https://www.scirp.org/journal/ns Natural Science, 2020, Vol. 12, (No. 6), pp: 388-399

# pLoc_Deep-mVirus: A CNN Model for Predicting Subcellular Localization of Virus Proteins by Deep Learning

**Yutao Shao[1], Kuo-Chen Chou[1,2]**

[1]Computer Science, Jingdezhen Ceramic Institute, Jingdezhen, China; [2]Gordon Life Science Institute, Boston, MA, USA

**Correspondence to:** Yutao Shao, 532606318@qq.com;
Kuo-Chen Chou, kcchou@gordonlifescience.org, kcchou38@gmail.com

**Keywords:** Coronavirus, Virus Proteins, Multi-Label System, Deep Learning, Five-Steps Rule, PseAAC
**Received:** June 10, 2020 **Accepted:** June 20, 2020 **Published:** June 23, 2020



## ABSTRACT

**The recent worldwide spreading of pneumonia-causing virus, such as Coronavirus, COVID-19, and H1N1, has been endangering the life of human beings all around the world. In order to really understand the biological process within a cell level and provide useful clues to develop antiviral drugs, information of virus protein subcellular localization is vitally important. In view of this, a CNN based virus protein subcellular localization predictor called "pLoc_Deep-mVirus" 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 97% and its local accuracy is over 98%. Both are transcending other existing state-of-the-art predictors significantly. It has not escaped our notice that the deep-learning treatment can be used to deal with many other biological systems as well. 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-mVirus/.**

## 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 2019, a very powerful predictor, called "pLoc_bal-mVirus" [4], was developed for predicting the subcellular localization of virus 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-mVirus 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-mVirus [4] is overwhelmingly higher, particularly in the absolute true rate.

The pLoc_bal-mVirus 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-mVirus [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-mVirus [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-mVirus [4]; *i.e.*,

$$\mathbb{S} = \mathbb{S}_1 \cup \mathbb{S}_2 \cup \mathbb{S}_3 \cup \mathbb{S}_4 \cup \mathbb{S}_5 \cup \mathbb{S}_6 \quad (1)$$

where $\mathbb{S}_1$ only contains the virus protein samples from the "Viral capsid" location (cf. **Table 1**), $\mathbb{S}_2$ only contains those from the "Host cell membrane" location, and so forth; $\cup$ denotes the symbol for "union" in the set theory. For readers' convenience, the detailed sequences of these protein samples and their accession numbers (or ID codes) are given in Supporting Information S1 that are also available at http://www.jci-bioinfo.cn/pLoc_bal-mVirus/Supp1.pdf, in which none of proteins included has ≥25% sequence identity to any other in the same subset (subcellular location). But such a cutoff treatment was not imposed for the protein sequences in the "viral capsid" subset; otherwise it would contain too few protein samples to be of statistical significance as explained in the original paper [18].

**Table 1. Comparison with the state-of-the-art method in predicting virus protein subcellular localizational.**

| Predictor | Aiming (↑)[a] | Coverage (↑)[a] | Accuracy (↑)[a] | Absolute true (↑)[a] | Absolute false (↓)[a] |
|---|---|---|---|---|---|
| pLoc_bal-mVirus[b] | 88.31% | 85.06% | 84.34% | 78.78% | 0.07% |
| pLoc_Deep-mVirus[c] | 99.47% | 99.47% | 98.95% | 97.89% | 0.00% |

[a]See Equation (4) for the definition of the metrics. [b]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. [c]The proposed predictor; to assure that the test was performed on exactly the same experimental data as reported in [4] for pLoc_bal-mVirus.

## 2.2. Proteins Sample Formulation

Now let us consider the 2nd 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 **P**, its most straightforward expression is

$$\mathbf{P} = \mathrm{R}_1\mathrm{R}_2\mathrm{R}_3\mathrm{R}_4\mathrm{R}_5\mathrm{R}_6\mathrm{R}_7\cdots\mathrm{R}_L \tag{2}$$

where $L$ denotes the protein's length or the number of its constituent amino acid residues, $\mathrm{R}_1$ is the 1st residue, $\mathrm{R}_2$ the 2nd residue, $\mathrm{R}_3$ the 3rd 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 [19] or PseAAC [20]. 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., [21-31] as well as a long list of references cited in [32]). Because it has been widely and increasingly used, recently three powerful open access soft-wares, called "PseAAC-Builder" [33], "propy" [34], and "PseAAC-General" [35], were established: the former two are for generating various modes of special PseAAC [36]; while the 3rd 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 [37] that have proved very successful as well (see, e.g., [38, 39].

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

$$\mathbf{P} = \left[\Psi_1\ \Psi_2\cdots\Psi_u\cdots\Psi_\Omega\right]^{\mathbf{T}} \tag{3}$$

where $\mathbf{T}$ is a transpose operator, while the integer $\Omega$ is a parameter and its value as well as the components $\Psi_u\left(u=1,2,\cdots,\Omega\right)$ will depend on how to extract the desired information from the amino acid sequence of **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 6-D numerical vector as given in Supporting Information S2, which can also be directly downloaded at http://www.jci-bioinfo.cn/pLoc_bal-mVirus/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 virus 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 virus protein with 6 features. The second layer is convolutional layer which extract dependency relationship between features subsequence of virus 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 virus 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 layer and convolutional layer is ReLU [40], and the activation function of output layer is sigmoid. 3) Convolutional Layer used the filter size 2 * 1 to extract features of virus proteins. 4) The batch size is 26. 5) The model is trained for 120 epochs. 6) The metrics is set as "accuracy".

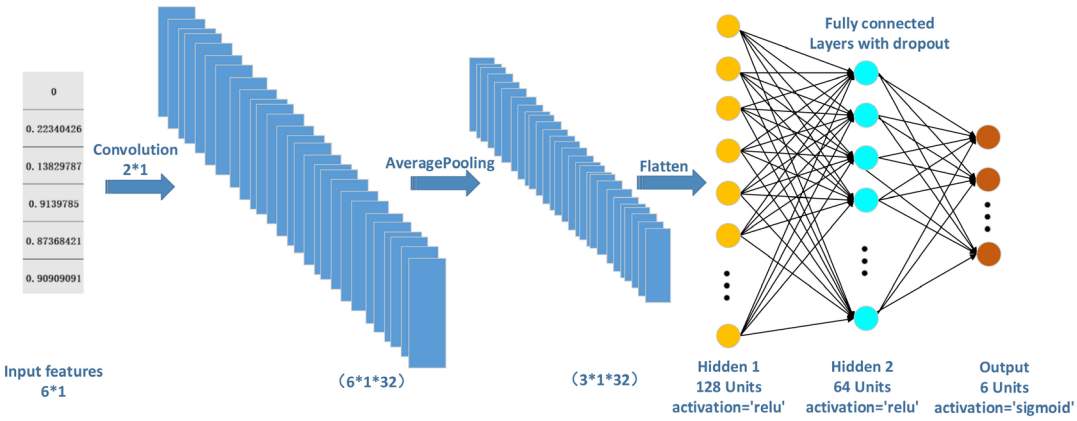


**Figure 1.** **An illustration to show the Architecture of the pLoc-Deep_mVirus model.**

The new predictor developed via the above procedures is called "pLoc_Deep-mVirus", where "pLoc_Deep" stands for "predict subcellular localization by deep learning", and "mVirus" for "multi-label virus 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 [41]. To make them more intuitive and easier to understand for most experimental scientists, here we use the following intuitive Chou's five metrics [42] or the "global metrics" that have recently been widely used for studying various multi-label systems (see, e.g., [43, 44]). For the current study, the set of global metrics can be formulated as:

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

where $N^{\mathrm{q}}$ 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 $L_{\text{cell}} = 6$), $\|\ \|$ 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

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

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 5th 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 1st 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 2nd 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 3rd 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 4th 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 5th 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-mVirus 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-mVirus [4], the existing most powerful predictor for identifying the subcellular localization of virus proteins with both single and multiple location sites. As shown in **Table 1**, the newly proposed predictor pLoc_Deep-mVirus is remarkably superior to the existing state-of-the-art predictor pLoc_bal-mVirus 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 97%, which is far beyond the reach of any other existing methods [45-50]. 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., [51-57]).

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" [41] or a set of four intuitive metrics that were derived in [58] based on the Chou's symbols introduced for studying protein signal peptides [59] and that have ever since been widely concurred or justified (see, e.g., [60-63]). For the current study, the set of local metrics can be formulated as:

**Table 2. Performance of pLoc_Deep-m Virus for each of the 6 subcellular locations.**

| $i$ | Location[a] | Sn($i$)[b] | Sp($i$)[b] | Acc($i$)[b] | MCC($i$)[b] |
|---|---|---|---|---|---|
| 1 | Viral caspid | 0.9909 | 1.0000 | 0.9979 | 0.9941 |
| 2 | Host call membrane | 0.9979 | 1.0000 | 0.9958 | 0.9908 |
| 3 | Host endoplasmic | 0.9852 | 1.0000 | 0.9958 | 0.9897 |
| 4 | Host cytoplasm | 0.9167 | 0.9920 | 0.9728 | 0.9251 |
| 5 | Host nucleu | 0.9286 | 0.9808 | 0.9623 | 0.9154 |
| 6 | Secreted | 0.9818 | 1.0000 | 0.9958 | 0.9882 |

[a]See **Table 1** and the relevant context for further explanation. [b]See Equation (6) for the metrics definition.

$$\begin{cases} \mathrm{Sn}(i)=1-\dfrac{N_{-}^{+}(i)}{N^{+}(i)} & 0\le \mathrm{Sn}\le 1 \\ \mathrm{Sp}(i)=1-\dfrac{N_{+}^{-}(i)}{N^{-}(i)} & 0\le \mathrm{Sp}\le 1 \\ \mathrm{Acc}(i)=1-\dfrac{N_{-}^{+}(i)+N_{+}^{-}(i)}{N^{+}(i)+N^{-}(i)} & 0\le \mathrm{Acc}\le 1 \\ \mathrm{MCC}(i)=\dfrac{1-\left(\dfrac{N_{-}^{+}(i)}{N^{+}(i)}+\dfrac{N_{+}^{-}(i)}{N^{-}(i)}\right)}{\sqrt{\left(1+\dfrac{N_{+}^{-}(i)-N_{-}^{+}(i)}{N^{+}(i)}\right)\left(1+\dfrac{N_{-}^{+}(i)-N_{+}^{-}(i)}{N^{-}(i)}\right)}} & -1\le \mathrm{MCC}\le 1 \\ (i=1,2,\cdots,6) \end{cases} \tag{6}$$

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-mVirus for the virus proteins in each of 12 subcellular locations. As we can see from the table, nearly all the success rates achieved by the new predictor for the virus 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

### 3.3. Web Server and User Guide

As pointed out in [64], 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 [32]. In view of this, the web-server of the current pLoc_Deep-mVirus pedictor has also been established. Moreover, to maximize users' convenience, a step-by-step guide is given below.

**Step 1**. Click the link at http://www.jci-bioinfo.cn/pLoc_Deep-mVirus/, the top page of the pLoc_Deep-mVirus web-server will appear on your computer screen, as shown in **Figure 2**. Click on the Read Me button to see a brief introduction about the predictor.

**Step 2.** Either type or copy/paste the sequences of query virus proteins into the input box at the center of **Figure 2**. The input sequence should be in the FASTA format. For the examples of sequences in FASTA format, click the Example button right above the input box.

**Step 3**. Click on the Submit button to see the predicted result. For instance, if you use the four protein sequences in the Example window as the input, after 10 seconds or so, you will see a new screen (**Figure 3**) occurring. On its upper part are listed the names of the subcellular locations numbered from (1) to (6) covered by the current predictor. On its lower part are the predicted results: the query protein "P01115" of example-1 corresponds to "2," meaning it belongings to "Viral capsid" only; the query protein "P03495" of example-2 corresponds to "4, 5" meaning it belonging to "Host cytoplasm" and "Cytoplasm"; the query protein "P89873" of example-3 corresponds to "4, 5, 6", meaning it belonging to "Host cytoplasm", "Host nucleus", and "Secreted". All these results are perfectly consistent with experimental observations.

**Step 4**. As shown on the lower panel of **Figure 2**, you may also choose the batch prediction by entering your e-mail address and your desired batch input file (in FASTA format of course) via the Browse

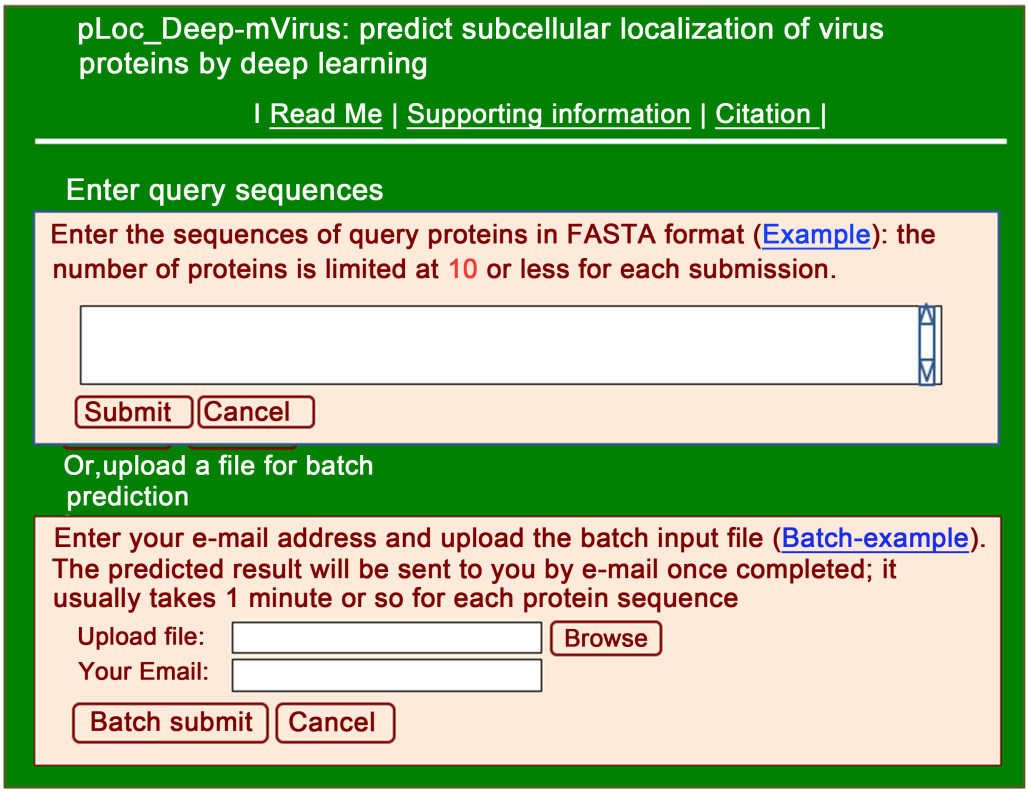


Figure 2. A semi screenshot for the top page of pLoc_Deep-mVirus.

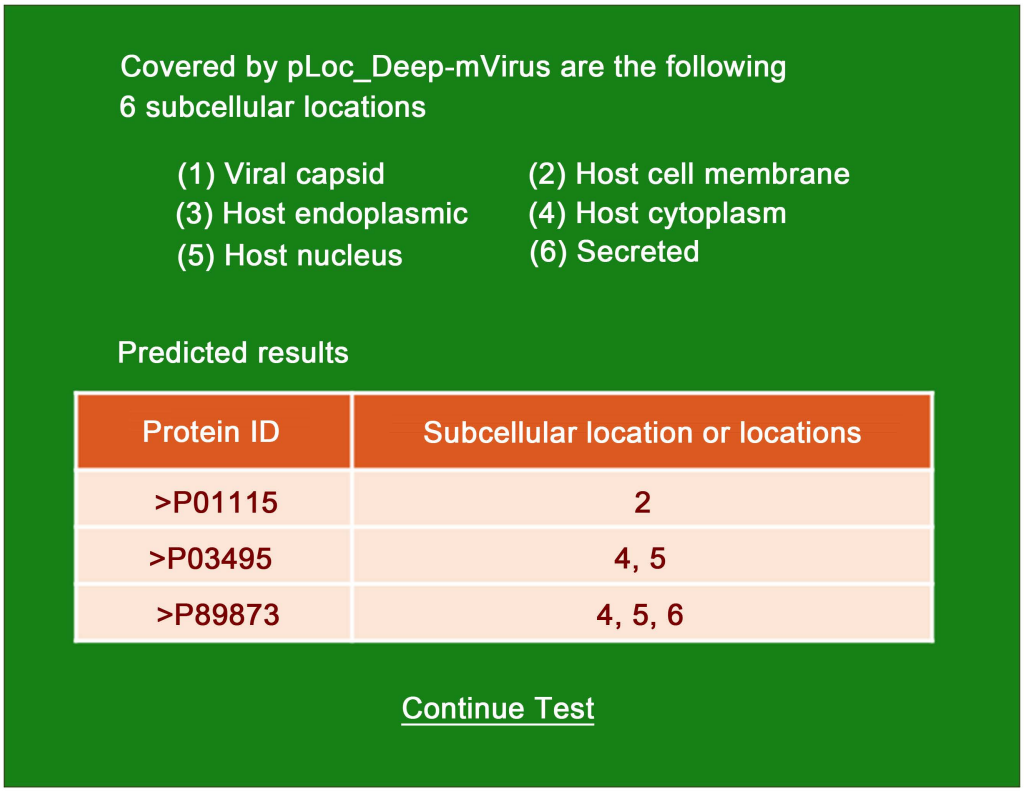


Figure 3. A semi screenshot for the webpage obtained by following Step 2 of Section 3.

button. To see the sample of batch input file, click on the button Batch-example. After clicking the button Batch-submit, you will see "Your batch job is under computation; once the results are available, you will be notified by e-mail."

**Step 5**. Click on the Citation button to find the papers that have played the key role in developing the current predictor of pLoc_Deep-mVirus.

**Step 6.** Click the Supporting Information button to download the Supporting Informations mentioned in this paper.

## 4. CONCLUSION

It is anticipated that the pLoc_Deep-mVirus predictor holds very high potential to become a useful high throughput tool in identifying the subcellular localization of virus 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
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. Xiao, X., Cheng, X., Chen, G., Mao, Q. and Chou, K.C. (2019) pLoc_bal-mVirus: Predict Subcellular Localization of Multi-Label Virus Proteins by Chou's General PseAAC and IHTS Treatment to Balance Training Dataset. *Medicinal Chemistry*, **15**, 496-509. https://doi.org/10.2174/1573406415666181217114710
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
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 Lo-

calization 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 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*), **189**, 169-180. 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**, 195-203. https://doi.org/10.1016/j.jtbi.2019.02.007
16. Charoenkwan, P., Schaduangrat, N., Nantasenamat, C., Piacham, T. and Shoombuatong, W. (2020) iQSP: A Sequence-Based Tool for the Prediction and Analysis of Quorum Sensing Peptides via Chou's 5-Steps Rule and Informative Physicochemical Properties. *International Journal of Molecular Sciences*, **21**, 75. https://doi.org/10.3390/ijms21010075
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. Shen, H.B. and Chou, K.C. (2010) Virus-mPLoc: A Fusion Classifier for Viral Protein Subcellular Location Prediction by Incorporating Multiple Sites. *Journal of Biomolecular Structure and Dynamics* (*JBSD*), **28**, 175-186. https://doi.org/10.1080/07391102.2010.10507351
19. Chou, K.C. (2001)d Prediction of Protein Cellular Attributes Using Pseudo Amino Acid Composition. *PROTEINS*: *Structure*, *Function*, *and Genetics*, **43**, 246-255. (Erratum: ibid., 2001, Vol. 44, 60) https://doi.org/10.1002/prot.1035
20. 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
21. 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
22. Zhang, S.W., Chen, W., Yang, F. and Pan, Q. (2008) Using Chou's Pseudo Amino Acid Composition to Predict Protein Quaternary Structure: A Sequence-Segmented PseAAC Approach. *Amino Acids*, **35**, 591-598. https://doi.org/10.1007/s00726-008-0086-x

23. Qiu, J.D., Huang, J.H., Liang, R.P. and Lu, X.Q. (2009) Prediction of G-Protein-Coupled Receptor Classes Based on the Concept of Chou's Pseudo Amino Acid Composition: An Approach from Discrete Wavelet Transform. *Analytical Biochemistry*, **390**, 68-73. https://doi.org/10.1016/j.ab.2009.04.009

24. Mohabatkar, H. (2010) Prediction of Cyclin Proteins Using Chou's Pseudo Amino Acid Composition. *Protein & Peptide Letters*, **17**, 1207-1214. https://doi.org/10.2174/092986610792231564

25. Qiu, J.D., Suo, S.B., Sun, X.Y., Shi, S.P. and Liang, R.P. (2011) OligoPred: A Web-Server for Predicting Homo-Oligomeric Proteins by Incorporating Discrete Wavelet Transform into Chou's Pseudo Amino Acid Composition. *Journal of Molecular Graphics & Modelling*, **30**, 129-134. https://doi.org/10.1016/j.jmgm.2011.06.014

26. Nanni, L., Lumini, A., Gupta, D. and Garg, A. (2012) Identifying Bacterial Virulent Proteins by Fusing a Set of Classifiers Based on Variants of Chou's Pseudo Amino Acid Composition and on Evolutionary Information. *IEEE-ACM Transaction on Computational Biolology and Bioinformatics*, **9**, 467-475. https://doi.org/10.1109/TCBB.2011.117

27. Khosravian, M., Faramarzi, F.K., Beigi, M.M., Behbahani, M. and Mohabatkar, H. (2013) Predicting Antibacterial Peptides by the Concept of Chou's Pseudo amino Acid Composition and Machine Learning Methods. *Protein & Peptide Letters*, **20**, 180-186. https://doi.org/10.2174/092986613804725307

28. Kumar, R., Srivastava, A., Kumari, B. and Kumar, M. (2015) Prediction of Beta-Lactamase and Its Class by Chou's Pseudo Amino Acid Composition and Support Vector Machine. *Journal of Theoretical Biology*, **365**, 96-103. https://doi.org/10.1016/j.jtbi.2014.10.008

29. Mei, J., Fu, Y. and Zhao, J. (2018) Analysis and Prediction of Ion Channel Inhibitors by Using Feature Selection and Chou's General Pseudo Amino Acid Composition. *Journal of Theoretical Biology*, **456**, 41-48. https://doi.org/10.1016/j.jtbi.2018.07.040

30. Zhang, S., Yang, K., Lei, Y. and Song, K. (2019) iRSpot-DTS: Predict Recombination Spots by Incorporating the Dinucleotide-Based Spare-Cross Covariance Information into Chou's Pseudo Components. *Genomics*, **111**, 1760-1770. https://doi.org/10.1016/j.ygeno.2018.11.031

31. 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

32. 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

33. 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

34. 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

35. 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

36. 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

37. 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

38. Chen, W., Lin, H. and Chou, K.C. (2015) Pseudo Nucleotide Composition or PseKNC: An Effective Formula-

tion for Analyzing Genomic Sequences. *Molecular BioSystems*, **11**, 2620-2634. https://doi.org/10.1039/C5MB00155B

39. 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
40. Glorot, X., Bordes, A. and Bengio, Y. (2011) Deep Sparse Rectifier Neural Networks. 14*th International Conference on Artificial Intelligence and Statistics*, Ft. Lauderdale, 11-13 April 2011, 315-323.
41. Chou, K.C. (2019) Two Kinds of Metrics for Computational Biology. *Genomics*. https://www.sciencedirect.com/science/article/pii/S0888754319304604?via%3Dihub
42. 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
43. 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
44. Zhang, M., Li, F., Marquez-Lago, T.T., Leier, A., Fan, C., Kwoh, 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
45. 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
46. 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
47. 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
48. 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
49. 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
50. 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
51. 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
52. 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
53. 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

54. 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

55. 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

56. 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. *Biosystems*, **113**, 50-57. https://doi.org/10.1016/j.biosystems.2013.04.005

57. 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

58. 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

59. 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

60. 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

61. 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

62. 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

63. 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

64. 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