

# pLoc\_Deep-mPlant: Predict Subcellular Localization of Plant Proteins by Deep Learning

Yu-Tao Shao<sup>1</sup>, Xin-Xin Liu<sup>1</sup>, Zhe Lu<sup>1</sup>, Kuo-Chen Chou<sup>1,2</sup>

<sup>1</sup>Computer Science, Jingdezhen Ceramic Institute, Jingdezhen, China; <sup>2</sup>Gordon Life Science Institute, Boston, MA, USA

**Correspondence to:** Yu-Tao Shao, 532606318@qq.com; Xin-Xin Liu, 1332250394@qq.com; Zhe Lu, 454170054@qq.com; Kuo-Chen Chou, kcchou@gordonlifescience.org, kcchou38@gmail.com

**Keywords:** Pandemic Coronavirus, Multi-Label System, Plant Proteins, Learning at Deeper Level, Five-Steps Rule, PseAAC

**Received:** May 3, 2020

**Accepted:** May 10, 2020

**Published:** May 13, 2020

Copyright © 2020 by author(s) and Scientific Research Publishing Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

## ABSTRACT

Current coronavirus pandemic has endangered mankind life. The reported cases are increasing exponentially. Information of plant protein subcellular localization can provide useful clues to develop antiviral drugs. To cope with such a catastrophe, a CNN based plant protein subcellular localization predictor called “pLoc\_Deep-mPlant” was developed. The predictor is particularly useful in dealing with the multi-sites systems in which some proteins may simultaneously occur in two or more different organelles that are the current focus of pharmaceutical industry. The global absolute true rate achieved by the new predictor is over 95% and its local accuracy is about 90% - 100%. Both have substantially exceeded the other existing state-of-the-art predictors. To maximize the convenience for most experimental scientists, a user-friendly web-server for the new predictor has been established at [http://www.jci-bioinfo.cn/pLoc\\_Deep-mPlant/](http://www.jci-bioinfo.cn/pLoc_Deep-mPlant/), by which the majority of experimental scientists can easily obtain their desired data without the need to go through the mathematical details.

## 1. INTRODUCTION

Knowledge of the subcellular localization of proteins is crucially important for fulfilling the following two important goals: 1) revealing the intricate pathways that regulate biological processes at the cellular level [1, 2]. 2) selecting the right targets [3] for developing new drugs.

With the avalanche of protein sequences in the post-genomic age, we are challenged to develop computational tools for effectively identifying their subcellular localization purely based on the sequence information.

In 2018, a very powerful predictor, called “pLoc\_bal-mPlant” [4], was developed for predicting the subcellular localization of plant proteins based on their sequences information alone. It has the following remarkable advantages. 1) Most existing protein subcellular location prediction methods were developed based on the single-label system in which it was assumed that each constituent protein had one, and only one, subcellular location (see, e.g., [5-7] and a long list of references cited in a review papers [8]). With more experimental data uncovered, however, the localization of proteins in a cell is actually a multi-label system, where some proteins may simultaneously occur in two or more different location sites. This kind of multiplex proteins often bears some exceptional functions worthy of our special notice [2]. And the pLoc\_bal-mPlant predictor [4] can cover this kind of important information missed by most other methods since it was established based on the multi-label benchmark dataset and theory. 2) Although there are a few methods (see, e.g., [9, 10]) that can be used to deal with multi-label subcellular localization for proteins, the prediction quality achieved by pLoc\_bal-mPlant [4] is overwhelmingly higher, particularly in the absolute true rate. 3) Although the pLoc\_bal-mPlant predictor [4] has the aforementioned merits, it has not been trained at a deeper level yet [11-14].

The present study was initiated in an attempt to address this problem. As done in pLoc\_bal-mPlant [4] as well as many other recent publications in developing new prediction methods (see, e.g., [15, 16]), the guidelines of the 5-step rule [17] are followed. They are about the detailed procedures for 1) benchmark dataset, 2) sample formulation, 3) operation engine or algorithm, 4) cross-validation, and 5) web-server. But here our attentions are focused on the procedures that significantly differ from those in developing the predictor pLoc\_bal-mPlant [4].

## 2. MATERIALS AND METHODS

### 2.1. Benchmark Dataset

The benchmark dataset used in this study is exactly the same as that in pLoc\_bal-mPlant [4]; *i.e.*,

$$\mathbb{S} = \mathbb{S}_1 \cup \mathbb{S}_2 \cup \dots \cup \mathbb{S}_u \cup \dots \cup \mathbb{S}_{11} \cup \mathbb{S}_{12} \quad (1)$$

where  $\mathbb{S}_1$  only contains the plant protein samples from the “Cell membrane” location (cf. **Table 1**),  $\mathbb{S}_2$  only contains those from the “Cell wall” location, and so forth;  $\cup$  denotes the symbol for “union” in the set theory [8]. For readers’ convenience, the detailed sequences of these protein samples and their accession numbers (or ID codes) are given in [Supporting Information S1](#) that is also available at [http://www.jci-bioinfo.cn/pLoc\\_bal-mPlant/Supp1.pdf](http://www.jci-bioinfo.cn/pLoc_bal-mPlant/Supp1.pdf) in which none of proteins included has  $\geq 25\%$  sequence identity to any other in the same subset (subcellular location).

### 2.2. Proteins Sample Formulation

Now let us consider the 2<sup>nd</sup> step of the 5-step rule [17]; *i.e.*, how to formulate the biological sequence samples with an effective mathematical expression that can truly reflect their essential correlation with the target concerned. Given a protein sequence  $\mathbf{P}$ , its most straightforward expression is

$$\mathbf{P} = R_1 R_2 R_3 R_4 R_5 R_6 R_7 \dots R_L \quad (2)$$

where  $L$  denotes the protein’s length or the number of its constituent amino acid residues,  $R_1$  is the 1<sup>st</sup> residue,  $R_2$  the 2<sup>nd</sup> residue,  $R_3$  the 3<sup>rd</sup> residue, and so forth. Since all the existing machine-learning algorithms can only handle vectors as elaborated in [3], one has to convert a protein sample from its sequential expression (Equation (2)) to a vector. But a vector defined in a discrete model might completely miss all the sequence-order or pattern information. To deal with this problem, the Pseudo Amino Acid Composition [18] or PseAAC [19] was proposed. Ever since then, the concept of “Pseudo Amino Acid Composition” has been widely used in nearly all the areas of computational proteomics with the aim to grasp various different sequence patterns that are essential to the targets investigated (see, e.g., [20, 21]) as well as a long list of references cited in [22]). Because it has been widely and increasingly used, recently three powerful open access soft-wares, called “PseAAC-Builder” [23], “propy” [24], and “PseAAC-General”

[25], were established: the former two are for generating various modes of special PseAAC [26]; while the 3<sup>rd</sup> one for those of general PseAAC [17], including not only all the special modes of feature vectors for proteins but also the higher level feature vectors such as “Functional Domain” mode, “Gene Ontology” mode, and “Sequential Evolution” or “PSSM” mode. Encouraged by the successes of using PseAAC to deal with protein/peptide sequences, its idea and approach were extended to PseKNC (Pseudo K-tuple Nucleotide Composition) to generate various feature vectors for DNA/RNA sequences [27] that have proved very successful as well (see, e.g., [28, 29]).

According to the concept of general PseAAC [17], any protein sequence can be formulated as a PseAAC vector given by

$$\mathbf{P} = [\Psi_1 \ \Psi_2 \ \dots \ \Psi_u \ \dots \ \Psi_\Omega]^T \quad (3)$$

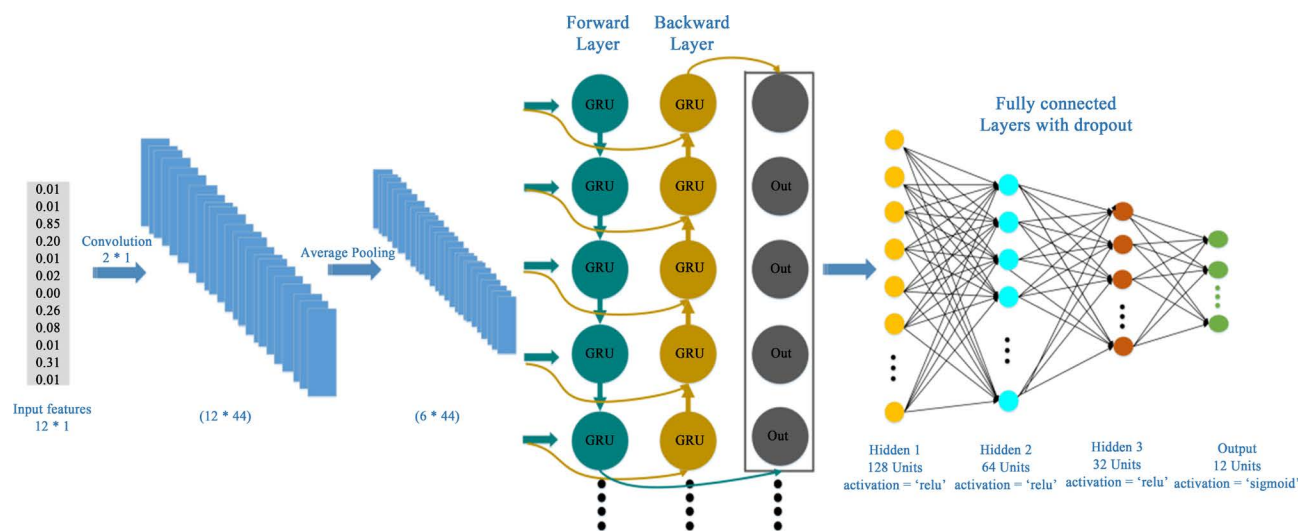
where  $\mathbf{T}$  is a transpose operator, while the integer  $\Omega$  is a parameter and its value as well as the components  $\Psi_u$  ( $u = 1, 2, \dots, \Omega$ ) will depend on how to extract the desired information from the amino acid sequence of  $\mathbf{P}$ , as elaborated in [4]. Thus, by following exactly the same procedures as described in the Section 2.2 of [4], each of the protein samples in the benchmark dataset can be uniquely defined as a 12-D numerical vector as given in Supporting Information S2, which can also be directly downloaded at [http://www.jci-bioinfo.cn/pLoc\\_bal-mPlant/Supp2.pdf](http://www.jci-bioinfo.cn/pLoc_bal-mPlant/Supp2.pdf).

### 2.3. Installing Deep-Learning for Three Deeper Levels

In this study, we use the CNN (Convolutional Neural Network) model to predict the subcellular localization of plant proteins, as illustrated in Figure 1.

The CNN model consists of input layer, convolutional layer, average-pooling layer and fully connected layer. The input layer represents each plant protein with 6 features. The second layer is convolutional layer which extract dependency relationship between features subsequence of plant proteins. The filter stride is set to one. The activation function is set as “relu”. The average-pooling layer down-samples the features and compute the average values of the features. The fully connected layer consists of 2 hidden layers. Finally, the output of connected layer was concatenated into output layer with sigmoid activation function. The label of plant protein was decided by the threshold  $\theta$ . If the output is greater than 0.5, the outcome was true; otherwise, false.

The other parameters of CNN model are as follows. 1) The algorithm of Adam was used to train the model and the loss function is set to binary cross-entropy. 2) The activation function of full connected



**Figure 1.** An illustration to show a dense neural network with 3 fully connected layers. Adapted from [11] with permission.

layer and convolutional layer is ReLU [30], and the activation function of output layer is sigmoid. 3) Convolutional Layer used the filter size 2 \* 1 to extract features of plant proteins. 4) The batch size is 26. 5) The model is trained for 120 epochs. 6) The metrics is set as ‘accuracy’.

The new predictor developed via the above procedures is called “pLoc\_Deep-mPlant”, where “pLoc\_Deep” stands for “predict subcellular localization by deep learning”, and “mPlant” for “multi-label plant proteins”.

### 3. RESULTS AND DISCUSSION

According to the 5-step rules [17], one of the important procedures in developing a new predictor is how to properly evaluate its anticipated accuracy. To deal with that, two issues need to be considered. 1) What metrics should be used to quantitatively reflect the predictor’s quality? 2) What test method should be applied to score the metrics?

#### 3.1. A Set of Five Metrics for Multi-Label Systems

Different from the metrics used to measure the prediction quality of single-label systems, the metrics for the multi-label systems are much more complicated [31]. To make them more intuitive and easier to understand for most experimental scientists, here we use the following intuitive Chou’s five metrics [32] or the “global metrics” that have recently been widely used for studying various multi-label systems (see, e.g., [33, 34]). For the current study, the set of global metrics can be formulated as:

$$\left\{ \begin{array}{l} \text{Aiming } \uparrow = \frac{1}{N^q} \sum_{k=1}^{N^q} \left( \frac{\|\mathbb{L}_k \cap \mathbb{L}_k^*\|}{\|\mathbb{L}_k^*\|} \right), [0,1] \\ \text{Coverage } \uparrow = \frac{1}{N^q} \sum_{k=1}^{N^q} \left( \frac{\|\mathbb{L}_k \cap \mathbb{L}_k^*\|}{\|\mathbb{L}_k\|} \right), [0,1] \\ \text{Accuracy } \uparrow = \frac{1}{N^q} \sum_{k=1}^{N^q} \left( \frac{\|\mathbb{L}_k \cap \mathbb{L}_k^*\|}{\|\mathbb{L}_k \cup \mathbb{L}_k^*\|} \right), [0,1] \\ \text{Absolute true } \uparrow = \frac{1}{N^q} \sum_{k=1}^{N^q} \Delta(\mathbb{L}_k, \mathbb{L}_k^*), [0,1] \\ \text{Absolute false } \downarrow = \frac{1}{N^q} \sum_{k=1}^{N^q} \left( \frac{\|\mathbb{L}_k \cup \mathbb{L}_k^*\| - \|\mathbb{L}_k \cap \mathbb{L}_k^*\|}{M} \right), [1,0] \end{array} \right. \quad (4)$$

where  $N^q$  is the total number of query proteins or tested proteins,  $M$  is the total number of different labels for the investigated system (for the current study it is  $L_{\text{cell}} = 12$ ),  $\|\cdot\|$  means the operator acting on the set therein to count the number of its elements,  $\cup$  means the symbol for the “union” in the set theory,  $\cap$  denotes the symbol for the “intersection”,  $\mathbb{L}_k$  denotes the subset that contains all the labels observed by experiments for the  $k$ -th tested sample,  $\mathbb{L}_k^*$  represents the subset that contains all the labels predicted for the  $k$ -th sample, and

$$\Delta(\mathbb{L}_k, \mathbb{L}_k^*) = \begin{cases} 1, & \text{if all the labels in } \mathbb{L}_k^* \text{ are identical to those in } \mathbb{L}_k \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

In Equation (4), the first four metrics with an upper arrow  $\uparrow$  are called positive metrics, meaning that the larger the rate is the better the prediction quality will be; the 5<sup>th</sup> metrics with a down arrow  $\downarrow$  is called negative metrics, implying just the opposite meaning.

From Equation (4) we can see the following: 1) the “Aiming” defined by the 1<sup>st</sup> sub-equation is for

checking the rate or percentage of the correctly predicted labels over the practically predicted labels; 2) the “Coverage” defined in the 2<sup>nd</sup> sub-equation is for checking the rate of the correctly predicted labels over the actual labels in the system concerned; 3) the “Accuracy” in the 3<sup>rd</sup> sub-equation is for checking the average ratio of correctly predicted labels over the total labels including correctly and incorrectly predicted labels as well as those real labels but are missed in the prediction; 4) the “Absolute true” in the 4<sup>th</sup> sub-equation is for checking the ratio of the perfectly or completely correct prediction events over the total prediction events; 5) the “Absolute false” in the 5<sup>th</sup> sub-equation is for checking the ratio of the completely wrong prediction over the total prediction events.

### 3.2. Comparison with the State-of-the-Art Predictor

Listed in **Table 1** are the rates achieved by the current pLoc\_Deep-mPlant predictor via the cross validations on the same experiment-confirmed dataset as used in [4]. For facilitating comparison, listed there are also the corresponding results obtained by the pLoc\_bal-mPlant [4], the existing most powerful predictor for identifying the subcellular localization of plant proteins with both single and multiple location sites. As shown in **Table 1**, the newly proposed predictor pLoc\_Deep-mPlant is remarkably superior to the existing state-of-the-art predictor pLoc\_bal-mPlant in all the five metrics. Particularly, it can be seen from the table that the absolute true rate achieved by the new predictor is over 95%, which is far beyond the reach of any other existing methods [35-40]. This is because it is extremely difficult to enhance the absolute true rate of a prediction method for a multi-label system as clearly elucidated in [4]. Actually, to avoid embarrassment, many investigators even chose not to mention the metrics of absolute true rate in dealing with multi-label systems (see, e.g., [41-47]).

Moreover, to in-depth examine the prediction quality of the new predictor for the proteins in each of the subcellular locations concerned (cf. **Table 2**), we used the “local metrics” [31] or a set of four intuitive metrics that were derived in [48] based on the Chou’s symbols introduced for studying protein signal peptides [49] and that have ever since been widely concurred or justified (see, e.g., [50-53]). For the current study, the set of local metrics can be formulated as:

$$\left\{ \begin{array}{l} \text{Sn}(i) = 1 - \frac{N_{-}^{+}(i)}{N^{+}(i)} \quad 0 \leq \text{Sn} \leq 1 \\ \text{Sp}(i) = 1 - \frac{N_{+}^{-}(i)}{N^{-}(i)} \quad 0 \leq \text{Sp} \leq 1 \\ \text{Acc}(i) = 1 - \frac{N_{-}^{+}(i) + N_{+}^{-}(i)}{N^{+}(i) + N^{-}(i)} \quad 0 \leq \text{Acc} \leq 1 \\ \text{MCC}(i) = \frac{1 - \left( \frac{N_{-}^{+}(i)}{N^{+}(i)} + \frac{N_{+}^{-}(i)}{N^{-}(i)} \right)}{\sqrt{\left( 1 + \frac{N_{+}^{-}(i) - N_{-}^{+}(i)}{N^{+}(i)} \right) \left( 1 + \frac{N_{-}^{+}(i) - N_{+}^{-}(i)}{N^{-}(i)} \right)}} \quad -1 \leq \text{MCC} \leq 1 \end{array} \right. \quad (6)$$

$(i = 1, 2, \dots, 12)$

where Sn, Sp, Acc, and MCC represent the sensitivity, specificity, accuracy, and Mathew’s correlation coefficient, respectively, and  $i$  denotes the  $i$ -th subcellular location (or subset) in the benchmark dataset.  $N^{+}(i)$  is the total number of the samples investigated in the  $i$ -th subset, whereas  $N_{-}^{+}(i)$  is the number of the samples in  $N^{+}(i)$  that are incorrectly predicted to be of other locations;  $N^{-}(i)$  is the total number of samples in any locations but not the  $i$ -th location, whereas  $N_{+}^{-}(i)$  is the number of the samples in  $N^{-}(i)$  that are incorrectly predicted to be of the  $i$ -th location.

Listed in **Table 2** are the results achieved by pLoc\_Deep-mPlant for the plant proteins in each of 12

**Table 1.** Comparison with the state-of-the-art method in predicting plant protein subcellular localization<sup>a</sup>.

Predictor	Aiming (↑) <sup>a</sup>	Coverage (↑) <sup>a</sup>	Accuracy (↑) <sup>a</sup>	Absolute true (↑) <sup>a</sup>	Absolute false (↓) <sup>a</sup>
pLoc_bal-mPlant <sup>b</sup>	93.90%	94.79%	93.29%	92.02%	93.90%
pLoc_Deep-mPlant <sup>c</sup>	96.57%	97.24%	96.40%	95.38%	96.57%

<sup>a</sup>See Equation (4) for the definition of the metrics. <sup>b</sup>See [4], where the reported metrics rates were obtained by the jackknife test on the benchmark dataset of [Supporting Information S1](#) that contains experiment-confirmed proteins only. <sup>c</sup>The proposed predictor; to assure that the test was performed on exactly the same experimental data as reported in [4] for pLoc\_bal-mPlant.

**Table 2.** Performance of pLoc\_Deep-mPlant for each of the 12 subcellular locations.

<i>i</i>	Location <sup>a</sup>	Sn( <i>i</i> ) <sup>b</sup>	Sp( <i>i</i> ) <sup>b</sup>	Acc( <i>i</i> ) <sup>b</sup>	MCC( <i>i</i> ) <sup>b</sup>
1	Acrosome	1.0000	0.9997	0.9997	0.9353
2	Cell membrane	0.9986	0.9907	0.9914	0.9505
3	Cell wall	0.9796	0.9990	0.9988	0.9158
4	Centrosome	1.0000	0.9961	0.9961	0.8712
5	Chloroplast	0.9948	0.9988	0.9986	0.9851
6	Cyanelle	1.0000	1.0000	1.0000	1.0000
7	Cytoplasm	0.8477	0.9559	0.9254	0.8137
8	Cytoskeleton	1.0000	0.9959	0.9960	0.9024
9	Endoplasmic reticulum	0.9978	0.9970	0.9970	0.9741
10	Endosome	1.0000	0.9992	0.9992	0.9336
11	Extracell	0.9962	0.9955	0.9956	0.9815
12	Golgi apparatus	0.9961	0.9963	0.9963	0.9452

<sup>a</sup>See [Table 1](#) and the relevant context for further explanation. <sup>b</sup>See Equation (6) for the metrics definition.

subcellular locations. As we can see from the table, nearly all the success rates achieved by the new predictor for the plant proteins in each of the 12 subcellular locations are within the range of 90% - 100%, which is once again far beyond the reach of any of its counterparts.

Meanwhile, as a byproduct, the present paper has also stimulated some very interesting or provoked papers (see, e.g., [54-59]).

### 3.3. Web Server and User Guide

As pointed out in [60], user-friendly and publicly accessible web-servers represent the future direction for developing practically more useful predictors. Actually, user-friendly web-servers will significantly enhance the impacts of theoretical work because they can attract the broad experimental scientists [22]. In

view of this, the web-server of the current pLoc\_Deep-mPlant predictor has also been established at [http://www.jci-bioinfo.cn/pLoc\\_Deep-mPlant/](http://www.jci-bioinfo.cn/pLoc_Deep-mPlant/), by which users can easily get their desired data without the need to go thru the mathematical details.

#### 4. CONCLUSION

It is anticipated that the pLoc\_Deep-mPlant predictor holds very high potential to become a useful high throughput tool in identifying the subcellular localization of plant proteins, particularly for finding multi-target drugs that is currently a very hot trend in drug development. Most important is that the predictor will become a very useful tool for fighting against the coronavirus to save the mankind in this planet.

#### ACKNOWLEDGEMENTS

This work was supported by the grants from the National Natural Science Foundation of China (No. 31560316, 61261027, 61262038, 61202313 and 31260273), the Province National Natural Science Foundation of JiangXi (No. 20132BAB201053), the Jiangxi Provincial Foreign Scientific and Technological Cooperation Project (No. 20120BDH80023), the Department of Education of JiangXi Province (GJJ160866).

#### CONFLICTS OF INTEREST

The authors declare no conflicts of interest regarding the publication of this paper.

#### REFERENCES

1. Ehrlich, J.S., Hansen, M.D. and Nelson, W.J. (2002) Spatio-Temporal Regulation of Rac1 Localization and Lamellipodia Dynamics during Epithelial Cell-Cell Adhesion. *Developmental Cell*, **3**, 259-270. [https://doi.org/10.1016/S1534-5807\(02\)00216-2](https://doi.org/10.1016/S1534-5807(02)00216-2)
2. Glory, E. and Murphy, R.F. (2007) Automated Subcellular Location Determination and High-Throughput Microscopy. *Developmental Cell*, **12**, 7-16. <https://doi.org/10.1016/j.devcel.2006.12.007>
3. Chou, K.C. (2015) Impacts of Bioinformatics to Medicinal Chemistry. *Medicinal Chemistry*, **11**, 218-234. <https://doi.org/10.2174/1573406411666141229162834>
4. Cheng, X., Xiao, X. and Chou, K.C. (2018) pLocbal-mPlant: Predict Subcellular Localization of Plant Proteins by General PseAAC and Balancing Training Dataset. *Current Pharmaceutical Design*, **24**, 4013-4022. <https://doi.org/10.2174/1381612824666181119145030>
5. Nakai, K. and Kanehisa, M. (1992) A Knowledge Base for Predicting Protein Localization Sites in Eukaryotic Cells. *Genomics*, **14**, 897-911. [https://doi.org/10.1016/S0888-7543\(05\)80111-9](https://doi.org/10.1016/S0888-7543(05)80111-9)
6. Cedano, J., Aloy, P., Perez-Pons, J.A. and Querol, E. (1997) Relation between Amino Acid Composition and Cellular Location of Proteins. *Journal of Molecular Biology*, **266**, 594-600. <https://doi.org/10.1006/jmbi.1996.0804>
7. Reinhardt, A. and Hubbard, T. (1998) Using Neural Networks for Prediction of the Subcellular Location of Proteins. *Nucleic Acids Research*, **26**, 2230-2236. <https://doi.org/10.1093/nar/26.9.2230>
8. Chou, K.C. and Shen, H.B. (2007) Recent Progresses in Protein Subcellular Location Prediction. *Analytical Biochemistry*, **370**, 1-16. <https://doi.org/10.1016/j.ab.2007.07.006>
9. Chou, K.C., Wu, Z.C. and Xiao, X. (2011) iLoc-Euk: A Multi-Label Classifier for Predicting the Subcellular Localization of Singleplex and Multiplex Eukaryotic Proteins. *PLoS ONE*, **6**, e18258. <https://doi.org/10.1371/journal.pone.0018258>
10. Mandal, M., Mukhopadhyay, A. and Maulik, U. (2015) Prediction of Protein Subcellular Localization by Incorporating

porating Multiobjective PSO-Based Feature Subset Selection into the General form of Chou's PseAAC. *Medical & Biological Engineering & Computing*, **53**, 331-344. <https://doi.org/10.1007/s11517-014-1238-7>

11. Maxwell, A., Li, R., Yang, B., Weng, H., Ou, A., Hong, H., Zhou, Z., Gong, P. and Zhang, C. (2017) Deep Learning Architectures for Multi-Label Classification of Intelligent Health Risk Prediction. *BMC Bioinformatics*, **18**, 523. <https://doi.org/10.1186/s12859-017-1898-z>
12. Khan, S., Khan, M., Iqbal, N., Hussain, T., Khan, S.A. and Chou, K.C. (2019) A Two-Level Computation Model Based on Deep Learning Algorithm for Identification of piRNA and Their Functions via Chou's 5-Steps Rule. *International Journal of Peptide Research and Therapeutics*, **26**, 795-809. <https://doi.org/10.1007/s10989-019-09887-3>
13. Khan, Z.U., Ali, F., Khan, I.A., Hussain, Y. and Pi, D. (2019) iRSpot-SPI: Deep Learning-Based Recombination Spots Prediction by Incorporating Secondary Sequence Information Coupled with Physio-Chemical Properties via Chou's 5-Step Rule and Pseudo Components. *Chemometrics and Intelligent Laboratory Systems (CHEMOLAB)*, **189**, 169-180. <https://doi.org/10.1016/j.chemolab.2019.05.003>
14. Nazari, I., Tahir, M., Tayari, H. and Chong, K.T. (2019) iN6-Methyl (5-Step): Identifying RNA N6-Methyladenosine Sites Using Deep Learning Mode via Chou's 5-Step Rules and Chou's General PseKNC. *Chemometrics and Intelligent Laboratory Systems (CHEMOLAB)*, **193**, Article ID: 103811. <https://doi.org/10.1016/j.chemolab.2019.103811>
15. Hussain, W., Khan, Y.D., Rasool, N., Khan, S.A. and Chou, K.C. (2019) SPrenylC-PseAAC: A Sequence-Based Model Developed via Chou's 5-Steps Rule and General PseAAC for Identifying S-Prenylation Sites in Proteins. *Journal of Theoretical Biology*, **468**, 1-11. <https://doi.org/10.1016/j.jtbi.2019.02.007>
16. Du, L., Meng, Q., Jiang, H. and Li, Y. (2020) Using Evolutionary Information and Multi-Label Linear Discriminant Analysis to Predict the Subcellular Location of Multi-Site Bacterial Proteins via Chou's 5-Steps Rule. *IEEE Access*, **8**, 56452-56461. <https://doi.org/10.1109/ACCESS.2020.2982160>
17. Chou, K.C. (2011) Some Remarks on Protein Attribute Prediction and Pseudo Amino Acid Composition (50th Anniversary Year Review, 5-Steps Rule). *Journal of Theoretical Biology*, **273**, 236-247. <https://doi.org/10.1016/j.jtbi.2010.12.024>
18. Chou, K.C. (2001) Prediction of Protein Cellular Attributes Using Pseudo Amino Acid Composition. *PROTEINS: Structure, Function, and Genetics*, **43**, 246-255. <https://doi.org/10.1002/prot.1035>
19. Chou, K.C. (2005) Using Amphiphilic Pseudo Amino Acid Composition to Predict Enzyme Subfamily Classes. *Bioinformatics*, **21**, 10-19. <https://doi.org/10.1093/bioinformatics/bth466>
20. Zhou, X.B., Chen, C., Li, Z.C. and Zou, X.Y. (2007) Using Chou's Amphiphilic Pseudo Amino Acid Composition and Support Vector Machine for Prediction of Enzyme Subfamily Classes. *Journal of Theoretical Biology*, **248**, 546-551. <https://doi.org/10.1016/j.jtbi.2007.06.001>
21. Akbar, S., Rahman, A.U. and Hayat, M. (2020) cACP: Classifying Anticancer Peptides Using Discriminative Intelligent Model via Chou's 5-Step Rules and General Pseudo Components. *Chemometrics and Intelligent Laboratory (CHEMOLAB)*, **196**, Article ID: 103912. <https://doi.org/10.1016/j.chemolab.2019.103912>
22. Chou, K.C. (2017) An Unprecedented Revolution in Medicinal Chemistry Driven by the Progress of Biological Science. *Current Topics in Medicinal Chemistry*, **17**, 2337-2358. <https://doi.org/10.2174/1568026617666170414145508>
23. Du, P., Wang, X., Xu, C. and Gao, Y. (2012) PseAAC-Builder: A Cross-Platform Stand-Alone Program for Generating Various Special Chou's Pseudo Amino Acid Compositions. *Analytical Biochemistry*, **425**, 117-119. <https://doi.org/10.1016/j.ab.2012.03.015>
24. Cao, D.S., Xu, Q.S. and Liang, Y.Z. (2013) Propy: A Tool to Generate Various Modes of Chou's PseAAC. *Bioinformatics*, **29**, 960-962. <https://doi.org/10.1093/bioinformatics/btt072>



25. Du, P., Gu, S. and Jiao, Y. (2014) PseAAC-General: Fast Building Various Modes of General form of Chou's Pseudo Amino Acid Composition for Large-Scale Protein Datasets. *International Journal of Molecular Sciences*, **15**, 3495-3506. <https://doi.org/10.3390/ijms15033495>
26. Chou, K.C. (2009) Pseudo Amino Acid Composition and Its Applications in Bioinformatics, Proteomics and System Biology. *Current Proteomics*, **6**, 262-274. <https://doi.org/10.2174/157016409789973707>
27. Chen, W., Lei, T.Y., Jin, D.C., Lin, H. and Chou, K.C. (2014) PseKNC: A Flexible Web-Server for Generating Pseudo K-Tuple Nucleotide Composition. *Analytical Biochemistry*, **456**, 53-60. <https://doi.org/10.1016/j.ab.2014.04.001>
28. Chen, W., Lin, H. and Chou, K.C. (2015) Pseudo Nucleotide Composition or PseKNC: An Effective Formulation for Analyzing Genomic Sequences. *Molecular BioSystems*, **11**, 2620-2634. <https://doi.org/10.1039/C5MB00155B>
29. Liu, B., Yang, F. and Chou, K.C. (2017) 2L-piRNA: A Two-Layer Ensemble Classifier for Identifying Piwi-Interacting RNAs and Their Function. *Molecular Therapy-Nucleic Acids*, **7**, 267-277. <https://doi.org/10.1016/j.omtn.2017.04.008>
30. Glorot, X., Bordes, A. and Bengio, Y. (2011) Deep Sparse Rectifier Neural Networks. *14th International Conference on Artificial Intelligence and Statistics*, Fort Lauderdale, FL, 11-13 April 2011, 315-323.
31. Chou, K.C. (2019) Two Kinds of Metrics for Computational Biology. *Genomics*. <https://www.sciencedirect.com/science/article/pii/S0888754319304604?via%3Dihub>
32. Chou, K.C. (2013) Some Remarks on Predicting Multi-Label Attributes in Molecular Biosystems. *Molecular Biosystems*, **9**, 1092-1100. <https://doi.org/10.1039/c3mb25555g>
33. Song, J., Wang, Y., Li, F., Akutsu, T., Rawlings, N.D., Webb, G.I. and Chou, K.C. (2018) iProt-Sub: A Comprehensive Package for Accurately Mapping and Predicting Protease-Specific Substrates and Cleavage Sites. *Brief in Bioinform*, **20**, 638-658. <https://doi.org/10.1093/bib/bby028>
34. Zhang, M., Li, F., Marquez-Lago, T.T., Leier, A., Fan, C., Kwok, C.K., Chou, K.C., Song, J. and Jia, C. (2019) MULTiPly: A Novel Multi-Layer Predictor for Discovering General and Specific Types of Promoters. *Bioinformatics*, **35**, 2957-2965. <https://doi.org/10.1093/bioinformatics/btz016>
35. Shen, H.B. and Chou K.C. (2007) Hum-mPLoc: An Ensemble Classifier for Large-Scale Human Protein Subcellular Location Prediction by Incorporating Samples with Multiple Sites. *Biochemical and Biophysical Research Communications (BBRC)*, **355**, 1006-1011. <https://doi.org/10.1016/j.bbrc.2007.02.071>
36. Chou, K.C. and Shen, H.B. (2008) Cell-PLoc: A Package of Web Servers for Predicting Subcellular Localization of Proteins in Various Organisms. *Nature Protocols*, **3**, 153-162. <https://doi.org/10.1038/nprot.2007.494>
37. Shen, H.B. and Chou, K.C. (2009) A Top-Down Approach to Enhance the Power of Predicting Human Protein Subcellular Localization: Hum-mPLoc 2.0. *Analytical Biochemistry*, **394**, 269-274. <https://doi.org/10.1016/j.ab.2009.07.046>
38. Chou, K.C. and Shen, H.B. (2010) Cell-PLoc 2.0: An Improved Package of Web-Servers for Predicting Subcellular Localization of Proteins in Various Organisms. *Natural Science*, **2**, 1090-1103. <https://doi.org/10.4236/ns.2010.210136>
39. Chou, K.C., Wu, Z.C. and Xiao, X. (2012) iLoc-Hum: Using Accumulation-Label Scale to Predict Subcellular Locations of Human Proteins with Both Single and Multiple Sites. *Molecular Biosystems*, **8**, 629-641. <https://doi.org/10.1039/C1MB05420A>
40. Cheng, X., Xiao, X. and Chou, K.C. (2018) pLoc-mHum: Predict Subcellular Localization of Multi-Location Human Proteins via General PseAAC to Winnow out the Crucial GO Information. *Bioinformatics*, **34**, 1448-1456. <https://doi.org/10.1093/bioinformatics/btx711>

41. Cao, J.Z., Liu, W.Q. and Gu, H. (2012) Predicting Viral Protein Subcellular Localization with Chou's Pseudo Amino Acid Composition and Imbalance-Weighted Multi-Label K-Nearest Neighbor Algorithm. *Protein and Peptide Letters*, **19**, 1163-1169. <https://doi.org/10.2174/092986612803216999>
42. He, J., Gu, H and Liu, W. (2012) Imbalanced Multi-Modal Multi-Label Learning for Subcellular Localization Prediction of Human Proteins with Both Single and Multiple Sites. *PLoS ONE*, **7**, e37155. <https://doi.org/10.1371/journal.pone.0037155>
43. Li, L.Q., Zhang, Y., Zou, L.Y., Zhou, Y. and Zheng, X.Q. (2012) Prediction of Protein Subcellular Multi-Localization Based on the General form of Chou's Pseudo Amino Acid Composition. *Protein & Peptide Letters*, **19**, 375-387. <https://doi.org/10.2174/092986612799789369>
44. Mei, S. (2012) Predicting Plant Protein Subcellular Multi-Localization by Chou's PseAAC Formulation Based Multi-Label Homolog Knowledge Transfer Learning. *Journal of Theoretical Biology*, **310**, 80-87. <https://doi.org/10.1016/j.jtbi.2012.06.028>
45. Wang, X. and Li, G.Z. (2012) A Multi-Label Predictor for Identifying the Subcellular Locations of Singleplex and Multiplex Eukaryotic Proteins. *PLoS ONE*, **7**, e36317. <https://doi.org/10.1371/journal.pone.0036317>
46. Huang, C. and Yuan, J. (2013) Using Radial Basis Function on the General Form of Chou's Pseudo Amino Acid Composition and PSSM to Predict Subcellular Locations of Proteins with Both Single and Multiple Sites. *Bio-systems*, **113**, 50-57. <https://doi.org/10.1016/j.biosystems.2013.04.005>
47. Pacharawongsakda, E. and Theeramunkong, T. (2013) Predict Subcellular Locations of Singleplex and Multiplex Proteins by Semi-Supervised Learning and Dimension-Reducing General Mode of Chou's PseAAC. *IEEE Transactions on Nanobioscience*, **12**, 311-320. <https://doi.org/10.1109/TNB.2013.2272014>
48. Chen, W., Feng, P.M., Lin, H. and Chou, K.C. (2013) iRSpot-PseDNC: Identify Recombination Spots with Pseudo Dinucleotide Composition. *Nucleic Acids Research*, **41**, e68. <https://doi.org/10.1093/nar/gks1450>
49. Chou, K.C. (2001) Using Subsite Coupling to Predict Signal Peptides. *Protein Engineering*, **14**, 75-79. <https://doi.org/10.1093/protein/14.2.75>
50. Qiu, W.R., Xiao, X. and Chou, K.C. (2014) iRSpot-TNCPseAAC: Identify Recombination Spots with Trinucleotide Composition and Pseudo Amino Acid Components. *International Journal of Molecular Sciences (IJMS)*, **15**, 1746-1766. <https://doi.org/10.3390/ijms15021746>
51. Xu, R., Zhou, J., Liu, B., He, Y.A., Zou, Q., Wang, X. and Chou, K.C. (2015) Identification of DNA-Binding Proteins by Incorporating Evolutionary Information into Pseudo Amino Acid Composition via the Top-n-Gram Approach. *Journal of Biomolecular Structure & Dynamics (JBSD)*, **33**, 1720-1730. <https://doi.org/10.1080/07391102.2014.968624>
52. Jia, J., Zhang, L., Liu, Z., Xiao, X. and Chou, K.C. (2016) pSumo-CD: Predicting Sumoylation Sites in Proteins with Covariance Discriminant Algorithm by Incorporating Sequence-Coupled Effects into General PseAAC. *Bioinformatics*, **32**, 3133-3141. <https://doi.org/10.1093/bioinformatics/btw387>
53. Liu, B., Wang, S., Long, R. and Chou, K.C. (2017) iRSpot-EL: Identify Recombination Spots with an Ensemble Learning Approach. *Bioinformatics*, **33**, 35-41. <https://doi.org/10.1093/bioinformatics/btw539>
54. Chou, K.C. (2020) The Development of Gordon Life Science Institute: Its Driving Force and Accomplishments. *Natural Science*, **12**, 202-217. <https://doi.org/10.4236/ns.2020.124018>
55. Chou, K.C. (2020) The Most Important Ethical Concerns in Science. *Natural Science*, **12**, 35-36. <https://doi.org/10.4236/ns.2020.122005>
56. Chou, K.C. (2020) Other Mountain Stones Can Attack Jade: The 5-Steps Rule. *Natural Science*, **12**, 59-64. <https://doi.org/10.4236/ns.2020.123011>
57. Chou, K.C. (2020) The Problem of Elsevier Series Journals Online Submission by Using Artificial Intelligence.

*Natural Science*, **12**, 37-38. <https://doi.org/10.4236/ns.2020.122006>

58. Chou, K.C. (2020) Proposing 5-Steps Rule Is a Notable Milestone for Studying Molecular Biology. *Natural Science*, **12**, 74-79. <https://doi.org/10.4236/ns.2020.123011>
59. Chou, K.C. (2020) Using Similarity Software to Evaluate Scientific Paper Quality Is a Big Mistake. *Natural Science*, **12**, 42-58. <https://doi.org/10.4236/ns.2020.123008>
60. Chou, K.C. and Shen, H.B. (2009) Recent Advances in Developing Web-Servers for Predicting Protein Attributes. *Natural Science*, **1**, 63-92. <https://doi.org/10.4236/ns.2009.12011>