[2,3]. Carboxylation has role in blood clotting and other biological processes[4,5]. The deficiency of vitamin K also results in deficiency of protein S and C which also formulate a Moyamoya disease. Carboxylation of glutamic acid causes disorders including bone resorption, osteoporosis, papilloma and plaque atherosclerotic[6–8].

Experimenting and identification of 4-carboxyglutamate residue sites at laboratory is costly and time-consuming. Therefore, it is necessary to formulate a computational model to identify 4-carboxyglutamate residue sites.

This study focuses on the post translational modification of glutamic acid into 4-carboxyglutamic acid within the glutamic acid domain modification. An accurate and efficient prediction model is devised to serve the purpose. The methodology is based on a Chou's 5-step rule[11]. These rules serve as a benchmark for dataset collection, mathematical formulation of samples, prediction-algorithm, and cross-validation of results and the development of web server. This methodology is further carried out one by one in the above said sequential order.

[1]Department of Computer Sciences, Bahria University Lahore Campus, Lahore 25000, Pakistan. [2]University of Management and Technology, Lahore 25000, Pakistan. ✉email: alishahsadiq@gmail.com

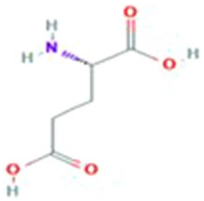

Figure 1. Structure of glutamic acid[9].

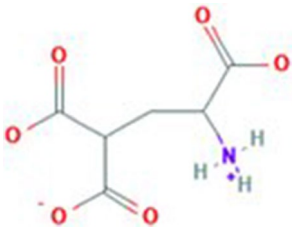

Figure 2. Structure of 4-carboxyglutamate[10].

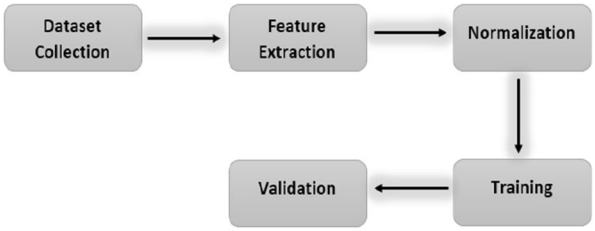

Figure 3. Flow chart of methodology.

Materials and methods

Chou's peptide formulation[11,12] is widely used in many studies[13–19]. In this study, Chou's formulation is also adopted to reach the solution. The operational flow chart of the chosen methodology is depicted in Fig. 3.

Benchmark dataset. 4-Carboxyglutamate sequences are extracted from a universal resource of protein (www.UniProt.org) through an advanced search query. The data is bifurcated as one with 4-carboxyglutamate modification and the other without 4-carboxyglutamate residues (also termed as positive and negative respectively). The redundancy and homology biases were excluded through CD-HIT web server (https://weizhongli-lab.org/cd-hit/) and the similarity threshold is 90%. Finally, a refined benchmark dataset of 261 proteins are constructed containing 560 positive and 600 negative samples. The total observations of obtained dataset are 560 + 600 = 1160. The dataset is represented by O. The positive observations are represented by O+, and negative observations within the data set are depicted by O−. U represents union according to the set theory.

$$\mathrm{O} = \mathrm{O}^{+} \mathrm{U}\, \mathrm{O}^{-} \tag{1}$$

Sequence logo. The PTM sequencing of the obtained dataset is graphically and visually represented in Figs. 4 and 5. Sequence conservation at a specific position is represented by the overall height of the stack.

Sample formulation. The formulation of biological sequencing is one of the most critical problems in computational biology. Vector quantification is a key to formulate the sequence by maintaining their sequence patterns and features that are required for targeted analysis. As vector quantification paves a way for addressing the formulated sequencing using machine learning algorithms[20]. In this work, a pseudo amino acid composition (PseAAC)[21] is chosen. According to the chosen composition, samples in the dataset can be described as[34]. Equation (2) depicts that each sample is a subsequence of fixed size while Eq. (3) depicts that 20 residues upstream and 20 residues downstream were extracted while R21 is the 4-carboxyglutamate site.

$$\boldsymbol{B}_{\xi=7}(\mathbb{K}) = [\Psi_1 \Psi_2 \ldots \Psi_u \ldots \Psi_\Omega]^T \tag{2}$$

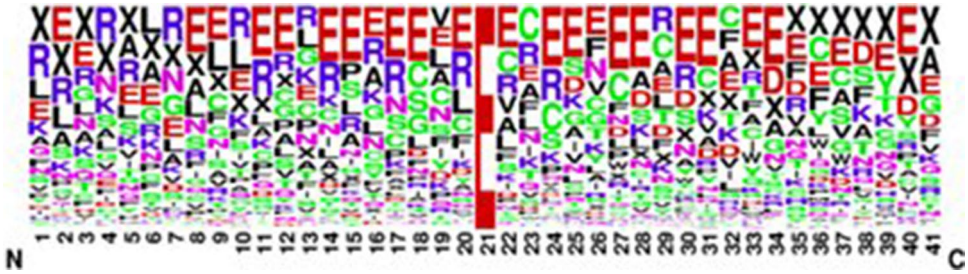

Figure 4. Sequence logo of positive 4-carboxyglutamate.

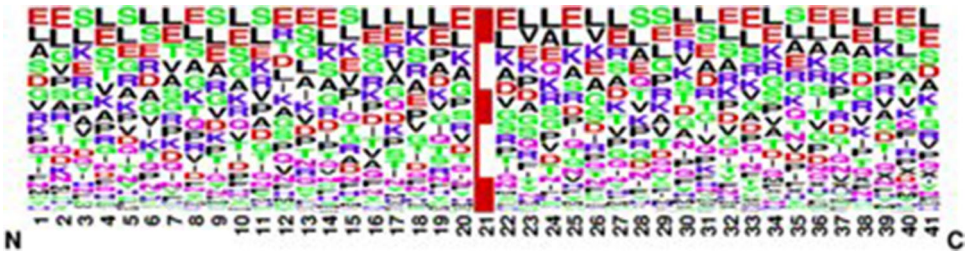

Figure 5. Sequence logo of negative 4-carboxyglutamate.

where u = 1, 2, 3 …Ω. It elaborates how useful features can be extracted from relevant peptide sequencing and T denotes transpose operator. Each sample peptide sequence is 41 in length due to which Eq. (2) can be formulated as.

$$\boldsymbol{B} = R_1 R_2 \ldots R_{19} R_{20} R_{21} \ldots R_{40} R_{41} \tag{3}$$

Statistical moment calculation. The composition of each sequence of proteins follows some specific pattern. Due to such distinction, each sequence is to be described with different statistical parameters. In previous work, statistical moments are used for feature extraction[22,23]. In order to have feature extraction, raw, central and Hahn moments are used. The composition of amino acids has a very important role in the functionality and nature of the proteins. The extraction of the feature can be location and scale variant. To address location variant features, raw moments are used to calculate mean, variance and asymmetry of sample distribution in the dataset. Central moments are also used for feature extraction by estimating mean, variance and asymmetry but it is location invariant as the estimations are made using centroid but central moments are actually scaled variant[24,25]. Hahn moments are used to estimate statistical parameters but these moments are both location and scale variant[26,27]. Therefore Hahn moments are computed using Hahn polynomials to estimate the mean in dataset and variance in dataset and asymmetry of the probability distribution. For the said method, moments are computed in a two-dimensional n × n matrix denoted by B′[28].

$$\mathbf{B}' = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1n} \\ b_{21} & b_{22} & \ldots & b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n1} & b_{n2} & \cdots & b_{nn} \end{bmatrix} \tag{4}$$

A function ω[29] is a mapping function used for matrix transformation of **B** as **B′**. It uses the element from this matrix **B′**. Moments were computed up to order three such as M01, M10, M11, M12, M21, M30 and M03. The raw moments are computed as given below.

$$M_{ij} = \sum_{b=1}^{n} \sum_{q=1}^{n} b^i q^j \beta_{bq} \tag{5}$$

The sum of i and j represents the order of the moments that is i + j and it can be less than or equal to three. The Central moments can be computed as given below.

$$n_{ij} = \sum_{b=1}^{n} \sum_{q=1}^{n} (b - \overline{x})^i (q - \overline{y})^j \beta_{bq} \tag{6}$$

Hahn moments can be easily computed for even dimensional data organization. Reversible property of Hahn moments is evident due to their orthogonality[28]. Hahn moments of order n are computed as following,

$$h_n^{u,v}(r,N) = (N+V-1)_n(N-1)_n \times \sum_{k=0}^{n} (-1)^k \frac{(-n)_k(-r)_k(2N+u+v-n-1)_k}{(N+v-1)_k(N-1)_k} \frac{1}{k!}. \tag{7}$$

Normalized orthogonal Hahn moments of two dimensional discrete are computed as

$$H_{ij} = \sum_{q=0}^{N-1}\sum_{b=0}^{N-1} \beta_{ij} \widetilde{h_i^{u,v}}(q,N) \widetilde{h_i^{u,v}}(b,N) m, n = 0, 1, \ldots N-1. \tag{8}$$

Determination of PRIM and RPRIM. The primary sequence and relative position of residues are key factors to predict the characteristics of proteins. Quantitative characterization of the relative position of amino acid is also necessary. In order to serve the said purpose, 20 × 20 matrix is constructed as representative of Position relative Incidence Matrix (PRIM) to extract information about the relative position of each amino acid residue in the protein as given in Eq. (9).

$$S_{PRIM} = \begin{bmatrix} S_{1\to1} & S_{1\to2} & \ldots. & S_{1\to j} & \ldots. & S_{1\to1} \\ S_{2\to1} & S_{2\to1} & \ldots. & S_{2\to1} & \ldots. & S_{2\to20} \\ S_{i\to1} & S_{i\to1} & \ldots. & S_{i\to1} & \ldots. & S_{i\to20} \\ S_{N\to1} & S_{N\to1} & \ldots. & S_{N\to1} & \ldots. & S_{N\to20} \end{bmatrix} \tag{9}$$

Information is extracted as 400 coefficients for PRIM. In order to reduce PRIM dimensionality, statistical moments are computed for PRIM which produces a set of 24 elements.

To make it more effective and better, identifying hidden features, Reverse Position Relative Incidence Matrix (RPRIM) is also computed as:

$$S_{RPRIM} = \begin{bmatrix} S_{1\to1} & S_{1\to2} & \ldots & S_{1\to j} & \cdots & S_{1\to1} \\ S_{2\to1} & S_{2\to1} & \ldots & S_{2\to1} & \ldots & S_{2\to20} \\ S_{i\to1} & S_{i\to1} & \ldots & S_{i\to1} & \ldots & S_{i\to20} \\ S_{N\to1} & S_{N\to1} & \ldots & S_{N\to1} & \cdots & S_{N\to20} \end{bmatrix}. \tag{10}$$

By adapting the procedure explained in PRIM, 400 coefficients are also obtained from RPRIM. Similarly, with the help of computing statistical parameters, a set of 24 elements is obtained by reducing the dimensionality of RPRIM.

Feature scaling. Feature scaling is actually used to provide all features an opportunity to give an equal contribution to detect and predict the 4-carboxyglutamate sequencing. In this work, a standard scaler function is used within the Python environment to scale all features[30]. The standard scaler is used to scale the given data such that each feature should have mean around zero and unit variance. The standard scaling formulation is given in Eq. (11).

$$\text{Min}-\text{Max scaling} : X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{11}$$

Prediction algorithm. In this work, Multilayered Perceptron (MLP), Logistic Regression and Random Forest classifiers are applied for the prediction of 4-carboxyglutamate residue sites. MLP classifier provides better prediction which is 94% in comparison to other methods. So MLP is discussed further in detail.

The dataset has consisted of a total of 1160 sequences including 560 positive samples and 600 negative samples including 194 features. A supervised learning approach is used in this work to predict 4-carboxyglutamate residue sites. The prediction algorithm has to predict between residue sites having 4-carboxyglutamate or not.

MLP is a feed-forward artificial neural network that is used to map input data against the most appropriate output. It is actually a directed graph consisting input and an output layer and multiple hidden layers in between them. All nodes are connected to all other nodes in the adjacent layer and therefore, it is called a fully connected network[31]. The graphical representation of the MLP classifier is given in Fig. 6.

MLP classifier consists of N neurons in the hidden layer and each neuron has R weights, which is described in the N × R matrix[33]. The input weight matrix has N elements and is denoted by I as described in Eq. (12). The functional processing of the hidden layer is explained with the help of Eqs. (12) – (14).

$$\mathbf{I} = \begin{bmatrix} w_{1,1} & w_{1,2} & \cdots & w_{1,R} \\ w_{2,1} & w_{2,2} & \ldots & w_{2,R} \\ \vdots & \vdots & \ddots & \vdots \\ w_{N,1} & w_{N,2} & \cdots & w_{N,R} \end{bmatrix} \tag{12}$$

$$\boldsymbol{n}_1 = \mathrm{I} \cdot \mathrm{V} + b_1 \tag{13}$$

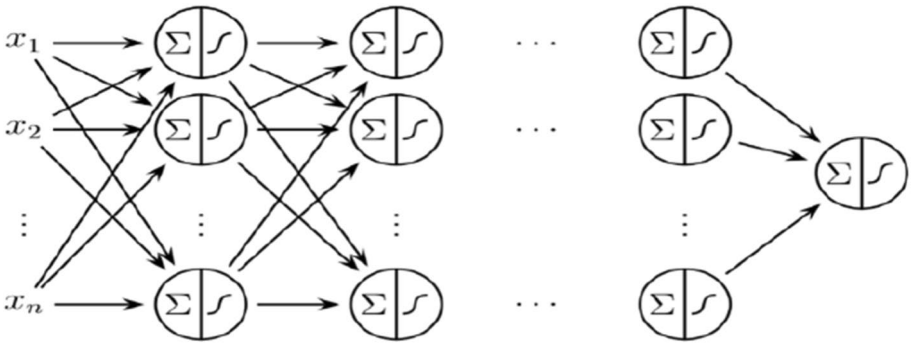

Figure 6. Graphical representation of MLP classifier[32].

n = 232	Predicted No	Predicted Yes	
Actual			
No	T N = 106	F P = 8	114
Actual			
Yes	F N = 6	T P = 112	118
	112	120	

Table 1. Confusion matrix of the proposed model.

$$\boldsymbol{a_1} = f_1(\boldsymbol{n_1}) \tag{14}$$

The sequential processing of output layer form hidden layer is explained with the help of Eqs. (15)–(17).

$$\mathbf{L} = \begin{bmatrix} w_{1,1} & w_{1,2} & \cdots & w_{1,S} \\ w_{2,1} & w_{2,2} & \ldots & w_{2,S} \\ \vdots & \vdots & \ddots & \vdots \\ w_{k,1} & w_{k,2} & \cdots & w_{k,S} \end{bmatrix} \tag{15}$$

$$\boldsymbol{n}_2 = \mathrm{I} \cdot \boldsymbol{a}_1 + b_2 \tag{16}$$

$$\boldsymbol{a_2} = f_2(\boldsymbol{n_2}) \tag{17}$$

Results

This study is first to predict 4-carboxyglutamate residue sites. Data samples are collected and formulated as described in "Materials and methods" section. The obtained data sets had non-numeric values having a series of alphabetic values. A featured set of numeric values is obtained as explained in "Sequence logo" section. As there were a lot of variations in obtained data so feature scaling technique is used so that each feature should have equal contribution in the prediction and detection of 4-carboxyglutamate residue sites. A neural network named MLP Classifier is used to train the obtained data sets and then based on training 4-carboxyglutamate residue sites are then predicted efficiently. The process of MLP classifier is well explained using graphical representation as shown in Fig. 6 and mathematically described in Eqs. (12) – (17) respectively.

The confusion matrix obtained from the MLP classifier is described in detail in Table 1. True positive, true negative, false positive, false negative is represented as TP, TN, FP and FN respectively.

The test set consists of 232 samples where 106 negative samples out of 114 negative samples are correctly predicted and 112 positive samples out of 118 are correctly identified, as shown in Table 1.

There is a number of metrics used to validate prediction accuracy. Correct and actual prediction can be validated by Sensitivity, Specificity, Accuracy and Mathew's Correlation Coefficient. Accuracy, Sensitivity Specificity and Mathew's Correlation Coefficient are represented at many places in this study by Acc, Sn, Sp and Mcc respectively. Their formulation is also given below[34–36] where Sensitivity is applied to measure the probability of the model to predict target values. Mathew's Correlation Coefficient is used to evaluate the quality of the classification framework[37].

	Independent set test				Self-consistency test				Tenfold cross validation test				Jack Knife test			
	Acc (%)	Sn (%)	Sp (%)	MCC	Acc (%)	Sn (%)	Sp (%)	Mc	Acc (%)	Sn (%)	Sp (%)	MCC	Acc (%)	Sn (%)	Sp (%)	MCC
MLP	94	95	93	0.88	99	99	99	0.99	85	92	79	0.71	94	93	96	0.88
LR	93	92	93	0.85	97	97	96	0.93	88	91	82	0.74	93	92	94	0.86
RF	91	90	91	0.81	89	90	88	0.78	81	86	77	0.62	88	88	89	0.76

Table 2. Combined results of Multilayered Perceptron (MLP), Logistic Regression (LR) and Random Forest (RF).

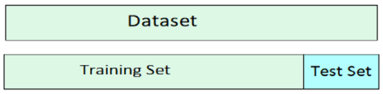

Figure 7. Sample dataset for independent set test.

$$\mathrm{Sn} = \frac{TP}{TP + FN} 0 \leq \mathrm{Sn} \leq 1 \tag{18}$$

$$\mathrm{Sp} = \frac{TN}{TN + FP} 0 \leq \mathrm{Sp} \leq 1 \tag{19}$$

$$\mathrm{ACC} = \frac{TP + TN}{TP + TN + FP + FN} 0 \leq \mathrm{Acc} \leq 1 \tag{20}$$

$$\mathrm{Mcc} = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}} - 1 \leq \mathrm{MCC} \leq 1 \tag{21}$$

The obtained sensitivity, specificity, accuracy and Mathew's Correlation Coefficient are 95%, 93%, 94% and 0.88 respectively. The obtained results validate the accuracy of the prediction model. Test methods are also applied for further validation which will be elaborated in "Test methods" section.

Test methods. There are many popular test methods in data mining and machine learning to evaluate the validity of the devised model. In this work, the independent set test, K-fold cross-validation test, and jackknife test are used to validate the devised model[38]. The independent test has 94% accuracy. K-fold cross-validation is performed with K = 10. The tenfold cross-validation test has 85% accuracy. Jackknife testing always gives you a unique value for the same dataset[8]. Jackknife testing is mostly used by an investigator to examine the quality of various predictors[38–50]. This study also uses a Jackknife test to check the quality of the predictor. The jackknife testing produced 94% accuracy. The result of all these test cases is given in Table 2. These test methods are also further explained in the coming subsections.

Independent set test. It is the basic performance metric of the proposed model in which obtained values from a confusion matrix are used to evaluate the accuracy of the model. The dataset is split into 80% training set and 20% test set and also shown in Fig. 7.

In this study, an independent set test has 94% Acc, 95% Sn, 93% Sp and is having 0.88 Mcc achieved by Multilayered Perceptron. The results of Logistic Regression and Random Forest results can also be seen in Table 2. Acc, Sn, Sp, and Mcc is mathematically described in Eqs. (18) – (21) respectively.

The area under the curve (AUC), obtained by Multilayered Perceptron, Logistic Regression and Random Forest are 97%, 97% and 95% respectively. The F1—score obtained by Multilayered Perceptron, Logistic Regression and Random Forest are 94%, 93% and 91% respectively. It also shows correctness of classifier. ROC-Curve is given in Fig. 8.

Self-consistency testing. This technique is used to have same data for both training and testing. The results are written in Table 2 and the ROC—Curve for Multilayered Perceptron, Logistic Regression and Random Forest is shown in Fig. 9.

K-fold cross-validation testing. It is a sampling technique used to validate the proposed models by using a limited number of data samples. It has a single parameter k which indicates the number of groups into which the data samples should be divided[51–53]. It is mostly used to evaluate the performance of the machine learning model to invisible data[54].

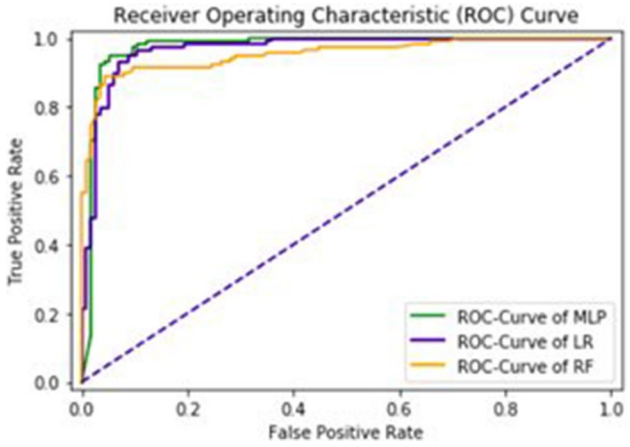

Figure 8. ROC-Curve of an independent set test.

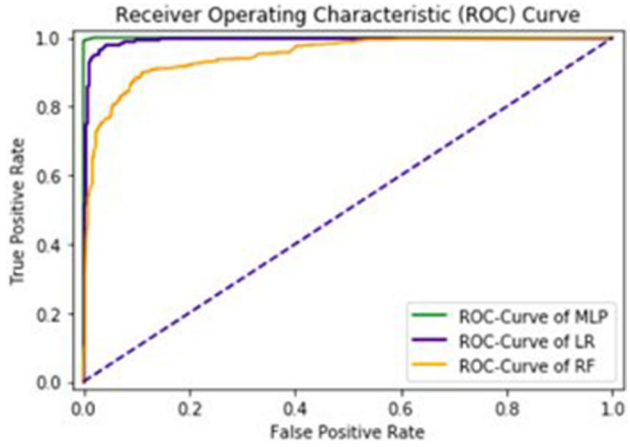

Figure 9. ROC-curve of self consistency test.

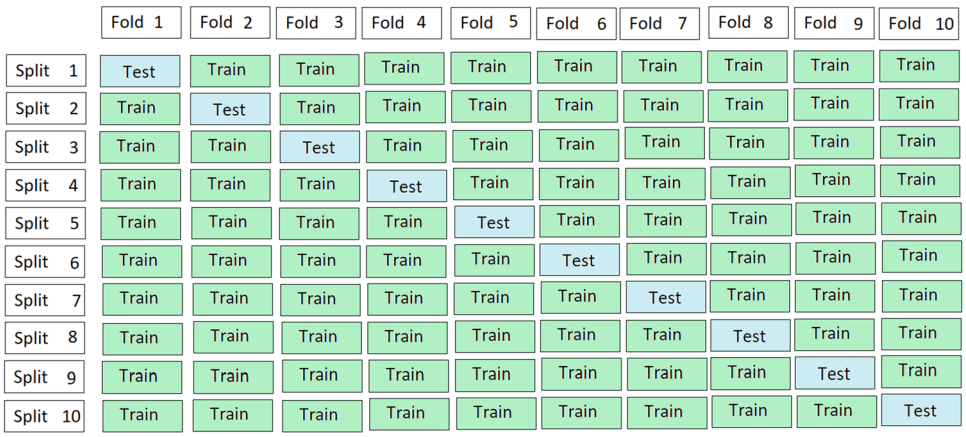

Figure 10. Tenfold cross validation process.

K can have any numeric value such as 5 or 10. In this work, tenfold cross validation sampling test is applied to evaluate the performance of the proposed model. The process of tenfold cross validation is also explained in Fig. 10. The data are divided into 10 equal observation sets (10 data samples). All the values such as Acc, An, Sp and Mcc are obtained for each observation set. The average of obtained accuracy for all observation sets is

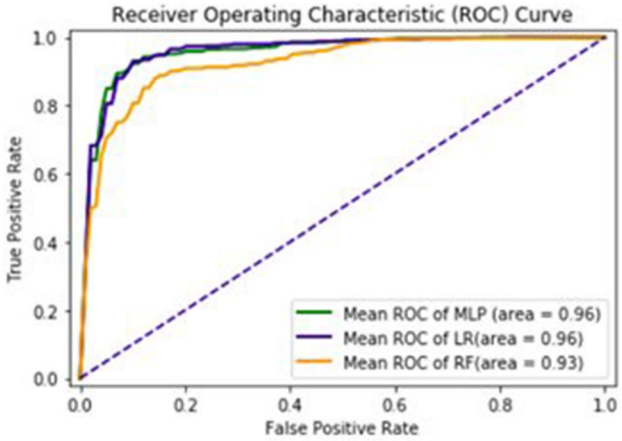

Figure 11. ROC-curve of tenfold cross validation test.

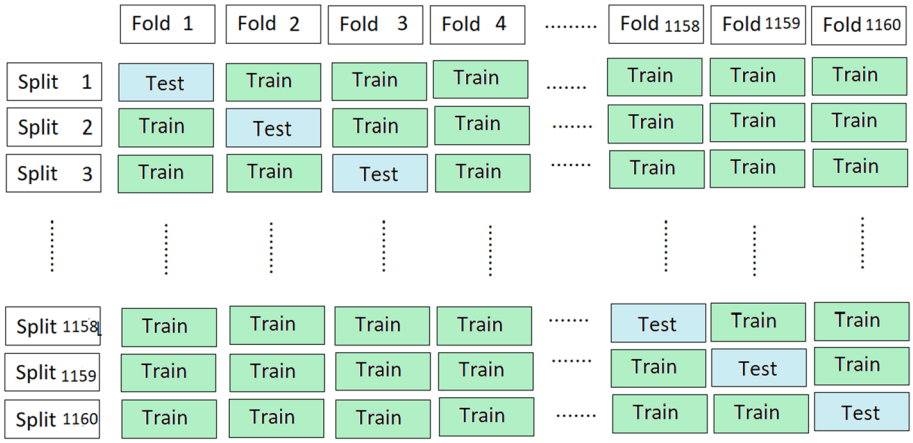

Figure 12. Jackknife sample test.

85%, average sensitivity is 92%, average specificity is 79% and average Mathew's correlation coefficient is 0.71 as given in Table 2.

The detailed of ROC-Curve of MLP, LR and RF is given in Fig. 11. The AUC of MLP, LR and RF are 0.96, 0.96 and 0.93 respectively.

Jackknife testing. It is considered a resample technique that is mostly used to compute the bias, mean and variance[55–57].

It evaluates the classification model sample by sample. The proposed classification model is validated on each sample using Jackknife testing and an average is computed of all the obtained results based on each sample. The process is also explained in Fig. 12. Overall observation samples are 1160 and therefore classification model is run 1160 times with obtained accuracy 94% along with sensitivity 93%, specificity 96% and Mathew's Correlation Coefficient 0.88.

The sequences are taken from a universal resource of protein (www.UniProt.org) through an advanced search. The chosen sequencings are streams of alphabets. It is difficult to process these sequences directly through the machine learning algorithm as they are unable to provide quantification measures. In order to address this issue, the feature vector is extracted from chosen sequences in a way that it has a strong correlation among features. In order to scale the obtained features, a standard normalization technique is used. A multilayered perceptron classifier is then applied to learn hidden patterns within observed features. Based on the said intelligent learning, observed features are going to be trained first which will then be a groundbreaking step for prediction. The validation of the proposed algorithm is carried out using a confusion matrix which is given in Table 1. Acc, Sn, Sp, and Mcc are estimated using FP, FN, TP, and TN within the confusion matrix which are 94%, 95%, 93% and 0.88 respectively as given in Table 2 and area under the curve is 0.97. Three different Machine learning algorithms are applied such as Multilayer Perceptron (MLP), Logistic Regression (LR) and Random Forest (RF). Four different types of tests are applied such as an independent set test, self-consistency test, cross validation test, and jackknife test. In this study it is clear from ROC curves that MLP is a better approach. The obtained results using different test cases validates the authenticity of our proposed model that it performs well even if the data set has large

variations. Along with independent set test, self-consistency test, tenfold cross-validation test and jackknife test also obtained very good results as given in Table 2.

Conclusion

Glutamate is an important type of common alpha-amino acid. 4-Carboxyglutamic acid is produced by a post-translational carboxylation of glutamic acid residues. This study is conducted to predict 4-carboxyglutamate following Chou's 5 steps rule. An MLP, RF and LR classification frameworks are adopted for the prediction of 4-carboxyglutamate residue sites. The accuracy of the independent set test, self-consistency test, tenfold cross-validation test, and Jackknife testing were determined to be 94%, 99%, 85% and 94%, respectively. A properly devised model will help in accurate detection of 4-carboxyglutamate which may be useful in evaluation of blood clotting, bone proteins, bone resorption, osteoporosis, papilloma and plaque atherosclerotic statuses.

Received: 8 March 2020; Accepted: 20 August 2020
Published online: 09 October 2020

References

1. Danbolt, N. C. Glutamate uptake. *Prog. Neurobiol.* **65**, 1–105 (2001).
2. Lee, C. A. *Textbook of Hemophilia* (Wiley, Hoboken, 2014).
3. Horava, S. D. & Peppas, N. A. Recent advances in hemophilia B therapy. *Drug Deliv. Transl. Res.* **7**, 359–371 (2017).
4. Suttie, J. W. Vitamin K-dependent carboxylase. *Annu. Rev. Biochem.* **54**, 459–477 (1985).
5. Burnier, J. P., Borowski, M., Furie, B. C. & Furie, B. Gamma-carboxyglutamic acid. *Mol. Cell. Biochem.* **39**, 91–207 (1981).
6. Pacifici, R. *et al.* Spontaneous release of interleukin 1 from human blood monocytes reflects bone formation in idiopathic osteoporosis. *Proc. Natl. Acad. Sci.* **84**, 4616–4620 (1987).
7. Malm, J., Cohen, E., Dackowski, W., Dahlback, B. & Wydro, R. Expression of completely gamma-carboxylated and beta-hydroxylated recombinant human vitamin-K-dependent protein S with full biological activity. *Eur. J. Biochem.* **187**, 737–743 (1990).
8. Gijsbers, B. L., Haarlem, L. J. V., Soute, B. A., Ebberink, R. H. & Vermeer, C. Characterization of a Gla-containing protein from calcified human atherosclerotic plaques. *Arteriosclerosis* **10**, 991–995 (1990).
9. Glutamic Acid. in*National Center for Biotechnology Information. PubChem Compound Database*. https://pubchem.ncbi.nlm.nih.gov/compound/Glutamic-acid. Accessed 26 Apr 2020.
10. -Carboxyglutamic acid. in*National Center for Biotechnology Information. PubChem Compound Database*. https://pubchem.ncbi.nlm.nih.gov/compound/4-Carboxyglutamic-acid#section=Structures. Accessed 26 Apr 2020.
11. Chou, K. C. Some remarks on protein attribute prediction and pseudo amino acid composition. *J. Theor. Biol.* **273**, 236–247 (2011).
12. Chou, K. C. Using subsite coupling to predict signal peptides. *Protein Eng.* **14**, 75–79 (2001).
13. Arif, M., Hayat, M. & Jan, Z. iMem-2LSAAC: A two-level model for discrimination of membrane proteins and their types by extending the notion of SAAC into Chou's pseudo amino acid composition. *J. Theor. Biol.* **442**, 11–21 (2018).
14. Contreras-Torres, E. Predicting structural classes of proteins by incorporating their global and local physicochemical and conformational properties into general Chous PseAAC. *J. Theor. Biol.* **454**, 139–145 (2018).
15. Feng, P.-M., Chen, W., Lin, H. & Chou, K.-C. iHSP-PseRAAAC: Identifying the heat shock protein families using pseudo reduced amino acid alphabet composition. *Anal. Biochem.* **442**, 118–125 (2013).
16. Javed, F. & Hayat, M. Predicting subcellular localization of multi-label proteins by incorporating the sequence features into Chous PseAAC. *Genomics* **111**, 1325–1332 (2018).
17. Krishnan, S. M. Using Chous general PseAAC to analyze the evolutionary relationship of receptor associated proteins (RAP) with various folding patterns of protein domains. *J. Theor. Biol.* **445**, 62–74 (2018).
18. Sankari, E. S. & Manimegalai, D. Predicting membrane protein types by incorporating a novel feature set into Chous general PseAAC. *J. Theor. Biol.* **455**, 319–328 (2018).
19 Khan, Y. D., Rasool, N., Hussain, W., Khan, S. A. & Chou, K. C. iphosY-PseAAC: Identify phosphotyrosine sites by incorporating sequence statistical moments into PseAAC. *Mol. Biol. Rep.* **45**, 2501–2509 (2018).
20. Chou, K. C. Impacts of bioinformatics to medical chemistry. *Med. Chem.* **11**, 218–234 (2015).
21. Chou, K. C. Impacts of bioinformatics to medical using pseudo-amino acid composition. *Proteins* **43**, 246–255 (2001).
22. Khan, Y. D., Ahmad, F. & Anwar, M. W. A neuro-cognitive approach for iris recognition using backpropagation. *World Appl. Sci. J.* **16**, 678–685 (2012).
23. Khan, Y. D., Ahmed, F. & Khan, S. A. Situation recognition using image moments and recurrent neural networks. *Neural Comput. Appl.* **24**, 1519–1529 (2013).
24. Butt, H., Khan, S. A., Jamil, H., Rasool, N. & Khan, Y. D. A prediction model for membrane proteins using moments based features. *Biomed. Res. Int.* **2016**, 1–7 (2016).
25. Butt, H., Rasool, N. & Khan, Y. D. A treatise to computational approaches towards prediction of membrane protein and its subtypes. *J. Membr. Biol.* **250**, 55–76 (2016).
26. Khan, Y. D. *et al.* An efficient algorithm for recognition of human actions. *Sci. World J.* **2014**, 1–11 (2014).
27 Khan, Y. D., Khan, S. A., Ahmad, F. & Islam, S. Iris recognition using image moments and k-means algorithm. *Sci. World J.* **2014**, 1–9 (2014).
28. Khan, Y. D., Rasool, N., Hussain, W., Khan, S. A. & Chou, K. C. iPhosT-PseAAC: Identify phosphothreonine sites by incorporating sequence statistical moments into PseAAC. *Anal. Biochem.* **550**, 109–116 (2018).
29. Akmal, M. A., Rasool, N. & Khan, Y. D. Prediction of N-linked glycosylation sites using position relative features and statistical moments. *PLoS ONE* https://doi.org/10.1371/journal.pone.0181966 (2017).
30. sklearn.preprocessing.StandardScaler. scikit. https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html. Accessed 8 Mar 2020.
31. Wan, S., Liang, Y., Zhang, Y. & Guizani, M. Deep multi-layer perceptron classifier for behavior analysis to estimate Parkinson's disease severity using smartphones. *IEEE Access.* **6**, 36825–36833 (2018).
32. Gajoui, K. E., Allah, F. A. & Oumsis, M. Diacritical language OCR based on neural network: Case of Amazigh language. *Procedia Comput. Sci.* **73**, 298–305 (2015).
33. Zhai, X., Ali, A. A. S., Amira, A. & Bensaali, F. MLP neural network based gas classification system on Zynq SoC. *IEEE Access.* **4**, 8138–8146 (2016).
34. Chen, J., Liu, H., Yang, J. & Chou, K.-C. Prediction of linear B-cell epitopes using amino acid pair antigenicity scale. *Amino Acids* **33**, 423–428 (2007).
35. Xu, Y., Ding, J., Wu, L.-Y. & Chou, K.-C. iSNO-PseAAC: Predict cysteine S-nitrosylation sites in proteins by incorporating position specific amino acid propensity into pseudo amino acid composition. *PLoS ONE* **8**, e55844 (2013).

36 Chen, W., Feng, P.-M., Lin, H. & Chou, K.-C. iRSpot-PseDNC: Identify recombination spots with pseudo dinucleotide composition. *Nucleic Acids Res.* **41**, e68 (2013).
37. Porter, J., Berkhahn, J. & Zhang, L. A comparative analysis of read mapping and indel calling pipelines for next-generation sequencing data. In *Emerging Trends in Computational Biology, Bioinformatics, and Systems Biology* (eds Tran, Q. N. & Arabnia, H.) 521–535 (Elsevier, Amsterdam, 2015).
38. Chou, K.-C. & Zhang, C.-T. Prediction of protein structural classes. *Crit. Rev. Biochem. Mol. Biol.* **30**, 275–349 (1995).
39. Ali, F. & Hayat, M. Classification of membrane protein types using voting feature interval in combination with Chou's pseudo amino acid composition. *J. Theor. Biol.* **384**, 78–83 (2015).
40. Zhou, G.-P. & Doctor, K. Subcellular location prediction of apoptosis proteins. *ProteinsStruct. Funct. Bioinform.* **50**, 44–48 (2002).
41. Mondal, S. & Pai, P. P. Chou's pseudo amino acid composition improves sequence-based antifreeze protein prediction. *J. Theor. Biol.* **356**, 30–35 (2014).
42. Feng, K.-Y., Cai, Y.-D. & Chou, K.-C. Boosting classifier for predicting protein domain structural class. *Biochem. Biophys. Res. Commun.* **334**, 213–217 (2005).
43. Nanni, L., Brahnam, S. & Lumini, A. Prediction of protein structure classes by incorporating different protein descriptors into general Chou's pseudo amino acid composition. *J. Theor. Biol.* **360**, 109–116 (2014).
44. Shen, H.-B., Yang, J. & Chou, K.-C. Euk-PLoc: An ensemble classifier for large-scale eukaryotic protein subcellular location prediction. *Amino Acids* **33**, 57–67 (2007).
45 Wu, Z.-C., Xiao, X. & Chou, K.-C. iLoc-Plant: A multi-label classifier for predicting the subcellular localization of plant proteins with both single and multiple sites. *Mol. BioSyst.* **7**, 3287 (2011).
46. Dehzangi, A. *et al.* Gram-positive and gram-negative protein subcellular localization by incorporating evolutionary-based descriptors into Chou's general PseAAC. *J. Theor. Biol.* **364**, 284–294 (2015).
47. Qiu, W.-R., Xiao, X. & Chou, K.-C. iRSpot-TNCPseAAC: Identify recombination spots with trinucleotide composition and pseudo amino acid components. *Int. J. Mol. Sci.* **15**, 1746–1766 (2014).
48. Kumar, R., Srivastava, A., Kumari, B. & Kumar, M. Prediction of β-lactamase and its class by Chou's pseudo-amino acid composition and support vector machine. *J. Theor. Biol.* **365**, 96–103 (2015).
49 Chen, J., Long, R., Wang, X.-L., Liu, B. & Chou, K.-C. dRHP-PseRA: Detecting remote homology proteins using profile-based pseudo protein sequence and rank aggregation. *Sci. Rep.* https://doi.org/10.1038/srep32333 (2016).
50. Ahmad, K., Waris, M. & Hayat, M. Prediction of protein submitochondrial locations by incorporating dipeptide composition into Chou's general pseudo amino acid composition. *J. Membr. Biol.* **249**, 293–304 (2016).
51. Duchesnay, E. & Löfstedt, T. *Statistics and Machine Learning in Python Release 0.2.* (2018).
52. Adams, R. P. *Model Selection and Cross Validation Evaluation Hygiene: The Train/Test Split,* 1–8.
53. Anguita, D. Ghelardoni, L. Ghio, A. Oneto, L & Ridella, S. The 'K' in K-fold cross validation. in*European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 441–446 (2012).
54. Rodríguez, J. D., Pérez, A. & Lozano, J. A. Sensitivity analysis of k-fold cross validation in prediction error estimation. *IEEE Trans. Pattern Anal. Mach. Intell.* **32**, 569–575 (2010).
55. Chapter 8 Bootstrap and Jackknife Estimation of Sampling. https://www.stat.washington.edu/jaw/COURSES/580s/581/LECTNOTES/ch8.pdf. Accessed 24 May 2019.
56. G Protein-Coupled Receptor 172A (GPR172A) ELISA Kit. Human GPR172A ELISA Kit (ABIN5654457). https://www.antibodies-online.com/kit/5654457/GProtein-CoupledReceptor172AGPR172AELISAKit/. Accessed 8 Mar 2020.
57. Lavergne, C. A Jackknife method for estimation of variance components. *Statistics* **27**, 1–13 (1995).

Author contributions

A.A.S.: Manuscript write up; Machine learning algorithm implementation; Obtaining results; Applying all types of test cases. Y.D.K.: Bench mark dataset; Sample formulation; Statistical Moment Calculation; Guidance in the whole process.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to A.A.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.