BIOENGINEERED, 2018
VOL. 9, NO. 1, 196–202
https://doi.org/10.1080/21655979.2017.1373536

Taylor & Francis
Taylor & Francis Group

RESEARCH PAPER

OPEN ACCESS

Check for updates

The effect of three novel feature extraction methods on the prediction of the subcellular localization of multi-site virus proteins

Lei Wang[a,b], Yaou Zhao[a,b], Yuehui Chen[a,b], and Dong Wang[a,b]

[a]School of Information Science and Engineering, University of Jinan, Jinan, China; [b]Shandong Provincial Key Laboratory of Network Based Intelligent Computing, Jinan, China

ABSTRACT

Experimental methods play a crucial role in identifying the subcellular localization of proteins and building high-quality databases. However, more efficient, automated computational methods are required to predict the subcellular localization of proteins on a large scale. Various efficient feature extraction methods have been proposed to predict subcellular localization, but challenges remain. In this paper, three novel feature extraction methods are established to improve multi-site prediction. The first novel feature extraction method utilizes repetitive information via moving windows based on a dipeptide pseudo amino acid composition method (R-Dipeptide). The second novel feature extraction method utilizes the impact of each amino acid residue on its following residues based on pseudo amino acids (I-PseAAC). The third novel feature extraction method provides local information about protein sequences that reflects the strength of the physicochemical properties of residues (PseAAC2). The multi-label k-nearest neighbor algorithm (MLKNN) is used to predict the subcellular localization of multi-site virus proteins. The best overall accuracy values of R-Dipeptide, I-PseAAC, and PseAAC2 when applied to dataset S from Virus-mPloc are 59.92%, 59.13%, and 57.94% respectively.

ARTICLE HISTORY
Received 5 July 2017
Accepted 5 July 2017

KEYWORDS
feature extraction; I-PseAAC; PseAAC2; R-Dipeptide; subcellular localization

Introduction

Knowledge about the subcellular localization of proteins is critical for understanding their functions and biological processes in cells.[1] High-quality databases of information on the subcellular localization of proteins are informed by wet laboratory experiments. However, such experiments are time-consuming, costly and laborious.[2] Experimental methods for handling proteins on a large scale have become increasingly difficult. It is necessary to develop effective computational methods to analyze subcellular localization.[3] The web servers proposed to identify the subcellular localization of proteins based on their sequence information can be classified into two series.[4] One is the "PLoc" series, and the other is the "iLoc" series. The "PLoc" series includes six web servers to handle eukaryotic, plant, human, gram-negative bacterial, gram-positive bacterial, and viral proteins, while the "iLoc" series includes seven web servers to handle eukaryotic, plant, human, animal, gram-negative bacterial, gram-positive bacterial, and viral proteins.[5] Many studies have indicated that greater progress in prediction systems is obtained by developing feature extraction methods than by improving the classifiers.[6,7]

In recent years, a wide range of feature extraction methods have been proposed to improve the performance of prediction: (1) amino acid composition (AAC) methods;[8-11] (2) homology-based methods;[7,12] (3) sorting signal-based methods;[13-14] and (4) pseudo amino acid-based feature methods (PseAAC).[15-17] All these methods have shown good performance but could be improved. AAC methods lack location information; homology-based methods are not suitable for low-homology protein sequences; and PseAAC can reflect some of the effects of sequence order but lacks the impact of each residue on the subsequent residues. Therefore, three feature extraction methods are proposed to improve the performance of multi-site prediction.

The three novel feature extraction methods proposed in this study are called R-Dipeptide, I-PseAAC and PseAAC2. Inspired by the long short-term memory

CONTACT Lei Wang ✉ 3027630499@qq.com

with attention mechanism (A-LSTM),[18] R-Dipeptide focuses on using repetitive information. First, the spacing between two windows is set by the user, often to a small number. In this study, the spacing is one. Then, two better protein sub-sequences are selected according to the prediction results and combined. This method makes up for the lack of extraction of key information by PseAAC. I-PseAAC computes the impact of each amino acid residue on the subsequent residues. This method offers global order information, rather than the local order information provided by PseAAC. PseAAC2 focuses on location information. This method not only offers global order information but also adds the relative strengths of the residues, whereas PseAAC lacks information on the relative strengths of residues.

Material and methods

Dataset

Dataset S, constructed by Shen in establishing Virus-mPloc, is the benchmark dataset for the study.[19] Dataset S offers three advantages. (1) The dataset is specialized for virus proteins. (2) None of the proteins included in S has $\geq$ 25% pairwise sequence identity to any other protein in the same location. (3) The dataset includes proteins with more than one location and thus can be utilized to address the subcellular localization of multi-site virus proteins.[20]

Dataset S includes 207 virus protein sequences, of which 165 belong to one subcellular location, 39 to two locations, and 3 to three locations.[20] The dataset is classified into 6 subcellular locations,[21] as expressed in Eq. 1:

$$S = S1 \cup S2 \cup S3 \cup S4 \cup S5 \cup S6 \tag{1}$$

where *S1* represents the subset for the subcellular location "viral capsid", *S2* the subset for "host cell membrane", and so forth (Table 1), while $\cup$ denotes "union" in set theory.[21]

Table 1. The benchmark dataset S taken from Virus-mPloc[21].

Subset	Subcellular location	Number of proteins
S1	Host viral capsid	8
S2	Host cell membrane	33
S3	Host endoplasmic reticulum	20
S4	Host cytoplasm	87
S5	Host nucleus	84
S6	Secreted	20
Total number of locative proteins		252
Total number of different proteins		207

Here, the locative protein sequences and different protein sequences are briefly described as follows. Locative proteins are described by Eq. 2:

$$N(locative) = N(different) + \sum_{m=1}^{M} (m-1)N(m) \tag{2}$$

where *N(locative)* represents the number of locative proteins and *N(different)* represents the number of different proteins. Here, *m* is the number of locations where the specific protein is identified, and *N(m)* is the number of proteins that are identified in *m* locations.

R-Dipeptide

R-Dipeptide utilizes repetitive information via moving windows based on a dipeptide pseudo amino acid composition method.

First, the number of each amino acid residue in every protein sequence is calculated in Eq. 3. Then, the number of residues is normalized in Eq. 4.

$$V = [v_1, v_2, v_3, \ldots, v_i, \ldots, v_{20}] \tag{3}$$

where v_i is the number of the *i-th* type of residue in every protein sequence.

$$v_i^* = \frac{v_i - \mu}{\sigma} \tag{4}$$

where v_i^* is the normalized value of v_i, μ denotes the mean of v_i, and σ represents the standard deviation of v_i.

Second, the spacing between two windows is set to one, and the window size is set to thirty. The sub-sequence of the first group is $\{R_1,R_2,\ldots,R_{30}\}$, the sub-sequence of the second group is $\{R_2,R_3,\ldots,R_{31}\}$, and so forth. For the last residue (R_L), L is smaller than the minimum length of all protein sequences. Then, two improved protein sub-sequences are combined to create a new database based on the prediction results. The new database contains important repetitive information. that contributes to the prediction of subcellular localization.

Lastly, a dipeptide pseudo amino acid composition method (Dipeptide) is used for the new database. Dipeptide will generate 400 components, i.e., AA, AC, AD, …, YV, YW, and YY. These 400 components are calculated for every protein sequence and then subjected to a standard conversion.

I-PseAAC

PseAAC is proposed by Chou and avoids losing the ordering information of protein sequences.[23]

A protein (P) including L amino acid residues can be described by Eq. 5:

$$P = R_1, R_2, R_3, \ldots\ldots, R_L \qquad (5)$$

where R_1 is the first residue of the protein sequence P, R_2 is the second residue of the protein sequence P, and so forth.

The sequence order information can be represented by Eq. 6.

$$\delta_\theta = \sum_{i=1}^{L-\theta} \Omega(R_i, R_{i+\theta})/(L-\theta),\ \ (\theta = 1, 2, \ldots, \text{n and } n < L) \qquad (6)$$

where δ_θ is the θ-*th* correlation factor, which provides the sequence order information between the θ most contiguous residues. $\Omega(R_i, R_{i+1})$ can be described by Eq. 7:

$$\begin{aligned}\Omega(R_i, R_{i+1}) = \frac{1}{6}\{&[H_1(R_j) - H_1(R_i)]^2 + [H_2(R_j) \\ &- H_2(R_i)]^2 + [Pk_1(R_j) - Pk_1(R_i)]^2 \\ &+ [Pk_2(R_j) - Pk_2(R_i)]^2 + [PI(R_j) \\ &- PI(R_i)]^2 + [M(R_j) - M(R_i)]^2\} \qquad (7)\end{aligned}$$

where $H_1(R_i)$, $H_2(R_i)$, $Pk_1(R_i)$, $Pk_2(R_i)$, $PI(R_i)$, and $M(R_i)$ denote the hydrophobicity value, the hydrophilicity value, *Pk1(-COOH), Pk2(-NH3), PI*, and the mass value of the amino acid residue R_i, respectively.

All physicochemical properties should be normalized before being used in the calculation of Eq. 7.

In contrast to PseAAC, I-PseAAC utilizes the impact of each residue on the subsequent residues. I-PseAAC is described in Fig. 1(a), Fig. 1(b) and Fig. 1(c).

Fig. 1(a), Fig. 1(b) and Fig. 1(c) show the process of I-PseAAC. PseAAC calculates the order information for (R_1,R_2), (R_2,R_3), (R_3,R_4) and so forth, while I-PseAAC calculates the or information for (R_1,R_2), (R_1,R_3),(R_1,R_4) and so forth. The details are shown in Fig. 1(a), Fig. 1 (b) and Fig. 1(c). In (R_i, R_j), j is greater than i.

PseAAC2

In contrast to PseAAC and I-PseAAC, PseAAC2 provides a different kind of local information to reflect the strength of the physicochemical properties of residues, as described in Eq. 8 and Eq. 9:

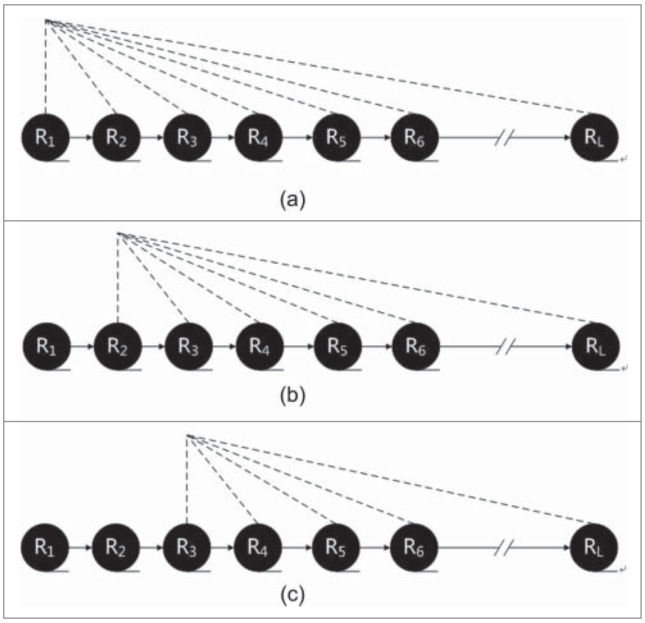

Figure 1. (a) The impact of each residue on the subsequent residues. Fig. 1(b). The impact of each residue on the subsequent residues. Fig. 1(c). The impact of each residue on the subsequent residues.

$$\begin{aligned}\Omega(R_i) = \frac{1}{6}&[H_1(R_i)^2 + H_2(R_i)^2 + Pk_1(R_i)^2 \\ &+ Pk_2(R_i)^2 + PI(R_i)^2 + M(R_i)^2] \qquad (8)\end{aligned}$$

$$\Omega(R_i, R_j) = \Omega(R_i) * R(R_j) \qquad (9)$$

MLKNN

MLKNN is a multi-label classifier that utilizes the k-nearest neighbor algorithm to collect the category tag information of neighbor samples and exploits the principle of maximum posterior probability to infer the "no example of label" set.[21,24] MLKNN can be described by Eq. 10 and Eq. 11:

$$C_j = \sum_{(x,Y)\in N(x)} \{y_j \in Y\} \qquad (10)$$

where C_j represents the number of neighbors of x belonging to class $N(x)$.[21]

$$h(x) = \left\{ y_j \,\middle|\, \frac{P(H_j \mid C_j)}{P(H\neg_j \mid C_j)} > 0.5, 1 \le j \le q \right\} \qquad (11)$$

where H_j denotes the event of x including the category y_j. $P(H_j|y_j)$ denotes the posterior probability set H_j that $N(x)$ contains the number C_j in the category y_j.

Evaluation

To provide a more intuitive and easier-to-understand measurement, a new scale, the so-called "absolute true" overall accuracy,[20] reflecting the accuracy of a predictor, is given in Eq. 12:

$$\Lambda = \frac{\sum_{i=1}^{N}\Delta(i)}{N} \tag{12}$$

where Λ represents the absolute true rate, N represents the number of total proteins investigated, and $\Delta(i) = 1$ or $\Delta(i) = 0$.

All subcellular locations of the *i-th* protein will be tested. If every subcellular location of the *i-th* protein is correctly predicted, $\Delta(i) = 1$; otherwise, $\Delta(i) = 0$.

Therefore, the absolute true scale is much stricter than the scale used previously to measure the overall accuracy.

In addition, a series of other evaluation functions are applied to evaluate the prediction performance.[22]

HammingLoss:

$$HammingLoss(h) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{C}\left|h(x_i)\Delta y_i\right| \tag{13}$$

output value of labels-one should be higher than the output value in labels-zero. RankingLoss is utilized to calculate how many times the output lacks an appropriate comparison. A lower value of RankingLoss indicates better algorithm performance.

One_error:

$$One\_error(h) = \frac{1}{N}\sum_{i=1}^{N}\left\{\left[\arg\max_{y\in Y} h(x_i, y)\right] \notin Y_i\right\} \tag{15}$$

One_error calculates how many times the top label is not in the appropriate label sets. A lower value of One_error represents better algorithm performance.

Coverage:

$$Coverage(h) = \frac{1}{N}\sum_{i=1}^{N}\frac{\max rank^h(x_i, y) - 1}{C} \tag{16}$$

Coverage is utilized to calculate how far down the label set of an instance it is necessary to go. A lower value of Coverage indicates better algorithm performance.

Average_Precision:

$$Average\_Precision(h) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{|C_i|}\sum_{y\in C_i}\frac{\left|\left\{y' \in C_i \mid rank^h(x_i, y') \le rank^h(x_i, y)\right\}\right|}{rank^h(x_i, y)} \tag{17}$$

HammingLoss is utilized to calculate how many times a label is misclassified. A lower value of HammingLoss represents better algorithm performance.

RankingLoss:

$$RankingLoss(h) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{|C_i|\,|\overline{C_i}|} \cdot\left\{(y_1, y_2)\middle|h(x_i, y_1) \le h(x_i, y_2)\right\} \tag{14}$$

C_i is the collection of labels with a value of one, denoted by labels-one. $\overline{C_i}$ is the collection of labels with a value of zero, denoted by labels-zero. If the predictive labels of an instance are completely correct, the

Average_Precision is utilized to calculate the average fraction of labels ranked. A higher value of Average_Precision represents better algorithm performance.

Results

In this study, the spacing between two windows is set to 1, and the window size is set to 30. The database is divided into 24 groups: (0,30) is the first group, (1,30) is the second group, and so forth. The number of each amino acid residue in every group is calculated in Eq. 3. and Eq. 4. The overall accuracy of each group is shown in Table 2.

Table 2. Sorting signals of database.

(0,30)	(1,30)	(2,30)	(3,30)	(4,30)	(5,30)	(6,30)	(7,30)
50%	45.63%	50%	54.76%	**57.54%**	53.57%	**57.54%**	47.22%
(8,30)	(9,30)	(10,30)	(11,30)	(12,30)	(13,30)	(14,30)	(15,30)
55.56%	46.43%	44.05%	41.27%	49.21%	50%	43.25%	50.79%
(16,30)	(17,30)	(18,30)	(19,30)	(20,30)	(21,30)	(22,30)	(23,30)
50.79%	51.98%	55.56%	56.35%	46.83%	46.03%	46.03%	44.05%

Table 3. Application of two methods to original database and new database.

AAC in original dataset	R-AAC	Dipeptide in original dataset	R-Dipeptide
55.16%	**58.33%**	54.76%	**59.92%**

Table 4. Six physicochemical properties.

hydrophobicity	hydrophilicity	Pk1	Pk2	PI	mass
0.62	−0.5	2.35	9.87	6.11	15
0.29	−1	1.71	10.78	5.02	47
−0.9	3	1.88	9.6	2.98	59
−0.74	3	2.19	9.67	3.08	73
1.19	−2.5	2.58	9.24	5.91	91
0.48	0	2.34	9.6	6.06	1
−0.4	−0.5	1.78	8.97	7.64	82
1.38	−1.8	2.32	9.76	6.04	57
−1.5	3	2.2	8.9	9.47	73
1.06	−1.8	2.36	9.6	6.04	57
0.64	−1.3	2.28	9.21	5.74	75
−0.78	0.2	2.18	9.09	10.76	58
0.12	0	1.99	10.6	6.3	42
−0.85	0.2	2.17	9.13	5.65	72
−2.53	3	2.18	9.09	10.76	101
−0.18	0.3	2.21	9.15	5.68	31
−0.05	−0.4	2.15	9.12	5.6	45
1.08	−1.5	2.29	9.74	6.02	43
0.81	−3.4	2.38	9.39	5.88	130
0.26	−2.3	2.2	9.11	5.63	107

Table 5. Application of PseAAC, R-Dipeptide, I-PseAAC and PseAAC2 to the new database.

PseAAC	R-Dipeptide	I-PseAAC	PseAAC2
57.14%	**59.92%**	**59.13%**	**57.94%**

The overall accuracy of the original database is 55.16%, while the best overall accuracy of the groups is 57.54%. Table 2 demonstrates the effect of the sorting-signal method. Group 5 and group 7 are combined to create a new database. For example, group 5 {ACDVY} and group 7 {DVYWY} are converted to the new database {ACDVYDVYWY}. The important repetitive information is {DVY}. If both group 5 and group 7 show good performance, we believe that the two groups share important information {DVY} for the prediction of subcellular localization.

As shown in Table 3, the two methods AAC and Dipeptide give better results when applied to the new database than when applied to the original database. The original database contains redundant information. Therefore, the methods cannot obtain better performance when applied to the original database. The new database utilizes the repetitive information from sub-sequences. This approach is equivalent to increasing the weight of key residues.

Six physicochemical properties are used in the PseAAC2 and I-PseAAC methods: the hydrophobicity, hydrophilicity, *Pk1(-COOH), Pk2(-NH3), PI* and mass values of each amino acid residue, as described in Table 4.

The three novel feature extraction methods are compared with PseAAC.[23] Group 5 and group 7 are combined to create a new database, and four feature extraction methods are used in the new database to identify the subcellular localization of multi-site virus proteins by MLKNN. The results of the PseAAC method are obtained via a web server called PseAAC at http://www.csbio.sjtu.edu.cn/bioinf/PseAAC/#. The weight factor is 0.05, and the Lambda parameter is 40.

As shown in Table 5, the three novel feature extraction methods show superior performance, achieving 59.92%, 59.13%, and 57.94% accuracy for the MLKNN algorithm. The PseAAC method shows 57.14% accuracy for MLKNN algorithm. Thus, the three novel feature extraction methods improve the performance of multi-site prediction.

As shown in Table 6, the number of correct predictions of every subcellular location is calculated by Eq. 12. The overall accuracy is the sum of the correct predictions.

Table 6. Overall accuracy of R-Dipeptide, I-PseAAC, and PseAAC2.

	Overall accuracy		
Subcellular location	R-Dipeptide	I-PseAAC	PseAAC2
Viral capsid	7/8 = 87.5%	7/8 = 87.5%	7/8 = 87.5%
Host cell membrane	12/33 = 36.36%	13/33 = 39.39%	13/33 = 39.39%
Host endoplasmic reticulum	11/20 = 55%	11/20 = 55%	10/20 = 50%
Host cytoplasm	49/87 = 56.32%	52/87 = 59.77%	51/87 = 58.62%
Host nucleus	59/84 = 70.24%	51/84 = 60.71%	52/84 = 61.9%
Secreted	13/20 = 65%	15/20 = 75%	13/20 = 65%
Overall accuracy	151/252 = 59.92%	149/252 = 59.13%	146/252 = 57.94%

Table 7. Evaluation functions for PseAAC, R-Dipeptide, I-PseAAC, and PseAAC2.

FEM	A	C	H	O	R
PseAAC	0.7662	0.1798	0.1428	0.3888	0.1396
R-Dipeptide	0.7838↑	0.1712↓	0.1296↓	0.3571↓	0.1293↓
I-PseAAC	0.7684↑	0.1798	0.1355↓	0.3809↓	0.1396
PseAAC2	0.7686↑	0.1772↓	0.1329↓	0.3849↓	0.1365↓

To simplify the representation of the evaluation functions, Average_Precision is denoted by A, Coverage is denoted by C, HammingLoss is denoted by H, One_error is denoted by O, and RankingLoss is denoted by R. The calculation details of the five evaluation functions are described in Eq. 13-17. The feature extraction method is denoted by FEM.

As shown in Table 7, R-Dipeptide, I-PseAAC, PseAAC2 all show better performance than PseAAC in general.

Conclusion and discussion

In this study, three novel feature extraction methods are proposed to improve the performance of multi-site prediction. In experimental comparisons, the R-Dipeptide, I-PseAAC, and PseAAC2 methods achieve higher accuracy rates for the MLKNN algorithm than does the PseAAC method. Thus, repetitive information, the impact of each residue on subsequent residues, and local information are critical for the performance of multi-site prediction. The advantage of R-Dipeptide is the extraction of key information using the repetitive information method. We are accustomed to extracting key information by weight adjustment of the algorithm. For a large-scale dataset, weight adjustment is an effective method for the extraction of key information. However, if the dataset is limited in scale, the repetitive information method is better than the weight adjustment method. The advantage of I-PseAAC is that it can reflect the difference in physicochemical properties between each amino acid residue and the subsequent residues. In addition, I-PseAAC provides global information on the residues. The disadvantage is that the difference between the *i-th* residue and the *j-th* residue may be the same as the difference between the *i-th* residue and the *k-th* residue. For example, two kinds of physicochemical properties are denoted by A and B, respectively. The difference in A between the *i-th* residue and the subsequent *j-th* residue is 0.2, and the difference in B is −0.2. The difference between the *i-th* residue and the subsequent *k-th* residue in A is 0.3, and the difference in B is −0.3. Thus, there is no difference between the *j-th* residue and the *k-th* residue. The advantage is that PseAAC2 amplifies the differences in the physicochemical properties of different residues by providing another source of local information about protein sequences. The disadvantage is how to choose a set of representative physicochemical properties. If the values of the physicochemical properties of different residues are different, this kind of physicochemical property is representative. If some of the residues have the same physicochemical property values, the performance of PseAAC2 will decline.

The three novel feature extraction methods have shown good performance but can still be improved. The first question is how to set an appropriate window size and spacing between two windows. If the window is too small, important information will be lost and a large number of groups will be generated. If the window is too large, too much redundant information will be generated. If the spacing between two windows is too large, repeat information will be lost. In addition, groups can be combined in a variety of ways, such as adjacent groups (group 4, group 5), interval groups (group 4, group 7), or more than two groups (group 4, group 5, group 7). Our future studies will focus on these questions with regard to subcellular localization.

Disclosure of potential conflicts of interest

No potential conflicts of interest were disclosed.

Funding

This research was supported by the National Key Research And Development Program of China (No. 2016YFC0106000), the National Natural Science Foundation of China (Grant No. 61302128, 61573166, 61572230, 61671220, 61640218), the Youth Science and Technology Star Program of Jinan City (201406003), the Natural Science Foundation of Shandong Province (ZR2013FL002), the Shandong Distinguished Middle-aged and Young Scientist Encourage and Reward Foundation, China (Grant No. ZR2016FB14), the Project of Shandong Province Higher Educational Science and Technology Program, China (Grant No. J16LN07), the Shandong Province Key Research and Development Program, China (Grant No. 2016GGX101022), and the Research Fund for the Doctoral Program of the University of Jinan (No. XBS1604).

References

[1] Wei L, Liao M, Gao X, Wang J, Lin W. mGOF-loc: A novel ensemble learning method for human protein subcellular localization prediction. Neurocomputing. 2016; 217:73-82. doi:10.1016/j.neucom.2015.09.137.

[2] Wan S, Mak MW, Kung SY. mLASSO-Hum: A LASSO-based interpretable human-protein subcellular localization predictor. J Theor Biol. 2015;382:223-34. doi:10.1016/j.jtbi.2015.06.042. PMID:26164062

[3] Xiao X, Wu ZC, Chou KC. iLoc-Virus: A multi-label learning classifier for identifying the subcellular localization of virus proteins with both single and multiple sites. J Theor Biol. 2011;284:42-51. doi:10.1016/j.jtbi.2011.06.005. PMID:21684290

[4] Chou KC. Impacts of bioinformatics to medicinal chemistry. Med Chem. 2015;11:218-34. doi:10.2174/1573406411666141229162834. PMID:25548930

[5] Wang X, Li H, Zhang Q, Wang R. Predicting Subcellular Localization of Apoptosis Proteins Combining GO Features of Homologous Proteins and Distance Weighted KNN Classifier. Biomed Res Int. 2016;2016:1-8.

[6] Sharma R, Dehzangi A, Lyons J, Paliwal K, Tsunoda T, Sharma A. Predict Gram-Positive and Gram-Negative Subcellular Localization via Incorporating Evolutionary Information and Physicochemical Features Into Chou's General PseAAC. IEEE Trans Nanobiosci. 2015;14:915-926. doi:10.1109/TNB.2015.2500186.

[7] Dehzangi A, Heffernan R, Sharma A, Lyons J, Paliwal K, Sattar A. Gram-positive and gram-negative protein subcellular localization by incorporating evolutionary-based descriptors into Chou's general PseAAC. Theor Biol. 2015;364:284-94. doi:10.1016/j.jtbi.2014.09.029.

[8] Nakashima H, Nishikawa K. Discrimination of Intracellular and Extracellular Proteins Using Amino Acid Composition and Residue-pair Frequencies. J Mol Biol 1994;238:54-61. doi:10.1006/jmbi.1994.1267. PMID:8145256

[9] Feng PM, Chen W, Lin H, Chou KC. iHSP-PseRAAAC: Identifying the heat shock protein families using pseudo reduced amino acid alphabet composition. Analyt Biochem. 2013;442:118-25. doi:10.1016/j.ab.2013.05.024. PMID:23756733

[10] Deng Y, Luo YL, Wang Y, Zhao Y. Effect of different drying methods on the myosin structure, amino acid composition, protein digestibility and volatile profile of squid fillets. Food Chem. 2015;171:168-76. doi:10.1016/j.foodchem.2014.09.002. PMID:25308657

[11] Mandal S, Das G, Askari H. Amino acid-type interactions of L-3,4-dihydroxyphenylalanine with transition metal ions: An experimental and theoretical investigation. J Mol Struct. 2015;1100:162-73. doi:10.1016/j.molstruc.2015.06.063.

[12] Mak MW, Guo J, Kung SY. PairProSVM: protein subcellular localization based on local pairwise profile alignment and SVM. IEEE/ACM Trans Comput Biol Bioinform. 2008;5:416-22. doi:10.1109/TCBB.2007.70256. PMID:18670044

[13] Nakai K, Kanehisa M. Expert system for predicting protein localization sites in gram-negative bacteria. Prot.: Struct Funct Genet 1991;11:95-110.

[14] Nielsen H, Engelbrecht J, Brunak S, von Heijne G. A neural network method for identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Int J Neural Syst 1997;8:581-99. doi:10.1142/S0129065797000537. PMID:10065837

[15] Chou KC. Prediction of Protein Cellular Attributes Using PseudoAmino Acid Composition. Proteins. 2001;44:246-55.

[16] Chou KC, Cai YD. Prediction of protein subcellular locations by GO-FunD-PseAA predictor. Biochem Bioph Res Co. 2004;320:1236-39. doi:10.1016/j.bbrc.2004.06.073.

[17] Shen HB, Chou KC. PseAAC: A flexible web server for generating various kindsof protein pseudo amino acid composition. Analyt Biochem. 2007;373:386-88. doi:10.1016/j.ab.2007.10.012. PMID:17976365

[18] Bahdanau D, Cho K, Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate. arXiv preprint arXiv. 2014;1049.0473.

[19] Shen HB, Chou KC. Virus-mPLoc: A Fusion Classifier for Viral Protein Subcellular Location Prediction by Incorporating Multiple Sites. J Biomol Struct Dyn. 2012;28:175. doi:10.1080/07391102.2010.10507351.

[20] Xiao X, Wu ZC, Chou KC. A Multi-Label Classifier for Predicting the Subcellular Localization of Gram-Negative Bacterial Proteins with Both Single and Multiple Sites. Plos One. 2011;6:e20592doi:10.1371/journal.pone.0020592. PMID:21698097

[21] Wang L, Wang D, Chen Y, Qiao S, Zhao Y, Cong H. Feature Combination Methods for Prediction of Subcellular Locations of Proteins with Both Single and Multiple Sites. Intelligent Computing Theories and Application. 2016;9771:192-201. doi:10.1007/978-3-319-42291-6_19.

[22] Qu X, Chen Y, Qiao S, Wang D, Zhao Q. Predicting the subcellular localization of proteins with multiple sites based on multiple features fusion. In: Huang, D-S, Han, K, Gromiha, M. (eds.) ICIC 2014. LNCS. vol. 8590. Heidelberg. Springer; pp. 456-65.

[23] Chou KC. Pseudo Amino Acid Composition and its Applications in Bioinformatics, Proteomics and System Biology. Curr Proteomics. 2009;6:262-74. doi:10.2174/157016409789973707.

[24] Han SY, Chen YH, Tang GY. Fault Diagnosis and Fault Tolerant Tracking Control for Discrete-Time Systems with Faults and Delays in Actuator and Measurement. Journal of the Franklin Institute. Online. 2017; (DOI: 10.1016/j.jfranklin.2017.05.027).