Scientific Research Publishing

pLoc_Deep-mGneg: Predict Subcellular Localization of Gram Negative Bacterial Proteins by Deep Learning

Xin-Xin Liu[1], Kuo-Chen Chou[1,2]

[1]Computer Science, Jingdezhen Ceramic Institute, Jingdezhen, China
[2]Gordon Life Science Institute, Boston, MA 02478, USA
Email: 1332250394@qq.com, kcchou@gordonlifescience.org, kcchou38@gmail.com

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 Gram negative bacterial protein subcellular localization is vitally important. In view of this, a CNN based protein subcellular localization predictor called "pLoc_Deep-mGnet" 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 98% and its local accuracy is around 94% - 100%. Both are transcending other existing state-of-the-art predictors significantly. 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-mGneg/, which will become a very useful tool for fighting pandemic coronavirus and save the mankind of this planet.

Keywords

Pandemic Coronavirus, Multi-Label System, Gram Negative Bacterial Proteins, Learning at Deeper Level, Five-Steps Rule, PseAAC

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-mGneg" [4], was developed for predicting the subcellular localization of Gram negative bacterial 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] [6] [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-mGneg 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-mGneg [4] is overwhelmingly higher, particularly in the absolute true rate. 3) Although the pLoc_bal-mGneg predictor [4] has the aforementioned merits, it has not been trained at a deeper level yet [11] [12] [13] [14].

The present study was initiated in an attempt to address this problem. As done in pLoc_bal-mGneg [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-mGneg [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-mGneg [4]; *i.e.*,

$$\mathbb{S} = \mathbb{S}_1 \bigcup \mathbb{S}_2 \bigcup \cdots \bigcup \mathbb{S}_u \bigcup \cdots \bigcup \mathbb{S}_7 \bigcup \mathbb{S}_8 \tag{1}$$

where $\mathbb{S}_1$ only contains the Gram-negative bacteria protein samples from the "Cell inner membrane" organelle (cf. Table 1), $\mathbb{S}_2$ only contains those from the "Cell wall", and so forth; $\bigcup$ denotes the symbol for "union" in the set theory. For readers' convenience, their detailed sequences and accession num-

bers (or ID codes) are given in Supporting Information S1 that is also available at http://www.jci-bioinfo.cn/pLoc_bal-mGneg/Supp1.pdf in which none of proteins included has ≥25% sequence identity to any other in the same subset (subcellular location).

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 $\mathbf{P}$, its most straightforward expression is

$$\mathbf{P} = R_1 R_2 R_3 R_4 R_5 R_6 R_7 \cdots R_L \tag{2}$$

where L denotes the protein's length or the number of its constituent amino acid residues, R_1 is the 1st residue, R_2 the 2nd residue, 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 [18] or PseAAC [19] has been 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]). Meanwhile, because it has been widely and increasingly used, four powerful open access soft-wares, called "PseAAC" [23], "PseAAC-Builder" [24], "propy" [25], and "PseAAC-General" [26], were established: the former three are for generating various modes of special PseAAC [27]; while the 4th 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 [28] that have proved very successful as well (see, e.g., [29] [30]).

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

$$\mathbf{P} = \begin{bmatrix} \Psi_1 & \Psi_2 & \cdots & \Psi_u & \cdots & \Psi_\Omega \end{bmatrix}^{\mathbf{T}} \tag{3}$$

where $\mathbf{T}$ is a transpose operator, while the integer Ω is a parameter and its value as well as the components $\Psi_u \ (u = 1, 2, \cdots, \Omega)$ 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 Sec-

tion 2.2 of [4], each of the protein samples in the benchmark dataset can be uniquely defined as an 8-D numerical vector as given in <u>Supporting Information S2</u>, which can also be directly downloaded at
<u>http://www.jci-bioinfo.cn/pLoc_bal-mGneg/Supp2.pdf</u>.

2.3. Installing Deep-Learning for Three Deeper Levels

In this study, we use multilayer perceptron neural network model, which consists of 3 fully connected layers and was used to predict subcellular localization of multi-label human proteins, as illustrated in **Figure 1**. We set input layer with 14 neural units which correspond to 14 features. Too many hidden layer would make network complexity bigger and suffer from the vanishing gradient problem while a model is constructed. Here, only two hidden layer is included. The hidden layer 1 is set as 200 neural units. The activation function is set as "relu". The second hidden layer has 100 neural units. The activation function is set the same as the hidden layer 1. We end the model with 14 neural units and Sigmoid activation. To go with it, we use the binary_crossentropy loss and the adam (adaptive moment estimation) optimizer to train the model. The metrics is set as "accuracy". The batch size is set as 28, and the epochs is 100. The predicted results were decided by the output of the threshold θ. If the output is greater than 0.5, the outcome was true; otherwise, false. For more information about this, see [11], where the details have been clearly elaborated and hence there is no need to repeat here.

The new predictor developed via the above procedures is called "pLoc_Deep-mGneg", where "pLoc_Deep" stands for "predict subcellular localization by deep learning", and "mGneg" for "multi-label Gram negative proteins".

3. Results and Discussion

According to the 5-step rules [17], one of the important procedures in developing
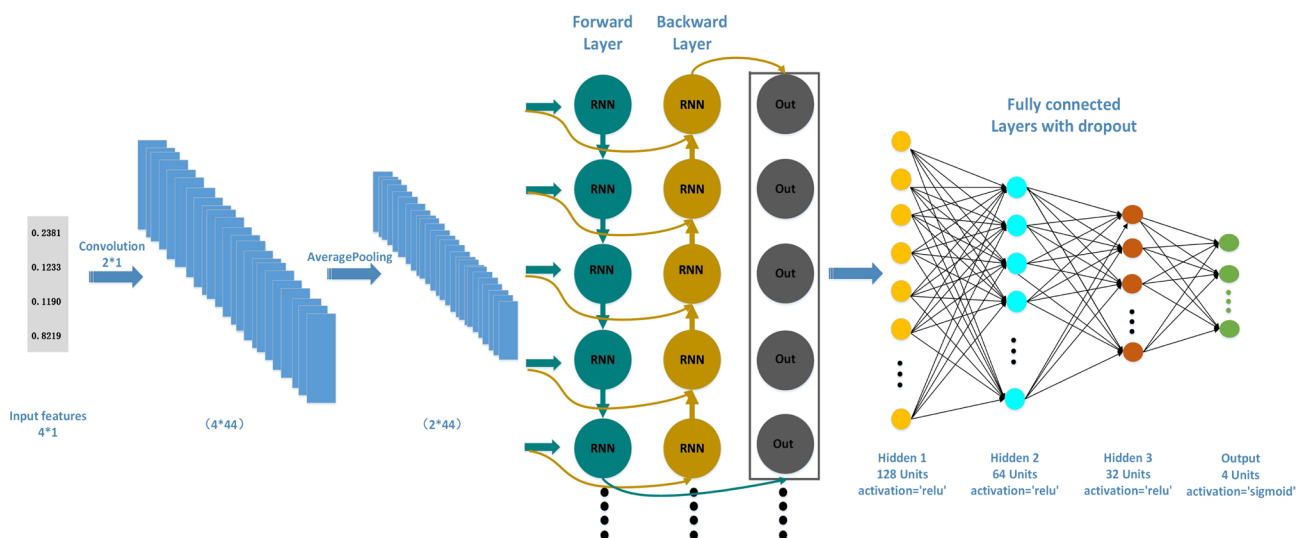

Figure 1. An illustration to show a dense neural network with 3 fully connected layers. Adapted from [11] with permission.

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:

$$\begin{cases} \text{Aiming} \uparrow = \dfrac{1}{N^q} \sum_{k=1}^{N^q} \left(\dfrac{\left\| \mathbb{L}_k \cap \mathbb{L}_k^* \right\|}{\left\| \mathbb{L}_k^* \right\|} \right), \quad [0,1] \\[3mm] \text{Coverage} \uparrow = \dfrac{1}{N^q} \sum_{k=1}^{N^q} \left(\dfrac{\left\| \mathbb{L}_k \cap \mathbb{L}_k^* \right\|}{\left\| \mathbb{L}_k \right\|} \right), \quad [0,1] \\[3mm] \text{Accuracy} \uparrow = \dfrac{1}{N^q} \sum_{k=1}^{N^q} \left(\dfrac{\left\| \mathbb{L}_k \cap \mathbb{L}_k^* \right\|}{\left\| \mathbb{L}_k \cup \mathbb{L}_k^* \right\|} \right), \quad [0,1] \\[3mm] \text{Absolute true} \uparrow = \dfrac{1}{N^q} \sum_{k=1}^{N^q} \Delta \left(\mathbb{L}_k, \mathbb{L}_k^* \right), \quad [0,1] \\[3mm] \text{Absolute false} \downarrow = \dfrac{1}{N^q} \sum_{k=1}^{N^q} \left(\dfrac{\left\| \mathbb{L}_k \cup \mathbb{L}_k^* \right\| - \left\| \mathbb{L}_k \cap \mathbb{L}_k^* \right\|}{M} \right), \quad [1,0] \end{cases} \quad (4)$$

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 $L_{\text{cell}} = 8$), $\| \ \|$ 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

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

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 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 la-

bels 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-mGneg 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-mGneg [4], the existing most powerful predictor for identifying the subcellular localization of Gram negative proteins with both single and multiple location sites. As shown in Table 1, the newly proposed predictor pLoc_Deep-mGneg is remarkably superior to the existing state-of-the-art predictor pLoc_bal-mGneg 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 98%, which is far beyond the reach of any other existing methods (see, e.g., [35] [36] [37]). 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., [38] [39]).

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 a set of four intuitive metrics that were derived in [40] based on the Chou's symbols introduced for studying protein signal peptides [41] and that have ever since been widely concurred or justified (see, e.g., [42] [43]). For the current study, the set of metrics can be formulated as:

Table 1. Comparison with the state-of-the-art method in predicting Gram negative protein subcellular localization[a].

Predictor	Aiming (↑)[a]	Coverage (↑)[a]	Accuracy (↑)[a]	Absolute true (↑)[a]	Absolute false (↓)[a]
pLoc_bal-mGneg[b]	97.07%	97.8%	97.27%	96.55%	0.19%
pLoc_Deep-mGneg[c]	98.52%	98.64%	98.40%	98.02%	00.00%

[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-mGneg.

146

Table 2. Performance of pLoc_Deep-mGneg for each of the 8 subcellular locations.

i	Location[a]	Sn(i)[b]	Sp(i)[b]	Acc(i)[b]	MCC(i)[b]
1	Cell inner membrane	0.9805	0.9910	0.9882	0.9709
2	Cell outer membrane	0.9465	0.9989	0.9921	0.9648
3	Cytoplasm	0.9707	0.9896	0.9842	0.9616
4	Extracellular	0.9786	0.9976	0.9951	0.9807
5	Fimbrium	0.9333	0.9985	0.9961	0.9358
6	Flagellum	1.0000	1.0000	1.0000	1.0000
7	Nucleoid	1.0000	1.0000	1.0000	1.0000
8	Periplasm	0.9814	0.9988	0.9956	0.9852

[a]See Equation (1) and relevant context as well as the Supporting Information S1 for further explanation.
[b]See Equation (6) for the metrics definition.

$$
\begin{cases}
\mathrm{Sn}(i) = 1 - \dfrac{N_-^+(i)}{N^+(i)} & 0 \le \mathrm{Sn} \le 1 \\[2mm]
\mathrm{Sp}(i) = 1 - \dfrac{N_+^-(i)}{N^-(i)} & 0 \le \mathrm{Sp} \le 1 \\[2mm]
\mathrm{Acc}(i) = 1 - \dfrac{N_-^+(i) + N_+^-(i)}{N^+(i) + N^-(i)} & 0 \le \mathrm{Acc} \le 1 \\[2mm]
\mathrm{MCC}(i) = \dfrac{1 - \left(\dfrac{N_-^+(i)}{N^+(i)} + \dfrac{N_+^-(i)}{N^-(i)} \right)}{\sqrt{\left(1 + \dfrac{N_+^-(i) - N_-^+(i)}{N^+(i)}\right)\left(1 + \dfrac{N_-^+(i) - N_+^-(i)}{N^-(i)}\right)}} & -1 \le \mathrm{MCC} \le 1 \\[2mm]
(i = 1, 2, \cdots, 8)
\end{cases} \tag{6}
$$

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-mGneg for the Gram negative proteins in each of 8 subcellular locations. As we can see from the table, nearly all the success rates achieved by the new predictor for the Gram negative proteins in each of the 8 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 a series of somewhat provocative but quite interesting papers (see, e.g., [44]-[49]).

3.3. Web Server and User Guide

As pointed out in [50], 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-mGneg predictor has also been established at http://www.jci-bioinfo.cn/pLoc_Deep-mGneg/, 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-mGneg predictor holds very high potential to become a useful high throughput tool in identifying the subcellular localization of Gram negative bacterial 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 mankind on this planet.

Acknowledgements

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] Cheng, X., Xiao, X. and Chou, K.C. (2018) pLoc_bal-mGneg: Predict Subcellular Localization of Gram-Negative Bacterial Proteins by Quasi-Balancing Training Dataset and General PseAAC. *Journal of Theoretical Biology*, **458**, 92-102. https://doi.org/10.1016/j.jtbi.2018.09.005

[5] Nakai, K. and Kanehisa, M. (1992) A Knowledge Base for Predicting Protein Loca-

lization 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 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 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*. 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] Dutta, A., Dalmia, A., R, A., Singh, K.K. and Anand, A. (2019) Using the Chou's 5-Steps Rule to Predict Splice Junctions with Interpretable Bidirectional Long Short-Term Memory Networks. *Computers in Biology and Medicine*, **116**, Article ID: 103558. https://doi.org/10.1016/j.compbiomed.2019.103558

[16] Ehsan, A., Mahmood, M.K., Khan, Y.D., Barukab, O.M., Khan, S.A. and Chou, K.C. (2019) iHyd-PseAAC (EPSV): Identify Hydroxylation Sites in Proteins by Extracting Enhanced Position and Sequence Variant Feature via Chou's 5-Step Rule and General Pseudo Amino Acid Composition. *Current Genomics*, **20**, 124-133.
https://doi.org/10.2174/1389202920666190325162307

[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. (Erratum: ibid., 2001, Vol. 44, 60) 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] 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

[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] Shen, H.B. and Chou, K.C. (2008) PseAAC: A Flexible Web-Server for Generating Various Kinds of Protein Pseudo Amino Acid Composition. *Analytical Biochemistry*, **373**, 386-388. https://doi.org/10.1016/j.ab.2007.10.012

[24] 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

[25] 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

[26] 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

[27] 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

[28] 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

[29] 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

[30] 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

[31] Chou, K.C. (2019) Two Kinds of Metrics for Computational Biology. *Genomics*. https://www.sciencedirect.com/science/article/pii/S0888754319304604?via%3Dihu

[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., 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

[35] 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

[36] 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

[37] 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

[38] 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

[39] 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

[40] 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

[41] 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

[42] Lin, H., Deng, E.Z., Ding, H., Chen, W. and Chou, K.C. (2014) iPro54-PseKNC: A Sequence-Based Predictor for Identifying Sigma-54 Promoters in Prokaryote with Pseudo k-Tuple Nucleotide Composition. *Nucleic Acids Research*, **42**, 12961-12972. https://doi.org/10.1093/nar/gku1019

[43] Feng, P., Yang, H., Ding, H., Lin, H., Chen, W. and Chou, K.C. (2019) iDNA6mA-PseKNC: Identifying DNA N(6)-methyladenosine Sites by Incorporating Nucleotide Physicochemical Properties into PseKNC. *Genomics*, **111**, 96-102. https://doi.org/10.1016/j.ygeno.2018.01.005

[44] 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

[45] Chou, K.C. (2020) The Most Important Ethical Concerns in Science. *Natural Science*, **12**, 35-36. https://doi.org/10.4236/ns.2020.122005

[46] 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

[47] 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

[48] Chou, K.C. (2020) Proposing 5-Steps Rule Is a Notable Milestone for Studying Mo-

lecular Biology. *Natural Science*, **12**, 74-79. https://doi.org/10.4236/ns.2020.123011

[49] 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

[50] 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