An Insightful Recollection for Predicting Protein Subcellular Locations in Multi-Label Systems

Kuo-Chen Chou

Gordon Life Science Institute, Boston, Massachusetts 02478, United States of America

Correspondence to: Kuo-Chen Chou, kcchou@gordonlifescience.org, kcchou38@gmail.com

Keywords: Chou's 5-Steps Rule, Chou's PseAAC, Web-Server GO Approach, FunD Approach

Received: July 4, 2020 **Accepted:** July 12, 2020 **Published:** July 15, 2020

ABSTRACT

A systematic introduction has been presented for the recent advances in predicting protein subcellular localization in the multi-label systems, where the constituent proteins may simultaneously occur or move between two or more location sites and hence have exceptional biological functions worthy of our special notice. All the predictors included in this review each have a user-friendly web-server, by which the majority of experimental scientists can very easily acquire their desired data without the need to go through the complicated mathematics involved.

1. INTRODUCTION

As elucidated in two recent comprehensive review papers [1, 2], to develop a really useful bioinformatics tool, one needs to observe the guidelines of the Chou's 5-steps rule [2-36] to go through the following five steps: 1) select or construct a valid benchmark dataset to train and test the predictor; 2) represent the samples with an effective formulation that can truly reflect their intrinsic correlation with the target to be predicted; 3) introduce or develop a powerful algorithm to conduct the prediction; 4) properly perform cross-validation tests to objectively evaluate the anticipated prediction accuracy; 5) establish a user-friendly web-server for the predictor that is accessible to the public. The bioinformatics or computational tool established by observing the guidelines of Chou's 5-step rules have the following remarkable merits: a) crystal clear in logic development, b) completely transparent in operation, c) easily to repeat the reported results by other investigators, d) with high potential in stimulating other new bioinformatics tools, and e) very convenient to be used by the majority of experimental scientists. As for more about the importance of the 5-steps rule, see an insightful Wikipedia article at https://en.wikipedia.org/wiki/5-step_rules. It is instructive to point out that, although the present minireview was focused on the recent development in subcellular prediction for the multi-label proteins [37, 38], the 5-steps rule can also be used to deal with many different systems, such as those in material science [39] and even those in commercial science (e.g., analyzing the effect of bank credit card versus mobile payment). The only difference between the biological systems and other disciplines' systems is how to formulate the statistical samples or events with an effective mathematical expression that can truly reflect their intrinsic correlation with the target to be predicted. This is just like the case of many machine-learning algorithms. They can be used in nearly all the areas of statistical analysis.

2. PREDICTING SUBCELLULAR LOCALIZATION OF PROTEINS

The smallest unit of life is a cell, which contains numerous protein molecules. Most of the functions critical to the cell's survival are performed by these proteins located in its different organelles, usually called "subcellular locations" (**Figure 1**). Information of subcellular localization for a protein can provide useful clues about its function. To reveal the intricate pathways at the cellular level, knowledge of the subcellular localization of proteins in a cell is prerequisite. Unfortunately, it is both time-consuming and costly to determine the subcellular locations of proteins purely based on experiments. With the avalanche of protein sequences generated in the post-genomic age, it is highly desired to develop computational tools for rapidly and effectively identifying the subcellular locations of uncharacterized proteins based on their sequences information alone. The demand has become even more challenging owing to the fact that many protein molecules may simultaneously exist or move between two or more subcellular location sites [40]. Actually, it is these multiplex proteins that are of significance for in-depth understanding the biological processes in a living cell.

3. FOUR SERIES OF PREDICTORS

In the last decade or so, a number of predictors were developed for predicting the subcellular localization of proteins with both single site and multiple sites based on their sequences information alone. They can be generally classified into four series: 1) 𝕏-mPLoc, 2) iLoc𝕏, 3) pLoc-m𝕏, and 4) pLoc_bal-m𝕏, where the wildcard may denote "Euk" (eukaryotic), "Hum" (human), "Animal", "Plant", "Virus", "Gneg"

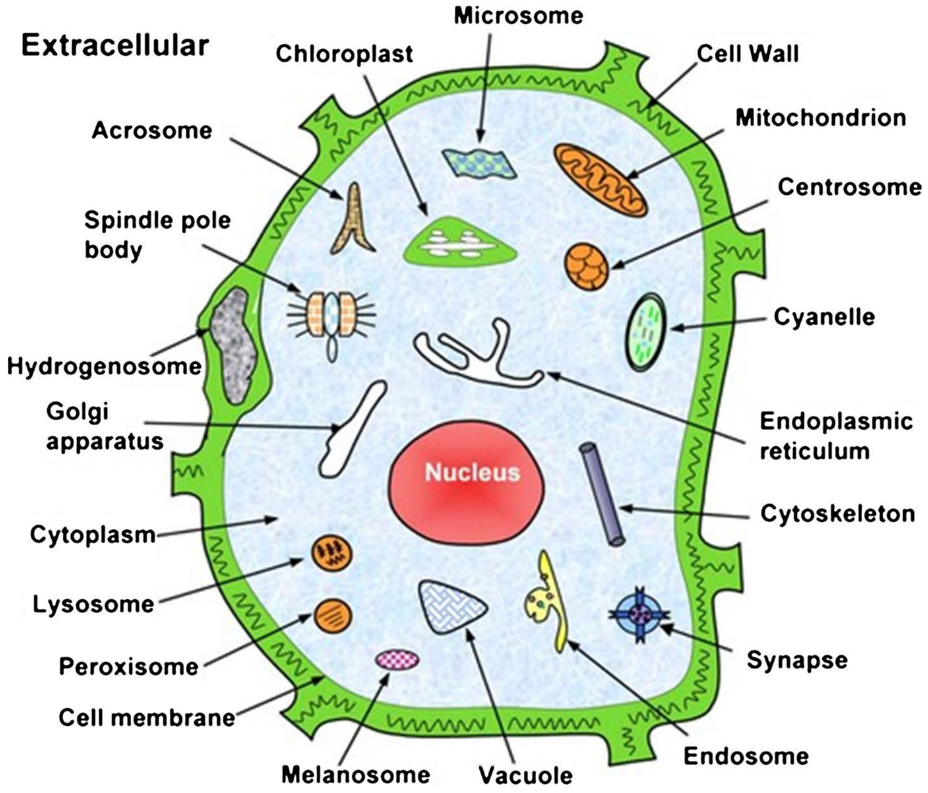


Figure 1. Schematic illustration to show the 22 organelles or subcellular locations in an eukaryotic cell. Adapted from Chou and Shen with permission [189].

(Gram-negative bacterial), “Gpos” (Gram-positive bacterial) proteins, respectively, as formulated by

$$\mathbb{X} \in \begin{cases} \text{Euk} \\ \text{Hum} \\ \text{Animal} \\ \text{Plant} \\ \text{Virus} \\ \text{Gneg} \\ \text{Gpos} \end{cases} \tag{1}$$

The protein samples in the $\mathbb{X}$-mPLoc series [41-46] were formulated by hybridizing the GO (Gene Ontology) information, FunD (Functional Domain) information, and PSSM (Sequential Evolutionary) information into the general PseAAC [3], which was extended from pseudo amino acid composition [47, 48].

The protein samples in the iLoc-$\mathbb{X}$series [49-55] were formulated by incorporating the GO information and PSSM information into the general PseAAC.

The protein samples in the pLoc-m $\mathbb{X}$series [56-62] were formulated by extracting the key or optimal GO information into the general PseAAC.

The protein samples in the pLoc_bal-m $\mathbb{X}$series [2, 26, 29, 63-65] were formulated by further balancing out the protein samples used in pLoc-m $\mathbb{X}$series.

As for the justification of using the GO information for predicting the subcellular localization of proteins, see Section 4 of a review paper [66], where an insightful analysis has been elaborated and there is no need to repeat here.

3.1. Benchmark Dataset

All the predictors in the above four series were developed based on a very stringent benchmark dataset in which none of proteins had ≥25% pairwise sequence identity to any other in a same subset. But such a strict cutoff treatment was not imposed for the protein sequences in the “viral capsid” subset because otherwise it would contain too few proteins to be of statistical significance as explained in [46].

3.2. Sample Formulation

The most straightforward expression for a protein sample is its sequential model as given by

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

where L denotes the protein’s length or the number of its constituent amino acid residues, 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 (e.g., “Support Vector Machine” or SVM algorithm [4, 5], “Covariance Discriminant” or CD algorithm [67-69], “Nearest Neighbor” or NN algorithm [70, 71], and “Random Forest” or RF algorithm [72, 73]) can only handle vectors as elaborated in [74], we have to convert the sequential expression of Equation (2) into a vector. But a vector defined in a discrete model might completely lose all the sequence order or pattern information. To deal with this problem, the concept of PseAAC (Pseudo Amino Acid Composition) was introduced [47, 48]. Ever since then, the concept of PseAAC 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, 28, 75-170] as well as a long list of references cited in [171]). Because it has been widely and increasingly used, four powerful open access soft-wares, called “PseAAC” [172], “PseAAC-Builder” [88], “propy” [98], and “PseAAC-General” [109], were established: the former three are for generating various modes of special PseAAC [173]; while the fourth one for those of general PseAAC [3], including not only all the special modes of feature vectors for proteins but also the higher level feature vectors such as “Functional Domain” or “FunD” mode, “Gene Ontology” or “GO” mode, and "Sequential Evolution" or "PSSM" [174] 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 [175-178] that have proved very successful as well [14, 179-190]. According to the concept of general PseAAC [3], 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 **T** is a transpose operator, while the integer Ω 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**.

3.3. Operation Engine

The operation engine for ×-mPLoc series was constructed by fusing an array of OET-KNN (Optimized Evidence-Theoretic K-Nearest Neighbor) classifiers [191-193].

The operation engine for iLoc-×series was the multi-labeled KNN (K-Nearest Neighbor) classifier [49].

The operation engine for the pLoc-m and pLoc_bal-m ×series was the ML-GKR (multi-label Gaussian kernel regression) classifier [56].

3.3.1. Metrics and Cross-Validation

In order to objectively evaluate the prediction quality of a multi-label predictor, one needs to consider the following two issues. 1) What metrics should be used to quantitatively reflect its accuracy? 2) What test approach should be adopted to score the metrics?

Quite different from the metrics used to measure the prediction quality of a single-label predictor, the metrics for a multi-label predictor are much more complicated. To quantitatively evaluate the power of a multi-label predictor, we need to use two sets of metrics: one for its global accuracy and the other for its local accuracy.

The global accuracy is defined by a set of five metrics as given in [66]

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

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, || || means the operator acting on the set therein to count the number of its elements, ⋃ means the symbol for the "union" in the set theory, ⋂ 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, represents the subset that contains all the labels predicted for the *k*-th sample, and $\mathbb{L}_k^*$ 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 ↑ 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 is called negative metrics, implying just the opposite meaning. As we can see from Equation (5): 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.

The five metrics in Equation (4) reflect the quality of a multi-label predictor from five different angles at the global level. It is instructive to point out, however, among the five global metrics the most important one and also the most difficult to improve its success rate is the "Absolute true" or "perfectly correct" rate [66]. Why? This is because the score standard for the absolute true rate is very harsh. According to its definition, for a protein sample that is actually simultaneously located at the subcellular locations ("A", "B", "C"). If the predicted result is not exactly the three locations but ("A", "B") or ("A", "B", "C", "D"), no score whatsoever will be given. In other words, when and only when the predicted localization for the protein sample is perfectly identical to its actual localization, can we add one point for the absolute true rate; otherwise, zero.

The set of metrics in Equation (4) are used to evaluate the prediction quality of a multi-label predictor for all the proteins in the entire cell, and hence is called the "set of metrics for the global accuracy" or the "set of global metrics".

To evaluate the local accuracy of a multi-label predictor, we use a set of Chou's four intuitive metrics that were derived by Chou *et al.* [4, 69] based on the symbols introduced by Chou [194-196] for studying the cleavage sites of signal peptides. The set of metrics are given below

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

where Sn, Sp, Acc, and MCC represent the sensitivity, specificity, accuracy, and Mathew's correlation coefficient, respectively [15], *i* denotes the *i*-th subcellular location (or subset) in the benchmark dataset, and *M* has exactly the same meaning as in Equation (5). $N^+(i)$ 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 location 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.

In addition to being widely used in proteome and genome analyses (see, e.g., [6, 8, 10, 13, 15, 33, 36, 180, 181, 185, 197-203]), the set of metrics in Equation (6) can be used to evaluate the prediction quality of a multi-label predictor for the proteins in each of subcellular locations concerned (see, e.g., [58, 62]), and hence is called the "set of metrics for local accuracy" or the "set of local metrics".

3.3.2. Cross-Validation and Jackknife Test

Three cross-validation methods are often used in statistical prediction. They are: 1) independent dataset test, 2) subsampling (or K-fold cross-validation) test, and 3) jackknife test [204]. Of these three, however, the jackknife test was deemed the least arbitrary that can always yield a unique result for a given benchmark dataset [37, 38], as clearly elucidated in a comprehensive review paper [3] and demonstrated by Eqs. (28)-(32) therein. Therefore, the jackknife test has been increasingly recognized and widely adopted by investigators to test the power of various prediction methods (see, e.g., [80, 82, 101, 110, 205-208]).

Therefore, all the predictors in Section 2 were examined by the jackknife tests.

3.4. Web Servers

The last but not least important guideline in the 5-step rules is about the web-server. As pointed out in [209] and demonstrated in a series of recent publications (see, e.g., [5-15, 17-19, 174, 180-185, 197-201, 203, 210-244]), user-friendly and publicly accessible web-servers represent the future direction for developing practically more useful predictors. Actually, many practically useful web-servers have significantly increased bioinformatics impacts on medicinal chemistry [74], driving medicinal chemistry into an unprecedented revolution [171].

All the multi-label predictors listed in Section 2 have their web-servers well established as summarized below.

3.4.1. mPLoc Series

This series contains six publicly accessible web-servers: (1) "Euk-mPLoc" at http://www.csbio.sjtu.edu.cn/bioinf/euk-multi-2/ [43] for predicting the subcellular localization of eukaryotic proteins. (2) "Hum-mPLoc" at http://www.csbio.sjtu.edu.cn/bioinf/hum-multi-2/ [41] for predicting the subcellular localization of human proteins. (3)"Plant-mPLoc" at http://www.csbio.sjtu.edu.cn/bioinf/plant-multi/ [44] for predicting the subcellular localization of plant proteins. (4)"Virus-mPLoc" at http://www.csbio.sjtu.edu.cn/bioinf/virus-multi/ [46] for predicting the subcellular localization of virus proteins. (5)"Gneg-mPLoc" at http://www.csbio.sjtu.edu.cn/bioinf/Gneg-multi/ [45] for predicting the subcellular localization of Gram-negative bacterial proteins. (6) "Gpos-mPLoc" at http://www.csbio.sjtu.edu.cn/bioinf/Gpos-multi/ [42] for predicting subcellular localization of Gram-positive bacterial proteins.

The aforementioned six web-servers have also been integrated into a package called "Cell-PLoc" at PLoc/ [37] and its updated version "Cell-PLoc 2.0" at http://www.csbio.sjtu.edu.cn/bioinf/Cell-PLoc-2/ [38].

3.4.2. iLoc-Series

It contains the following seven web-servers. 1) "iLoc-Euk" at http://www.jci-bioinfo.cn/iLoc-Euk [49] for predicting the subcellular localization of eukaryotic proteins. 2) "iLoc-Hum" at http://www.jcibioinfo.cn/iLoc-Hum [52] for predicting the subcellular localization of human proteins. 3) "iLoc-Animal" at Animal [55] for predicting the subcellular localization of animal proteins. 4) "iLoc-Plant" at http://www.jci-bioinfo.cn/iLoc-Plant [50] for predicting the subcellular localization of plant proteins. 5) "iLoc-Virus" at http://www.jci-bioinfo.cn/iLoc-Virus [51] for predicting the subcellular localization of virus proteins. 6) "iLoc-Gneg" at http://www.jcibioinfo.cn/iLoc-Gneg [53] for predicting the subcellular localization of Gram-negative proteins. 7) "iLoc-Gpos" at http://www.jci-bioinfo.cn/iLoc-Gpos [54].

3.4.3. pLoc-m Series

There are seven web-servers in this series as listed below. 1) "pLoc-mEuk" at http://www.jci-bioinfo.cn/pLoc-mEuk/ [60] for predicting the subcellular localization of eukaryotic proteins. 2) "pLoc-mHum" at http://www.jci-bioinfo.cn/pLoc-mHum/ [62] for predicting the subcellular localization of human proteins. 3) "pLoc-mAnimal" at http://www.jci-bioinfo.cn/pLoc-mAnimal/ [58] for predicting the subcellular localization of animal proteins. 4) "pLoc-mPlant" at http://www.jci-bioinfo.cn/pLoc-mPlant/ [56] for predicting the subcellular localization of plant proteins. 5) "pLoc-mVirus" at http:// www.jci-bioinfo.cn/pLoc-mVirus/ [57] for predicting the subcellular localization of virus proteins. 6) "pLoc-mGneg" at http://www.jcibioinfo.cn/pLoc-mGneg/ [61] for predicting the subcellular localization of Gram-negative proteins. 7) "pLoc-mGpos" at http://www.jcibioinfo.cn/pLoc-mGpos/ [59] for predicting the subcellular localization of Gram-positive proteins.

3.4.4. pLoc_bal-m Series

There are seven web-servers in this series as listed below. 1) "pLoc_bal-mEuk" at [2]. 2) "pLoc_bal-mHum" [2]. 3) "pLoc_bal-mAnimal" [65]. 4) "pLoc_bal-mPlant" [26]. 5) "pLoc_bal-mVirus" [29]. 6) "pLoc_bal-mGneg" [63]. 7) "pLoc_bal-mGpos" [29].

Listed in **Table 1** are the global accuracy rates (cf. Equation (4)) predicted with the aforementioned seven web-servers, while the corresponding the local accuracy rates (cf. Equation (6)) are given in **Table 2**. As shown from the rates in the two tables, all the seven web-servers have yielded very high prediction quality in both the global and local cases. Therefore, using these web-servers, the majority of experimental scientists can easily obtain their desired results without the need to go through the detailed mathematics involved.

Below, let us take the multi-label predictor of pLoc_bal-mEuk [2] as a showcase. 1) Click the link at mEuk/, you'll see the top page for predicting the eukaryotic protein subcellular localization prompted on your computer screen (**Figure 2**). 2) You can either type or copy/paste the sequences of query eukaryotic proteins into the input box at the center of **Figure 2**. The input sequence should be in the FASTA format. You can click the Example button right above the input box to see the sequences in FASTA

Table 1. List of the five global metrics rates reported from each of the seven predictors in the pLoc_bal-m series.

No	Predictor[a]	Aiming[b]	Coverage[b]	Accuracy[b]	Absolutetrue[b]	Absolutefalse[b]
1	pLoc_bal-mEuk	88.31%	85.06%	84.34%	78.78%	0.07%
2	pLoc_bal-mHum	90.57%	82.75%	84.39%	79.14%	1.20%
3	pLoc_bal-mAnimal	87.96%	85.33%	84.64%	73.11%	1.65%
4	pLoc_bal-mPlant	91.74%	87.39%	88.02%	84.87%	0.78%
5	pLoc_bal-mVirus	88.97%	92.86%	89.77%	82.13%	2.66%
6	pLoc_bal-mGneg	96.61%	95.81%	96.05%	94.68%	0.36%
7	pLoc_bal- mGpos	97.69%	97.13 %	97.40 %	97.11%	0.14%

[a] See Equation (1) of Section 2 for further explanation. [b] See Equation (5) for the definition of the global metrics.

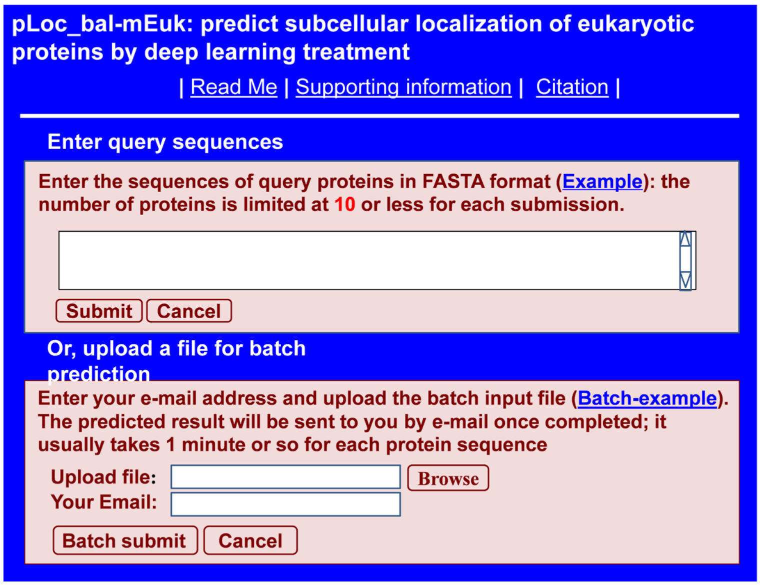
Table 2. Performance of pLoc_bal-mEuk for each of the 22 subcellular locations.

i	Location[a]	Sn(*i*)[b]	Sp(*i*)[b]	Acc(*i*)[b]	MCC(*i*)[b]
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
13	Hydrogenosome	1.0000	1.0000	1.0000	1.0000
14	Lysosome	1.0000	0.9999	0.9999	0.9913
15	Melanosome	1.0000	1.0000	1.0000	1.0000
16	Microsome	1.0000	0.9995	0.9995	0.8742
17	Mitochondrion	1.0000	0.9940	0.9945	0.9636
18	Nucleus	0.8858	0.9550	0.9343	0.8429
19	Peroxisome	1.0000	0.9988	0.9988	0.9609
20	Spindle pole body	1.0000	0.9991	0.9991	0.9518
21	Synapse	1.0000	0.9994	0.9994	0.9504
22	Vacuole	1.0000	0.9984	0.9985	0.9657

[a]See Table 1 and the relevant context for further explanation. [b]See Equation (7) for the metrics definition.

format. 3) Click on the Submit button to see the predicted result; e.g., 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 shown up (Figure 3). Listed on its upper part are the names of the subcellular locations numbered from "1" to "22" that are covered by the predictor for the eukaryotic proteins. Shown in its lower part is a table of two columns. Listed in the left-column are the IDs of query proteins; listed in the right column are the predicted subcellular locations denoted by the integer numbers within the range of 1 to 22. As we can see from the figure, the output for the query protein Q63564 of example-1 is "1," meaning it belonging to "acrosome" only; the output for the query protein P23276 of example-2 is "2, 8" meaning it belonging to "cell membrane" and "cytoskeleton"; the output for the query protein Q9VVV9 of example-3 is "2, 7, 18", meaning it belonging to "cell membrane", "cytoplasm", and "nucleus"; the output for the query protein Q673G8 of example-4 is "2, 7, 10, 18", meaning it belonging to "cell membrane", "cytoplasm", "endosome", and "nucleus". All these results are perfectly consistent with experimental observations.

As shown on the lower panel of Figure 2, you may also choose the batch prediction by entering your



pLoc_bal-mEuk: predict subcellular localization of eukaryotic proteins by deep learning treatment

Enter query sequences

Enter the sequences of query proteins in FASTA format (Example): the number of proteins is limited at 10 or less for each submission.

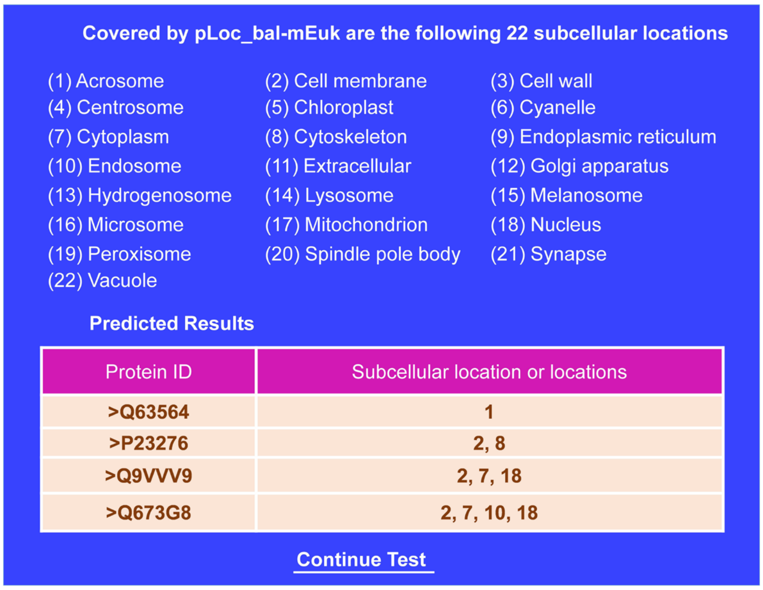
Or, upload a file for batch prediction

Enter your e-mail address and upload the batch input file (Batch-example). The predicted result will be sent to you by e-mail once completed; it usually takes 1 minute or so for each protein sequence

Upload file:

Your Email:

Figure 2. A semi screenshot for the top page of pLoc_bal-mEuk.



Covered by pLoc_bal-mEuk are the following 22 subcellular locations

(1) Acrosome (2) Cell membrane (3) Cell wall
(4) Centrosome (5) Chloroplast (6) Cyanelle
(7) Cytoplasm (8) Cytoskeleton (9) Endoplasmic reticulum
(10) Endosome (11) Extracellular (12) Golgi apparatus
(13) Hydrogenosome (14) Lysosome (15) Melanosome
(16) Microsome (17) Mitochondrion (18) Nucleus
(19) Peroxisome (20) Spindle pole body (21) Synapse
(22) Vacuole

Predicted Results

Protein ID	Subcellular location or locations
>Q63564	1
>P23276	2, 8
>Q9VVV9	2, 7, 18
>Q673G8	2, 7, 10, 18

Figure 3. A semi screenshot for the webpage obtained by following Step 3 of Section 3.5.4.

e-mail addresses and your batch input file (in FASTA format of course) via the Browse 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."

4. CONCLUSIONS AND PERSPECTIVE

The development of protein subcellular location prediction can be separated into two stages. In the early stage, all the prediction methods were developed with the assumption that each of the constituent proteins in a cell was located in one and only one location (organelle). Although those methods did play important roles in stimulating the development of such a fundamental area in cell molecular biology and proteomics, the aforementioned original hypothesis has been proved not completely correct. With more experimental data available, it has been found that many protein molecules may simultaneously exist or move between two or more subcellular location sites. It is these multiplex proteins that are of significance for in-depth understanding the biological processes in a living cell.

Since a multiplex protein needs the multiple labels to mark its locations, the multi-label theory and techniques [66] have been introduced into this frontier area of molecular biology. Meanwhile, to examine the power of a multi-label predictor, two sets of metrics have been introduced: one is the set of global metrics for evaluating its accuracy for an entire cell or in the global level, and the other is the set of local metrics for evaluating its accuracy for a specific subcellular location or in the local level. Of these metrics, the most important is the one for measuring the success rate of "absolute true" at the global level, which is also the harshest one for improvement.

The predictors introduced in this review paper have been all established by following the 5-steps rule [3], and hence they each have a user-friendly web server for the majority of experimental scientists to easily get their desired data. Also, their cornerstones are based on PseAAC [3, 47, 48, 173, 245], and hence their prediction quality is usually higher than the other methods.

It has not escaped our notice that since multi-label proteins usually have some unique or exceptional functions [37, 38, 74, 246], the advance in predicting this kind of proteins is far beyond the meaning of merely understanding the biological process concerned. It will play increasingly important roles for designing multi-target drugs [247-251], which represents a very hot trend currently in drug development [252].

It is instructive to point out that, in comparison with their counterparts, the benchmark datasets in Section 3.1 have the following two merits: 1) more stringent in excluding homology bias, and 2) cover more location sites. It is expected, however, with more experimental data available in future, they will also need updated in both the stringent criteria and coverage scope, so as to further empower the multi-label predictors in Section 3.5.4.

Finally, it is illuminative to point out that using graphic approaches to study biological and medical systems can provide an intuitive vision and useful insights for helping analyze complicated relations therein as shown in the systems of enzyme fast reaction [253-255], graphical rules in molecular biology [256-259], and low-frequency internal motion in biomacromolecules (such as protein and DNA) [260]. Particularly, what happened is that this kind of insightful implication has also been demonstrated in [261] and many follow-up publications [262-285].

CONFLICTS OF INTEREST

The author declares no conflicts of interest regarding the publication of this paper.

REFERENCES

1. Chou, K.C. (2019) Progresses in Predicting Post-Translational Modification. *International Journal of Peptide Research and Therapeutics* (*IJPRT*). https://doi.org/10.1007/s10989-019-09893-5

2. Chou, K.C. (2019) Advance in Predicting Subcellular Localization of Multi-Label Proteins and Its Implication for Developing Multi-Target Drugs. *Current Medicinal Chemistry*. https://doi.org/10.2174/0929867326666190507082559

3. 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

4. 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

5. Feng, P.M., Chen, W., Lin, H. and Chou, K.C. (2013) iHSP-PseRAAAC: Identifying the Heat Shock Protein Families Using Pseudo Reduced Amino Acid Alphabet Composition. *Analytical Biochemistry*, **442**, 118-125.

https://doi.org/10.1016/j.ab.2013.05.024

6. 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

7. Chen, W., Feng, P.M., Deng, E.Z., Lin, H. and Chou, K.C. (2014) iTIS-PseTNC: A Sequence-Based Predictor for Identifying Translation Initiation Site in Human Genes Using Pseudo Trinucleotide Composition. *Analytical Biochemistry*, **462**, 76-83. https://doi.org/10.1016/j.ab.2014.06.022

8. Ding, H., Deng, E.Z., Yuan, L.F., Liu, L., Lin, H., Chen, W. and Chou, K.C. (2014) iCTX-Type: A Sequence-Based Predictor for Identifying the Types of Conotoxins in Targeting Ion Channels. *BioMed Research International*, **2014**, Article ID: 286419. https://doi.org/10.1155/2014/286419

9. Liu, B., Fang, L., Wang, S., Wang, X., Li, H. and Chou, K.C. (2015) Identification of microRNA Precursor with the Degenerate K-Tuple or Kmer Strategy. *Journal of Theoretical Biology*, **385**, 153-159. https://doi.org/10.1016/j.jtbi.2015.08.025

10. Liu, Z., Xiao, X., Qiu, W.R. and Chou, K.C. (2015) iDNA-methyl: Identifying DNA Methylation Sites via Pseudo Trinucleotide Composition. *Analytical Biochemistry*, **474**, 69-77. https://doi.org/10.1016/j.ab.2014.12.009

11. Xiao, X., Min, J.L., Lin, W.Z., Liu, Z., Cheng, X. and Chou, K.C. (2015) iDrug-Target: Predicting the Interactions between Drug Compounds and Target Proteins in Cellular Networking via the Benchmark Dataset Optimization Approach. *Journal of Biomolecular Structure and Dynamics*, **33**, 2221-2233. https://doi.org/10.1080/07391102.2014.998710

12. Jia, J., Liu, Z., Xiao, X., Liu, B. and Chou, K.C. (2016) iSuc-PseOpt: Identifying Lysine Succinylation Sites in Proteins by Incorporating Sequence-Coupling Effects into Pseudo Components and Optimizing Imbalanced Training Dataset. *Analytical Biochemistry*, **497**, 48-56. https://doi.org/10.1016/j.ab.2015.12.009

13. 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

14. Liu, B., Fang, L., Long, R., Lan, X. and Chou, K.C. (2016) iEnhancer-2L: A Two-Layer Predictor for Identifying Enhancers and Their Strength by Pseudo k-Tuple Nucleotide Composition. *Bioinformatics*, **32**, 362-369. https://doi.org/10.1093/bioinformatics/btv604

15. Chen, W., Feng, P., Yang, H., Ding, H., Lin, H. and Chou, K.C. (2017) iRNA-AI: Identifying the Adenosine to Inosine Editing Sites in RNA Sequences. *Oncotarget*, **8**, 4208-4217. https://doi.org/10.18632/oncotarget.13758

16. Chen, W., Ding, H., Zhou, X., Lin, H. and Chou, K.C. (2018) iRNA(m6A)-PseDNC: Identifying N6-Methyladenosine Sites Using Pseudo Dinucleotide Composition. *Analytical Biochemistry*, **561**-**562**, 59-65. https://doi.org/10.1016/j.ab.2018.09.002

17. Chen, W., Feng, P., Yang, H., Ding, H., Lin, H. and Chou, K.C. (2018) iRNA-3typeA: Identifying 3-Types of Modification at RNA's Adenosine Sites. *Molecular Therapy*, **11**, 468-474. https://doi.org/10.1016/j.omtn.2018.03.012

18. Qiu, W.R., Sun, B.Q., Xiao, X., Xu, Z.C., Jia, J.H. and Chou, K.C. (2018) iKcr-PseEns: Identify Lysine Crotonylation Sites in Histone Proteins with Pseudo Components and Ensemble Classifier. *Genomics*, **110**, 239-246. https://doi.org/10.1016/j.ygeno.2017.10.008

19. 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

20. Hussain, W., Khan, S.D., Rasool, N., Khan, S.A. and Chou, K.C. (2019) SPalmitoylC-PseAAC: A Se-

quence-Based Model Developed via Chou's 5-Steps Rule and General PseAAC for Identifying S-palmitoylation Sites in Proteins. *Analytical Biochemistry*, **568**, 14-23. https://doi.org/10.1016/j.ab.2018.12.019

21. 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

22. Jia, J., Li, X., Qiu, W., Xiao, X. and Chou, K.C. (2019) iPPI-PseAAC(CGR): Identify Protein-Protein Interactions by Incorporating Chaos Game Representation into PseAAC. *Journal of Theoretical Biology*, **460**, 195-203. https://doi.org/10.1016/j.jtbi.2018.10.021

23. Khan, Y.D., Jamil, M., Hussain, W., Rasool, N., Khan, S.A. and Chou, K.C. (2019) pSS-bond-PseAAC: Prediction of Disulfide Bonding Sites by Integration of PseAAC and Statistical Moments. *Journal of Theoretical Biology*, **463**, 47-55. https://doi.org/10.1016/j.jtbi.2018.12.015

24. Lu, Y., Wang, S., Wang, J., Zhou, G., Zhang, Q., Zhou, X., Niu, B., Chen, Q. and Chou, K.C. (2019) An Epidemic Avian Influenza Prediction Model Based on Google Trends. *Letters in Organic Chemistry*, **16**, 303-310. https://doi.org/10.2174/1570178615666180724103325

25. Khan, Y.D., Batool, A., Rasool, N., Khan, A. and Chou, K.C. (2019) Prediction of Nitrosocysteine Sites Using Position and Composition Variant Features. *Letters in Organic Chemistry*, **16**, 283-293. https://doi.org/10.2174/1570178615666180802122953

26. Cheng, X., Xiao, X. and Chou, K.C. (2018) pLoc_bal-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

27. Li, J.X., Wang, S.Q., Du, Q.S., Wei, H., Li, X.M., Meng, J.Z., Wang, Q.Y., Xie, N.Z., Huang, R.B. and Chou, K.C. (2018) Simulated Protein Thermal Detection (SPTD) for Enzyme Thermostability Study and an Application Example for Pullulanase from *Bacillus deramificans*. *Current Pharmaceutical Design*, **24**, 4023-4033. https://doi.org/10.2174/1381612824666181113120948

28. Ghauri, A.W., Khan, Y.D., Rasool, N., Khan, S.A. and Chou, K.C. (2018) pNitro-Tyr-PseAAC: Predict Nitrotyrosine Sites in Proteins by Incorporating Five Features into Chou's General PseAAC. *Current Pharmaceutical Design*, **24**, 4034-4043. https://doi.org/10.2174/1381612825666181127101039

29. Xiao, X., Cheng, X., Chen, G., Mao, Q. and Chou, K.C. (2019) pLoc_bal-mGpos: Predict Subcellular Localization of Gram-Positive Bacterial Proteins by Quasi-Balancing Training Dataset and PseAAC. *Genomics*, **111**, 886-892. https://doi.org/10.1016/j.ygeno.2018.05.017

30. 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*. https://doi.org/10.1093/bioinformatics/btz016

31. Chen, Z., Zhao, P., Li, F., Marquez-Lago, T.T., Leier, A., Revote, J., Zhu, Y., Powell, D.R., Akutsu, T., Webb, G.I., Chou, K.C., Smith, A.I., Daly, R.J., Li, J. and Song, J. (2019) iLearn: An Integrated Platform and Meta-Learner for Eature Engineering, Machine-Learning Analysis and Modeling of DNA, RNA and Protein Sequence Data, Brief. *Bioinform*. https://doi.org/10.1093/bib/bbz041

32. Zhang, Y., Xie, R., Wang, J., Leier, A., Marquez-Lago, T.T., Akutsu, T., Webb, G.I., Chou, K.C. and Song, J. (2018) Computational Analysis and Prediction of Lysine Malonylation Sites by Exploiting Informative Features in an Integrative Machine-Learning Framework, Brief. *Bioinform*. https://doi.org/10.1093/bib/bby079

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. *Bioinform*, **20**, 638-658. https://doi.org/10.1093/bib/bby028

34. Song, J., Li, F., Takemoto, K., Haffari, G., Akutsu, T., Chou, K.C. and Webb, G.I. (2018) PREvaIL, an Integrative Approach for Inferring Catalytic Residues Using Sequence, Structural and Network Features in a Machine Learning Framework. *Journal of Theoretical Biology*, **443**, 125-137. https://doi.org/10.1016/j.jtbi.2018.01.023

35. Li, F., Wang, Y., Li, C., Marquez-Lago, T.T., Leier, A., Rawlings, N.D., Haffari, G., Revote, J., Akutsu, T., Chou, K.C., Purcell, A.W., Pike, R.N., Webb, G.I., Ian Smith, A., Lithgow, T., Daly, R.J., Whisstock, J.C. and Song, J. (2018) Twenty Years of Bioinformatics, Research for Protease-Specific Substrate and Cleavage Site Prediction: A Comprehensive Revisit and Benchmarking of Existing Methods, Brief. *Bioinform*. https://doi.org/10.1093/bib/bby077

36. Li, F., Li, C., Marquez-Lago, T.T., Leier, A., Akutsu, T., Purcell, A.W., Smith, A.I., Lightow, T., Daly, R.J., Song, J. and Chou, K.C. (2018) Quokka: A Comprehensive Tool for Rapid and Accurate Prediction of Kinase Family-Specific Phosphorylation Sites in the Human Proteome. *Bioinformatics*, **34**, 4223-4231. https://doi.org/10.1093/bioinformatics/bty522

37. 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

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 Sciences*, **2**, 1090-1103. https://doi.org/10.4236/ns.2010.210136

39. Zhai, X., Chen, M. and Lu, W. (2018) Accelerated Search for Perovskite Materials with Higher Curie Temperature Based on the Machine Learning Methods. *Computational Materials Science*, **151**, 41-48. https://doi.org/10.1016/j.commatsci.2018.04.031

40. 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

41. 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

42. Shen, H.B. and Chou, K.C. (2009) Gpos-mPLoc: A Top-Down Approach to Improve the Quality of Predicting Subcellular Localization of Gram-Positive Bacterial Proteins. *Protein & Peptide Letters*, **16**, 1478-1484. https://doi.org/10.2174/092986609789839322

43. Chou, K.C. and Shen, H.B. (2010) A New Method for Predicting the Subcellular Localization of Eukaryotic Proteins with Both Single and Multiple Sites: Euk-mPLoc 2.0. *PLoS ONE*, **5**, e9931. https://doi.org/10.1371/journal.pone.0009931

44. Chou, K.C. and Shen, H.B. (2010) Plant-mPLoc: A Top-Down Strategy to Augment the Power for Predicting Plant Protein Subcellular Localization. *PLoS ONE*, **5**, e11335. https://doi.org/10.1371/journal.pone.0011335

45. Shen, H.B. and Chou, K.C. (2010) Gneg-mPLoc: A Top-Down Strategy to Enhance the Quality of Predicting Subcellular Localization of Gram-Negative Bacterial Proteins. *Journal of Theoretical Biology*, **264**, 326-333. https://doi.org/10.1016/j.jtbi.2010.01.018

46. 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*, **28**, 175-186. https://doi.org/10.1080/07391102.2010.10507351

47. Chou, K.C. (2001) Prediction of Protein Cellular Attributes Using Pseudo Amino Acid Composition. *Proteins*, **43**, 246-255. (Erratum: ibid., 2001, Vol. 44, 60) https://doi.org/10.1002/prot.1035

48. 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

49. 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

50. Wu, Z.C., Xiao, X. and Chou, K.C. (2011) iLoc-Plant: A Multi-Label Classifier for Predicting the Subcellular Localization of Plant Proteins with Both Single and Multiple Sites. *Molecular BioSystems*, **7**, 3287-3297. https://doi.org/10.1039/c1mb05232b

51. Xiao, X., Wu, Z.C. and Chou, K.C. (2011) iLoc-Virus: A Multi-Label Learning Classifier for Identifying the Subcellular Localization of Virus Proteins with Both Single and Multiple Sites. *Journal of Theoretical Biology*, **284**, 42-51. https://doi.org/10.1016/j.jtbi.2011.06.005

52. 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

53. Xiao, X., Wu, Z.C. and Chou, K.C. (2011) A Multi-Label Classifier for Predicting the Subcellular Localization of Gram-Negative Bacterial Proteins with Both Single and Multiple Sites. *PLoS ONE*, **6**, e20592. https://doi.org/10.1371/journal.pone.0020592

54. Wu, Z.C., Xiao, X. and Chou, K.C. (2012) iLoc-Gpos: A Multi-Layer Classifier for Predicting the Subcellular Localization of Singleplex and Multiplex Gram-Positive Bacterial Proteins. *Protein & Peptide Letters*, **19**, 4-14. https://doi.org/10.2174/092986612798472839

55. Lin, W.Z., Fang, J.A., Xiao, X. and Chou, K.C. (2013) iLoc-Animal: A Multi-Label Learning Classifier for Predicting Subcellular Localization of Animal Proteins. *Molecular BioSystems*, **9**, 634-644. https://doi.org/10.1039/c3mb25466f

56. Cheng, X., Xiao, X. and Chou, K.C. (2017) pLoc-mPlant: Predict Subcellular Localization of Multi-Location Plant Proteins via Incorporating the Optimal GO Information into General PseAAC. *Molecular BioSystems*, **13**, 1722-1727. https://doi.org/10.1039/C7MB00267J

57. Cheng, X., Xiao, X. and Chou, K.C. (2017) pLoc-mVirus: Predict Subcellular Localization of Multi-Location Virus Proteins via Incorporating the Optimal GO Information into General PseAAC. *Gene*, **628**, 315-321. (Erratum: ibid., 2018, Vol. 644, 156-156) https://doi.org/10.1016/j.gene.2017.10.042

58. Cheng, X., Zhao, S.G., Lin, W.Z., Xiao, X. and Chou, K.C. (2017) pLoc-mAnimal: Predict Subcellular Localization of Animal Proteins with Both Single and Multiple Sites. *Bioinformatics*, **33**, 3524-3531. https://doi.org/10.1093/bioinformatics/btx476

59. Xiao, X., Cheng, X., Su, S., Nao, Q. and Chou, K.C. (2017) pLoc-mGpos: Incorporate Key Gene Ontology Information into General PseAAC for Predicting Subcellular Localization of Gram-Positive Bacterial Proteins. *Natural Sciences*, **9**, 331-349. https://doi.org/10.4236/ns.2017.99032

60. Cheng, X., Xiao, X. and Chou, K.C. (2017) pLoc-mEuk: Predict Subcellular Localization of Multi-Label Eukaryotic Proteins by Extracting the Key GO Information into General PseAAC. *Genomics*, **110**, 50-58. https://doi.org/10.1016/j.ygeno.2017.08.005

61. Cheng, X., Xiao, X. and Chou, K.C. (2018) pLoc-mGneg: Predict Subcellular Localization of Gram-Negative Bacterial Proteins by Deep Gene Ontology Learning via General PseAAC. *Genomics*, **110**, 231-239. https://doi.org/10.1016/j.ygeno.2017.10.002

62. 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

63. 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

64. Chou, K.C., Cheng, X. and Xiao, X. (2018) pLoc_bal-mHum: Predict Subcellular Localization of Human Proteins by PseAAC and Quasi-Balancing Training Dataset. *Genomics*. https://doi.org/10.1016/j.ygeno.2018.08.007

65. Cheng, X., Lin, W.Z., Xiao, X. and Chou, K.C. (2019) pLoc_bal-mAnimal: Predict Subcellular Localization of Animal Proteins by Balancing Training Dataset and PseAAC. *Bioinformatics*, **35**, 398-406. https://doi.org/10.1093/bioinformatics/bty628

66. 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

67. Chou, K.C. and Elrod, D.W. (2002) Bioinformatical Analysis of G-Protein-Coupled Receptors. *Journal of Proteome Research*, **1**, 429-433. https://doi.org/10.1021/pr025527k

68. Chen, W., Lin, H., Feng, P.M., Ding, C., Zuo, Y.C. and Chou, K.C. (2012) iNuc-PhysChem: A Sequence-Based Predictor for Identifying Nucleosomes via Physicochemical Properties. *PLoS ONE*, **7**, e47843. https://doi.org/10.1371/journal.pone.0047843

69. Xu, Y., Ding, J., Wu, L.Y. and Chou, K.C. (2013) iSNO-PseAAC: Predict Cysteine S-Nitrosylation Sites in Proteins by Incorporating Position Specific Amino Acid Propensity into Pseudo Amino Acid Composition. *PLoS ONE*, **8**, e55844. https://doi.org/10.1371/journal.pone.0055844

70. Cai, Y.D. and Chou, K.C. (2004) Predicting Subcellular Localization of Proteins in a Hybridization Space. *Bioinformatics*, **20**, 1151-1156. https://doi.org/10.1093/bioinformatics/bth054

71. Chou, K.C. and Cai, Y.D. (2006) Prediction of Protease Types in a Hybridization Space. *Biochemical and Biophysical Research Communications*, **339**, 1015-1020. https://doi.org/10.1016/j.bbrc.2005.10.196

72. Lin, W.Z., Fang, J.A., Xiao, X. and Chou, K.C. (2011) iDNA-Prot: Identification of DNA Binding Proteins Using Random Forest with Grey Model. *PLoS ONE*, **6**, e24756. https://doi.org/10.1371/journal.pone.0024756

73. Kandaswamy, K.K., Chou, K.C., Martinetz, T., Moller, S., Suganthan, P.N., Sridharan, S. and Pugalenthi, G. (2011) AFP-Pred: A Random Forest Approach for Predicting Antifreeze Proteins from Sequence-Derived Properties. *Journal of Theoretical Biology*, **270**, 56-62. https://doi.org/10.1016/j.jtbi.2010.10.037

74. Chou, K.C. (2015) Impacts of Bioinformatics, to Medicinal Chemistry. *Medicinal Chemistry*, **11**, 218-234. https://doi.org/10.2174/1573406411666141229162834

75. Fang, Y., Guo, Y., Feng, Y. and Li, M. (2008) Predicting DNA-Binding Proteins: Approached from Chou's Pseudo Amino Acid Composition and Other Specific Sequence Features. *Amino Acids*, **34**, 103-109. https://doi.org/10.1007/s00726-007-0568-2

76. 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

77. Chen, C., Chen, L., Zou, X. and Cai, P. (2009) Prediction of Protein Secondary Structure Content by Using the Concept of Chou's Pseudo Amino Acid Composition and Support Vector Machine. *Protein & Peptide Letters*, **16**, 27-31. https://doi.org/10.2174/092986609787049420

78. Lin, H., Wang, H., Ding, H., Chen, Y.L. and Li, Q.Z. (2009) Prediction of Subcellular Localization of Apoptosis Protein Using Chou's Pseudo Amino Acid Composition. *Acta Biotheoretica*, **57**, 321-330. https://doi.org/10.1007/s10441-008-9067-4

79. Esmaeili, M., Mohabatkar, H. and Mohsenzadeh, S. (2010) Using the Concept of Chou's Pseudo Amino Acid Composition for Risk Type Prediction of Human Papillomaviruses. *Journal of Theoretical Biology*, **263**, 203-209. https://doi.org/10.1016/j.jtbi.2009.11.016

80. 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

81. Qiu, J.D., Huang, J.H., Shi, S.P. and Liang, R.P. (2010) Using the Concept of Chou's Pseudo Amino Acid Composition to Predict Enzyme Family Classes: An Approach with Support Vector Machine Based on Discrete Wavelet Transform. *Protein & Peptide Letters*, **17**, 715-722. https://doi.org/10.2174/092986610791190372

82. Sahu, S.S. and Panda, G. (2010) A Novel Feature Representation Method Based on Chou's Pseudo Amino Acid Composition for Protein Structural Class Prediction. *Computational Biology and Chemistry*, **34**, 320-327. https://doi.org/10.1016/j.compbiolchem.2010.09.002

83. Yu, L., Guo, Y., Li, Y., Li, G., Li, M., Luo, J., Xiong, W. and Qin, W. (2010) SecretP: Identifying Bacterial Secreted Proteins by Fusing New Features into Chou's Pseudo Amino Acid Composition. *Journal of Theoretical Biology*, **267**, 1-6. https://doi.org/10.1016/j.jtbi.2010.08.001

84. Guo, J., Rao, N., Liu, G., Yang, Y. and Wang, G. (2011) Predicting Protein Folding Rates Using the Concept of Chou's Pseudo Amino Acid Composition. *Journal of Computational Chemistry*, **32**, 1612-1617. https://doi.org/10.1002/jcc.21740

85. Li, J.N. and Wang, Y. (2011) Using a Novel AdaBoost Algorithm and Chou's Pseudo Amino Acid Composition for Predicting Protein Subcellular Localization. *Protein & Peptide Letters*, **18**, 1219-1225. https://doi.org/10.2174/092986611797642797

86. Mohammad, B.M., Behjati, M. and Mohabatkar, H. (2011) Prediction of Metalloproteinase Family Based on the Concept of Chou's Pseudo Amino Acid Composition Using a Machine Learning Approach. *Journal of Structural and Functional Genomics*, **12**, 191-197. https://doi.org/10.1007/s10969-011-9120-4

87. Zou, D., He, Z., He, J. and Xia, Y. (2011) Supersecondary Structure Prediction Using Chou's Pseudo Amino Acid Composition. *Journal of Computational Chemistry*, **32**, 271-278. https://doi.org/10.1002/jcc.21616

88. 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

89. Hayat, M. and Khan, A. (2012) Discriminating Outer Membrane Proteins with Fuzzy K-Nearest Neighbor Algorithms Based on the General Form of Chou's PseAAC. *Protein & Peptide Letters*, **19**, 411-421. https://doi.org/10.2174/092986612799789387

90. 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

91. Liao, B., Xiang, Q. and Li, D. (2012) Incorporating Secondary Features into the General Form of Chou's PseAAC for Predicting Protein Structural Class. *Protein & Peptide Letters*, **19**, 1133-1138. https://doi.org/10.2174/092986612803217051

92. Mei, S. (2012) Multi-Kernel Transfer Learning Based on Chou's PseAAC Formulation for Protein Submitochondria Localization. *Journal of Theoretical Biology*, **293**, 121-130. https://doi.org/10.1016/j.jtbi.2011.10.015

93. 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

94. Qin, Y.F., Wang, C.H., Yu, X.Q., Zhu, J., Liu, T.G. and Zheng, X.Q. (2012) Predicting Protein Structural Class by Incorporating Patterns of Overrepresented k-Mers into the General Form of Chou's PseAAC. *Protein & Peptide Letters*, **19**, 388-397. https://doi.org/10.2174/092986612799789350

95. Sun, X.Y., Shi, S.P., Qiu, J.D., Suo, S.B., Huang, S.Y. and Liang, R.P. (2012) Identifying Protein Quaternary

Structural Attributes by Incorporating Physicochemical Properties into the General Form of Chou's PseAAC via Discrete Wavelet Transform. *Molecular BioSystems*, **8**, 3178-3184. https://doi.org/10.1039/c2mb25280e

96. Zhao, X.W., Li, X.T., Ma, Z.Q. and Yin, M.H. (2012) Identify DNA-Binding Proteins with Optimal Chou's Amino Acid Composition. *Protein & Peptide Letters*, **19**, 398-405. https://doi.org/10.2174/092986612799789404

97. Zhao, X.W., Ma, Z.Q. and Yin, M.H. (2012) Predicting Protein-Protein Interactions by Combing Various Sequence-Derived Features into the General Form of Chou's Pseudo Amino Acid Composition. *Protein & Peptide Letters*, **19**, 492-500. https://doi.org/10.2174/092986612800191080

98. 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

99. Chang, T.H., Wu, L.C., Lee, T.Y., Chen, S.P., Huang, H.D. and Horng, J.T. (2013) EuLoc: A Web-Server for Accurately Predict Protein Subcellular Localization in Eukaryotes by Incorporating Various Features of Sequence Segments into the General Form of Chou's PseAAC. *Journal of Computer-Aided Molecular Design*, **27**, 91-103. https://doi.org/10.1007/s10822-012-9628-0

100. Fan, G.-L., Li, Q.-Z. and Zuo, Y.-C. (2013) Predicting Acidic and Alkaline Enzymes by Incorporating the Average Chemical Shift and Gene Ontology Informations into the General Form of Chou's PseAAC. *Process Biochemistry*, **48**, 1048-1053. https://doi.org/10.1016/j.procbio.2013.05.012

101. Fan, G.L. and Li, Q.Z. (2013) Discriminating Bioluminescent Proteins by Incorporating Average Chemical Shift and Evolutionary Information into the General Form of Chou's Pseudo Amino Acid Composition. *Journal of Theoretical Biology*, **334**, 45-51. https://doi.org/10.1016/j.jtbi.2013.06.003

102. 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

103. Mohabatkar, H., Beigi, M.M., Abdolahi, K. and Mohsenzadeh, S. (2013) Prediction of Allergenic Proteins by Means of the Concept of Chou's Pseudo Amino Acid Composition and a Machine Learning Approach. *Medicinal Chemistry*, **9**, 133-137. https://doi.org/10.2174/157340613804488341

104. 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

105. Sarangi, A.N., Lohani, M. and Aggarwal, R. (2013) Prediction of Essential Proteins in Prokaryotes by Incorporating Various Physico-Chemical Features into the General form of Chou's Pseudo Amino Acid Composition. *Protein & Peptide Letters*, **20**, 781-795. https://doi.org/10.2174/0929866511320070008

106. Wang, X., Li, G.Z. and Lu, W.C. (2013) Virus-ECC-mPLoc: A Multi-Label Predictor for Predicting the Subcellular Localization of Virus Proteins with Both Single and Multiple Sites Based on a General Form of Chou's Pseudo Amino Acid Composition. *Protein & Peptide Letters*, **20**, 309-317. https://doi.org/10.2174/092986613804910608

107. Niu, X.H., *et al.* (2013) Using the Concept of Chou's Pseudo Amino Acid Composition to Predict Protein Solubility: An Approach with Entropies in Information Theory. *Journal of Theoretical Biology*, **332**, 211-217. https://doi.org/10.1016/j.jtbi.2013.03.010

108. Xie, H.L., Fu, L. and Nie, X.D. (2013) Using Ensemble SVM to Identify Human GPCRs N-Linked Glycosylation Sites Based on the General Form of Chou's PseAAC. *Protein Engineering, Design and Selection*, **26**, 735-742. https://doi.org/10.1093/protein/gzt042

109. 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

110. Hajisharifi, Z., Piryaiee, M., Mohammad Beigi, M., Behbahani, M. and Mohabatkar, H. (2014) Predicting Anticancer Peptides with Chou's Pseudo Amino Acid Composition and Investigating Their Mutagenicity via Ames Test. *Journal of Theoretical Biology*, **341**, 34-40. https://doi.org/10.1016/j.jtbi.2013.08.037

111. Han, G.S., Yu, Z.G. and Anh, V. (2014) A Two-Stage SVM Method to Predict Membrane Protein Types by Incorporating Amino Acid Classifications and Physicochemical Properties into a General Form of Chou's PseAAC. *Journal of Theoretical Biology*, **344**, 31-39. https://doi.org/10.1016/j.jtbi.2013.11.017

112. Hayat, M. and Iqbal, N. (2014) Discriminating Protein Structure Classes by Incorporating Pseudo Average Chemical Shift to Chou's General PseAAC and Support Vector Machine. *Computer Methods and Programs in Biomedicine*, **116**, 184-192. https://doi.org/10.1016/j.cmpb.2014.06.007

113. Jia, C., Li, X.N. and Wang, Z. (2014) Prediction of Protein S-Nitrosylation Sites Based on Adapted Normal Distribution Bi-Profile Bayes and Chou's Pseudo Amino Acid Composition. *International Journal of Molecular Sciences*, **15**, 10410-10423. https://doi.org/10.3390/ijms150610410

114. Li, L., Yu, S., Xiao, W., Li, Y., Li, M., Huang, L., Zheng, X., Zhou, S. and Yang, H. (2014) Prediction of Bacterial Protein Subcellular Localization by Incorporating Various Features into Chou's PseAAC and a Backward Feature Selection Approach. *Biochimie*, **104**, 100-107. https://doi.org/10.1016/j.biochi.2014.06.001

115. Mondal, S. and Pai, P.P. (2014) Chou's Pseudo Amino Acid Composition Improves Sequence-Based Antifreeze Protein Prediction. *Journal of Theoretical Biology*, **356**, 30-35. https://doi.org/10.1016/j.jtbi.2014.04.006

116. Zhang, J., Zhao, X., Sun, P. and Ma, Z. (2014) PSNO: Predicting Cysteine S-Nitrosylation Sites by Incorporating Various Sequence-Derived Features into the General Form of Chou's PseAAC. *International Journal of Molecular Sciences*, **15**, 11204-11219. https://doi.org/10.3390/ijms150711204

117. Ahmad, S., Kabir, M. and Hayat, M. (2015) Identification of Heat Shock Protein Families and J-Protein Types by Incorporating Dipeptide Composition into Chou's General PseAAC. *Computer Methods and Programs in Biomedicine*, **122**, 165-174. https://doi.org/10.1016/j.cmpb.2015.07.005

118. Dehzangi, A., Heffernan, R., Sharma, A., Lyons, J., Paliwal, K. and Sattar, A. (2015) Gram-Positive and Gram-Negative Protein Subcellular Localization by Incorporating Evolutionary-Based Descriptors into Chou's General PseAAC. *Journal of Theoretical Biology*, **364**, 284-294. https://doi.org/10.1016/j.jtbi.2014.09.029

119. Khan, Z.U., Hayat, M. and M.Khan, A. (2015) Discrimination of Acidic and Alkaline Enzyme Using Chou's Pseudo Amino Acid Composition in Conjunction with Probabilistic Neural Network Model. *Journal of Theoretical Biology*, **365**, 197-203. https://doi.org/10.1016/j.jtbi.2014.10.014

120. Liu, B., Chen, J. and Wang, X. (2015) Protein Remote Homology Detection by Combining Chou's Distance-Pair Pseudo Amino Acid Composition and Principal Component Analysis. *Molecular Genetics and Genomics*, **290**, 1919-1931. https://doi.org/10.1007/s00438-015-1044-4

121. Liu, B., Xu, J., Fan, S., Xu, R., Zhou, J. and Wang, X. (2015) PseDNA-Pro: DNA-Binding Protein Identification by Combining Chou's PseAAC and Physicochemical Distance Transformation. *Molecular Informatics*, **34**, 8-17. https://doi.org/10.1002/minf.201400025

122. 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

123. Sanchez, V., Peinado, A.M., Perez-Cordoba, J.L. and Gomez, A.M. (2015) A New Signal Characterization and Signal-Based Chou's PseAAC Representation of Protein Sequences. *Journal of Bioinformatics and Computational Biology*, **13**, Article ID: 1550024. https://doi.org/10.1142/S0219720015500249

124. Sharma, R., Dehzangi, A., Lyons, J., Paliwal, K., Tsunoda, T. and Sharma, A. (2015) Predict Gram-Positive and

Gram-Negative Subcellular Localization via Incorporating Evolutionary Information and Physicochemical Features into Chou's General PseAAC. *IEEE Transactions on NanoBioscience*, **14**, 915-926. https://doi.org/10.1109/TNB.2015.2500186

125. Zhang, M., Zhao, B. and Li, X.U. (2015) Predicting Industrial Polymer Melt Index via Incorporating Chaotic Characters into Chou's General PseAAC. *Chemometrics and Intelligent Laboratory Systems* (*CHEMOLAB*), **146**, 232-240. https://doi.org/10.1016/j.chemolab.2015.05.028

126. Zhang, S.L. (2015) Accurate Prediction of Protein Structural Classes by Incorporating PSSS and PSSM into Chou's General PseAAC. *Chemometrics and Intelligent Laboratory Systems*, **142**, 28-35. https://doi.org/10.1016/j.chemolab.2015.01.004

127. Behbahani, M., Mohabatkar, H. and Nosrati, M. (2016) Analysis and Comparison of Lignin Peroxidases between Fungi and Bacteria Using Three Different Modes of Chou's General Pseudo Amino Acid Composition. *Journal of Theoretical Biology*, **411**, 1-5. https://doi.org/10.1016/j.jtbi.2016.09.001

128. Jiao, Y.S. and Du, P.F. (2016) Prediction of Golgi-Resident Protein Types Using General Form of Chou's Pseudo Amino Acid Compositions: Approaches with Minimal Redundancy Maximal Relevance Feature Selection. *Journal of Theoretical Biology*, **402**, 38-44. https://doi.org/10.1016/j.jtbi.2016.04.032

129. Ju, Z., Cao, J.Z. and Gu, H. (2016) Predicting Lysine Phosphoglycerylation with Fuzzy SVM by Incorporating k-Spaced Amino Acid Pairs into Chou's General PseAAC. *Journal of Theoretical Biology*, **397**, 145-150. https://doi.org/10.1016/j.jtbi.2016.02.020

130. Kabir, M. and Hayat, M. (2016) iRSpot-GAEnsC: Identifing Recombination Spots via Ensemble Classifier and Extending the Concept of Chou's PseAAC to Formulate DNA Samples. *Molecular Genetics and Genomics*, **291**, 285-296. https://doi.org/10.1007/s00438-015-1108-5

131. Tahir, M. and Hayat, M. (2016) iNuc-STNC: A Sequence-Based Predictor for Identification of Nucleosome Positioning in Genomes by Extending the Concept of SAAC and Chou's PseAAC. *Molecular BioSystems*, **12**, 2587-2593. https://doi.org/10.1039/C6MB00221H

132. Tiwari, A.K. (2016) Prediction of G-Protein Coupled Receptors and Their Subfamilies by Incorporating Various Sequence Features into Chou's General PseAAC. *Computer Methods and Programs in Biomedicine*, **134**, 197-213. https://doi.org/10.1016/j.cmpb.2016.07.004

133. Ju, Z. and He, J.J. (2017) Prediction of Lysine Propionylation Sites Using Biased SVM and Incorporating Four Different Sequence Features into Chou's PseAAC. *Journal of Molecular Graphics and Modelling*, **76**, 356-363. https://doi.org/10.1016/j.jmgm.2017.07.022

134. Ju, Z. and He, J.J. (2017) Prediction of Lysine Crotonylation Sites by Incorporating the Composition of k-Spaced Amino Acid Pairs into Chou's General PseAAC. *Journal of Molecular Graphics and Modelling*, **77**, 200-204. https://doi.org/10.1016/j.jmgm.2017.08.020

135. Khan, M., Hayat, M., Khan, S.A. and Iqbal, N. (2017) Unb-DPC: Identify Mycobacterial Membrane Protein Types by Incorporating Un-Biased Dipeptide Composition into Chou's General PseAAC. *Journal of Theoretical Biology*, **415**, 13-19. https://doi.org/10.1016/j.jtbi.2016.12.004

136. Liang, Y. and Zhang, S. (2017) Predict Protein Structural Class by Incorporating Two Different Modes of Evolutionary Information into Chou's General Pseudo Amino Acid Composition. *Journal of Molecular Graphics and Modelling*, **78**, 110-117. https://doi.org/10.1016/j.jmgm.2017.10.003

137. Meher, P.K., Sahu, T.K., Saini, V. and Rao, A.R. (2017) Predicting Antimicrobial Peptides with Improved Accuracy by Incorporating the Compositional, Physico-Chemical and Structural Features into Chou's General PseAAC. *Scientific Reports*, **7**, Article No. 42362. https://doi.org/10.1038/srep42362

138. Qiu, W.R., Zheng, Q.S., Sun, B.Q. and Xiao, X. (2017) Multi-iPPseEvo: A Multi-Label Classifier for Identifying

Human Phosphorylated Proteins by Incorporating Evolutionary Information into Chou's General PseAAC via Grey System Theory. *Molecular Informatics*, **36**, UNSP 1600085. https://doi.org/10.1002/minf.201600085

139. Tripathi, P. and Pandey, P.N. (2017) A Novel Alignment-Free Method to Classify Protein Folding Types by Combining Spectral Graph Clustering with Chou's Pseudo Amino Acid Composition. *Journal of Theoretical Biology*, **424**, 49-54. https://doi.org/10.1016/j.jtbi.2017.04.027

140. Xu, C., Ge, L., Zhang, Y., Dehmer, M. and Gutman, I. (2017) Prediction of Therapeutic Peptides by Incorporating q-Wiener Index into Chou's General PseAAC. *Journal of Biomedical Informatics*.

141. Yu, B., Li, S., Qiu, W.Y., Chen, C., Chen, R.X., Wang, L., Wang, M.H. and Zhang, Y. (2017) Accurate Prediction of Subcellular Location of Apoptosis Proteins Combining Chou's PseAAC and PsePSSM Based on Wavelet Denoising. *Oncotarget*, **8**, 107640-107665. https://doi.org/10.18632/oncotarget.22585

142. Yu, B., Lou, L., Li, S., Zhang, Y., Qiu, W., Wu, X., Wang, M. and Tian, B. (2017) Prediction of Protein Structural Class for Low-Similarity Sequences Using Chou's Pseudo Amino Acid Composition and Wavelet Denoising. *Journal of Molecular Graphics and Modelling*, **76**, 260-273. https://doi.org/10.1016/j.jmgm.2017.07.012

143. Ahmad, J. and Hayat, M. (2018) MFSC: Multi-Voting Based Feature Selection for Classification of Golgi Proteins by Adopting the General Form of Chou's PseAAC Components. *Journal of Theoretical Biology*, **463**, 99-109. https://doi.org/10.1016/j.jtbi.2018.12.017

144. Akbar, S. and Hayat, M. (2018) iMethyl-STTNC: Identification of N(6)-Methyladenosine Sites by Extending the Idea of SAAC into Chou's PseAAC to Formulate RNA Sequences. *Journal of Theoretical Biology*, **455**, 205-211. https://doi.org/10.1016/j.jtbi.2018.07.018

145. Arif, M., Hayat, M. and Jan, Z. (2018) iMem-2LSAAC: A Two-Level Model for Discrimination of Membrane Proteins and Their Types by Extending the Notion of SAAC into Chou's Pseudo Amino Acid Composition. *Journal of Theoretical Biology*, **442**, 11-21. https://doi.org/10.1016/j.jtbi.2018.01.008

146. Butt, A.H., Rasool, N. and Khan, Y.D. (2018) Predicting Membrane Proteins and Their Types by Extracting Various Sequence Features into Chou's General PseAAC. *Molecular Biology Reports*.

147. Contreras-Torres, E. (2018) Predicting Structural Classes of Proteins by Incorporating Their Global and Local Physicochemical and Conformational Properties into General Chou's PseAAC. *Journal of Theoretical Biology*, **454**, 139-145. https://doi.org/10.1016/j.jtbi.2018.05.033

148. Fu, X., Zhu, W., Liso, B., Cai, L., Peng, L. and Yang, J. (2018) Improved DNA-Binding Protein Identification by Incorporating Evolutionary Information into the Chou's PseAAC. *IEEE Access*, **20**.

149. Javed, F. and Hayat, M. (2018) Predicting Subcellular Localizations of Multi-Label Proteins by Incorporating the Sequence Features into Chou's PseAAC. *Genomics*. https://doi.org/10.1016/j.ygeno.2018.09.004

150. Krishnan, M.S. (2018) Using Chou's General PseAAC to Analyze the Evolutionary Relationship of Receptor Associated Proteins (RAP) with Various Folding Patterns of Protein Domains. *Journal of Theoretical Biology*, **445**, 62-74. https://doi.org/10.1016/j.jtbi.2018.02.008

151. Liang, Y. and Zhang, S. (2018) Identify Gram-Negative Bacterial Secreted Protein Types by Incorporating Different Modes of PSSM into Chou's General PseAAC via Kull-Back-Leibler Divergence. *Journal of Theoretical Biology*, **454**, 22-29. https://doi.org/10.1016/j.jtbi.2018.05.035

152. Mousavizadegan, M. and Mohabatkar, H. (2018) Computational Prediction of Antifungal Peptides via Chou's PseAAC and SVM. *Journal of Bioinformatics and Computational Biology*, **16**, Article ID: 1850016. https://doi.org/10.1142/S0219720018500166

153. Qiu, W., Li, S., Cui, X., Yu, Z., Wang, M., Du, J., Peng, Y. and Yu, B. (2018) Predicting Protein Submitochondrial Locations by Incorporating the Pseudo-Position Specific Scoring Matrix into the General Chou's Pseudo-Amino Acid Composition. *Journal of Theoretical Biology*, **450**, 86-103.

https://doi.org/10.1016/j.jtbi.2018.04.026

154. Rahman, S.M., Shatabda, S., Saha, S., Kaykobad, M. and Sohel Rahman, M. (2018) DPP-PseAAC: A DNA-Binding Protein Prediction Model Using Chou's General PseAAC. *Journal of Theoretical Biology*, **452**, 22-34. https://doi.org/10.1016/j.jtbi.2018.05.006

155. Sankari, E.S. and Manimegalai, D.D. (2018) Predicting Membrane Protein Types by Incorporating a Novel Feature Set into Chou's General PseAAC. *Journal of Theoretical Biology*, **455**, 319-328. https://doi.org/10.1016/j.jtbi.2018.07.032

156. Srivastava, A., Kumar, R. and Kumar, M. (2018) BlaPred: Predicting and Classifying Beta-Lactamase Using a 3-Tier Prediction System via Chou's General PseAAC. *Journal of Theoretical Biology*, **457**, 29-36. https://doi.org/10.1016/j.jtbi.2018.08.030

157. Zhang, S. and Duan, X. (2018) Prediction of Protein Subcellular Localization with Oversampling Approach and Chou's General PseAAC. *Journal of Theoretical Biology*, **437**, 239-250. https://doi.org/10.1016/j.jtbi.2017.10.030

158. Zhang, S. and Liang, Y. (2018) Predicting Apoptosis Protein Subcellular Localization by Integrating Auto-Cross Correlation and PSSM into Chou's PseAAC. *Journal of Theoretical Biology*, **457**, 163-169. https://doi.org/10.1016/j.jtbi.2018.08.042

159. Adilina, S., Farid, D.M. and Shatabda, S. (2019) Effective DNA Binding Protein Prediction by Using Key Features via Chou's General PseAAC. *Journal of Theoretical Biology*, **460**, 64-78. https://doi.org/10.1016/j.jtbi.2018.10.027

160. Ahmad, J. and Hayat, M. (2019) MFSC: Multi-Voting Based Feature Selection for Classification of Golgi Proteins by Adopting the General Form of Chou's PseAAC Components. *Journal of Theoretical Biology*, **463**, 99-109. https://doi.org/10.1016/j.jtbi.2018.12.017

161. Awais, M., Hussain, W., Khan, Y.D., Rasool, N., Khan, S.A. and Chou, K.C. (2019) iPhosH-PseAAC: Identify Phosphohistidine Sites in Proteins by Blending Statistical Moments and Position Relative Features According to the Chou's 5-Step Rule and General Pseudo Amino Acid Composition. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. https://doi.org/10.1109/TCBB.2019.2919025

162. Butt, A.H., Rasool, N. and Khan, Y.D. (2019) Prediction of Antioxidant Proteins by Incorporating Statistical Moments Based Features into Chou's PseAAC. *Journal of Theoretical Biology*, **473**, 1-8. https://doi.org/10.1016/j.jtbi.2019.04.019

163. Chen, G., Cao, M., Yu, J., Guo, X. and Shi, S. (2019) Prediction and Functional Analysis of Prokaryote Lysine Acetylation Site by Incorporating Six Types of Features into Chou's General PseAAC. *Journal of Theoretical Biology*, **461**, 92-101. https://doi.org/10.1016/j.jtbi.2018.10.047

164. 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

165. Kabir, M., Ahmad, S., Iqbal, M. and Hayat, M. (2019) iNR-2L: A Two-Level Sequence-Based Predictor Developed via Chou's 5-Steps Rule and General PseAAC for Identifying Nuclear Receptors and Their Families. *Genomics*. https://doi.org/10.1016/j.ygeno.2019.02.006

166. Ning, Q., Ma, Z. and Zhao, X. (2019) dForml(KNN)-PseAAC: Detecting Formylation Sites from Protein Sequences Using K-Nearest Neighbor Algorithm via Chou's 5-Step Rule and Pseudo Components. *Journal of Theoretical Biology*, **470**, 43-49. https://doi.org/10.1016/j.jtbi.2019.03.011

167. Shen, Y., Tang, J. and Guo, F. (2019) Identification of Protein Subcellular Localization via Integrating Evolutionary and Physicochemical Information into Chou's General PseAAC. *Journal of Theoretical Biology*, **462**,

230-239. https://doi.org/10.1016/j.jtbi.2018.11.012

168. Tahir, M., Hayat, M. and Khan, S.A. (2019) iNuc-ext-PseTNC: An Efficient Ensemble Model for Identification of Nucleosome Positioning by Extending the Concept of Chou's PseAAC to Pseudo-Tri-Nucleotide Composition. *Molecular Genetics and Genomics*, **294**, 199-210. https://doi.org/10.1007/s00438-018-1498-2

169. Tian, B., Wu, X., Chen, C., Qiu, W., Ma, Q. and Yu, B. (2019) Predicting Protein-Protein Interactions by Fusing Various Chou's Pseudo Components and Using Wavelet Denoising Approach. *Journal of Theoretical Biology*, **462**, 329-346. https://doi.org/10.1016/j.jtbi.2018.11.011

170. Wang, L., Zhang, R. and Mu, Y. (2019) Fu-SulfPred: Identification of Protein S-Sulfenylation Sites by Fusing Forests via Chou's General PseAAC. *Journal of Theoretical Biology*, **461**, 51-58. https://doi.org/10.1016/j.jtbi.2018.10.046

171. 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

172. 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

173. 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

174. Wang, J., Yang, B., Revote, J., Leier, A., Marquez-Lago, T.T., Webb, G., Song, J., Chou, K.C. and Lithgow, T. (2017) POSSUM: A Bioinformatics, Toolkit for Generating Numerical Sequence Feature Descriptors Based on PSSM Profiles. *Bioinformatics*, **33**, 2756-2758. https://doi.org/10.1093/bioinformatics/btx302

175. 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

176. Liu, B., Li, F.U., Fang, L., Wang, X. and Chou, K.C. (2014) repDNA: A Python Package to Generate Various Modes of Feature Vectors for DNA Sequences by Incorporating User-Defined Physicochemical Properties and Sequence-Order Effects. *Bioinformatics*, **31**, 1307-1309. https://doi.org/10.1093/bioinformatics/btu820

177. Liu, B., Li, F.U., Fang, L., Wang, X. and Chou, K.C. (2016) repRNA: A Web Server for Generating Various Feature Vectors of RNA Sequences. *Molecular Genetics and Genomics*, **291**, 473-481. https://doi.org/10.1007/s00438-015-1078-7

178. Liu, B., Li, F.U., Wang, X., Chen, J., Fang, L. and Chou, K.C. (2015) Pse-in-One: A Web Server for Generating Various Modes of Pseudo Components of DNA, RNA, and Protein Sequences. *Nucleic Acids Research*, **43**, W65-W71. https://doi.org/10.1093/nar/gkv458

179. 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

180. Chen, W., Feng, P.M., Lin, H. and Chou, K.C. (2014) iSS-PseDNC: Identifying Splicing Sites Using Pseudo Dinucleotide Composition. *Biomed Research International* (*BMRI*), **2014**, Article ID: 623149. https://doi.org/10.1155/2014/623149

181. Chen, W., Tang, H., Ye, J., Lin, H. and Chou, K.C. (2016) iRNA-PseU: Identifying RNA Pseudouridine Sites. *Molecular Therapy, Nucleic Acids*, **5**, e332.

182. Liu, B., Long, R. and Chou, K.C. (2016) iDHS-EL: Identifying DNase I Hypersensitivesites by Fusing Three Different Modes of Pseudo Nucleotide Composition into an Ensemble Learning Framework. *Bioinformatics*, **32**,

2411-2418. https://doi.org/10.1093/bioinformatics/btw186

183. Feng, P., Ding, H., Yang, H., Chen, W., Lin, H. and Chou, K.C. (2017) iRNA-PseColl: Identifying the Occurrence Sites of Different RNA Modifications by Incorporating Collective Effects of Nucleotides into PseKNC. *Molecular Therapy Nucleic Acids*, **7**, 155-163. https://doi.org/10.1016/j.omtn.2017.03.006

184. 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

185. 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

186. Al Maruf, M.A. and Shatabda, S. (2018) iRSpot-SF: Prediction of Recombination Hotspots by Incorporating Sequence Based Features into Chou's Pseudo Components. *Genomics.* https://doi.org/10.1016/j.ygeno.2018.06.003

187. Sabooh, M.F., Iqbal, N., Khan, M., Khan, M. and Maqbool, H.F. (2018) Identifying 5-Methylcytosine Sites in RNA Sequence Using Composite Encoding Feature into Chou's PseKNC. *Journal of Theoretical Biology*, **452**, 1-9. https://doi.org/10.1016/j.jtbi.2018.04.037

188. Zhang, L. and Kong, L. (2018) iRSpot-ADPM: Identify Recombination Spots by Incorporating the Associated Dinucleotide Product Model into Chou's Pseudo Components. *Journal of Theoretical Biology*, **441**, 1-8. https://doi.org/10.1016/j.jtbi.2017.12.025

189. Zhang, L. and Kong, L. (2019) iRSpot-PDI: Identification of Recombination Spots by Incorporating Dinucleotide Property Diversity Information into Chou's Pseudo Components. *Genomics*, **111**, 457-464. https://doi.org/10.1016/j.ygeno.2018.03.003

190. Liu, B., Wu, H. and Chou, K.C. (2017) Pse-in-One 2.0: An Improved Package of Web Servers for Generating Various Modes of Pseudo Components of DNA, RNA, and Protein Sequences. *Natural Sciences*, **9**, 67-91. https://doi.org/10.4236/ns.2017.94007

191. Shen, H.B. and Chou, K.C. (2005) Using Optimized Evidence-Theoretic K-Nearest Neighbor Classifier and Pseudo Amino Acid Composition to Predict Membrane Protein Types. *Biochemical and Biophysical Research Communications*, **334**, 288-292. https://doi.org/10.1016/j.bbrc.2005.06.087

192. Chou, K.C. and Shen, H.B. (2007) Euk-mPLoc: A Fusion Classifier for Large-Scale Eukaryotic Protein Subcellular Location Prediction by Incorporating Multiple Sites. *Journal of Proteome Research*, **6**, 1728-1734. https://doi.org/10.1021/pr060635i

193. Shen, H.B. and Chou, K.C. (2009) QuatIdent: A Web Server for Identifying Protein Quaternary Structural Attribute by Fusing Functional Domain and Sequential Evolution Information. *Journal of Proteome Research*, **8**, 1577-1584. https://doi.org/10.1021/pr800957q

194. Chou, K.C. (2001) Prediction of Protein Signal Sequences and Their Cleavage Sites. *Proteins*: *Structure*, *Function and Genetics*, **42**, 136-139. https://doi.org/10.1002/1097-0134(20010101)42:1<136::AID-PROT130>3.0.CO;2-F

195. 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

196. Chou, K.C. (2001) Prediction of Signal Peptides Using Scaled Window. *Peptides*, **22**, 1973-1979. https://doi.org/10.1016/S0196-9781(01)00540-X

197. 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*, **15**, 1746-1766. https://doi.org/10.3390/ijms15021746

198. Xu, Y., Wen, X., Wen, L.S., Wu, L.Y., Deng, N.Y. and Chou, K.C. (2014) iNitro-Tyr: Prediction of Nitrotyrosine Sites in Proteins with General Pseudo Amino Acid Composition. *PLoS ONE*, **9**, e105018. https://doi.org/10.1371/journal.pone.0105018

199. Chen, W., Feng, P., Ding, H., Lin, H. and Chou, K.C. (2015) iRNA-methyl: Identifying N6-Methyladenosine Sites Using Pseudo Nucleotide Composition. *Analytical Biochemistry*, **490**, 26-33. https://doi.org/10.1016/j.ab.2015.08.021

200. Jia, J., Liu, Z., Xiao, X. and Chou, K.C. (2015) iPPI-Esml: An Ensemble Classifier for Identifying the Interactions of Proteins by Incorporating Their Physicochemical Properties and Wavelet Transforms into PseAAC. *Journal of Theoretical Biology*, **377**, 47-56. https://doi.org/10.1016/j.jtbi.2015.04.011

201. Chen, W., Ding, H., Feng, P., Lin, H. and Chou, K.C. (2016) iACP: A Sequence-Based Tool for Identifying Anticancer Peptides. *Oncotarget*, **7**, 16895-16909. https://doi.org/10.18632/oncotarget.7815

202. Chen, W., Feng, P., Ding, H., Lin, H. and Chou, K.C. (2016) Using Deformation Energy to Analyze Nucleosome Positioning in Genomes. *Genomics*, **107**, 69-75. https://doi.org/10.1016/j.ygeno.2015.12.005

203. Jia, J., Liu, Z., Xiao, X., Liu, B. and Chou, K.C. (2016) pSuc-Lys: Predict Lysine Succinylation Sites in Proteins with PseAAC and Ensemble Random Forest Approach. *Journal of Theoretical Biology*, **394**, 223-230. https://doi.org/10.1016/j.jtbi.2016.01.020

204. Chou, K.C. and Zhang, C.T. (1995) Review: Prediction of Protein Structural Classes. *Critical Reviews in Biochemistry and Molecular Biology*, **30**, 275-349. https://doi.org/10.3109/10409239509083488

205. Zhou, G.P. and Assa-Munt, N. (2001) Some Insights into Protein Structural Class Prediction. *Proteins*: *Structure*, *Function and Genetics*, **44**, 57-59. https://doi.org/10.1002/prot.1071

206. Zhou, G.P. and Doctor, K. (2003) Subcellular Location Prediction of Apoptosis Proteins. *Proteins*: *Structure*, *Function and Genetics*, **50**, 44-48. https://doi.org/10.1002/prot.10251

207. Khan and Zia-ur-Rehman, A. (2012) Identifying GPCRs and Their Types with Chou's Pseudo Amino Acid Composition: An Approach from Multi-Scale Energy Representation and Position Specific Scoring Matrix. *Protein & Peptide Letters*, **19**, 890-903. https://doi.org/10.2174/092986612801619589

208. Huang, C. and Yuan, J.Q. (2013) Predicting Protein Subchloroplast Locations with Both Single and Multiple Sites via Three Different Modes of Chou's Pseudo Amino Acid Compositions. *Journal of Theoretical Biology*, **335**, 205-212. https://doi.org/10.1016/j.jtbi.2013.06.034

209. Chou, K.C. and Shen, H.B. (2009) Recent Advances in Developing Web-Servers for Predicting Protein Attributes. *Natural Sciences*, **1**, 63-92. https://doi.org/10.4236/ns.2009.12011

210. Xu, Y., Shao, X.J., Wu, L.Y., Deng, N.Y. and Chou, K.C. (2013) iSNO-AAPair: Incorporating Amino Acid Pairwise Coupling into PseAAC for Predicting Cysteine S-Nitrosylation Sites in Proteins. *PeerJ*, **1**, e171. https://doi.org/10.7717/peerj.171

211. Liu, B., Xu, J., Lan, X., Xu, R., Zhou, J., Wang, X. and Chou, K.C. (2014) iDNA-Protdis: Identifying DNA-Binding Proteins by Incorporating Amino Acid Distance-Pairs and Reduced Alphabet Profile into the General Pseudo Amino Acid Composition. *PLoS ONE*, **9**, e106691. https://doi.org/10.1371/journal.pone.0106691

212. Xu, Y., Wen, X., Shao, X.J., Deng, N.Y. and Chou, K.C. (2014) iHyd-PseAAC: Predicting Hydroxyproline and Hydroxylysine in Proteins by Incorporating Dipeptide Position-Specific Propensity into Pseudo Amino Acid Composition. *International Journal of Molecular Sciences*, **15**, 7594-7610. https://doi.org/10.3390/ijms15057594

213. Qiu, W.R., Xiao, X., Lin, W.Z. and Chou, K.C. (2014) iMethyl-PseAAC: Identification of Protein Methylation Sites via a Pseudo Amino Acid Composition Approach. *BioMed Research International*, **2014**, Article ID: 947416. https://doi.org/10.1155/2014/947416

214. Fan, Y.N., Xiao, X., Min, J.L. and Chou, K.C. (2014) iNR-Drug: Predicting the Interaction of Drugs with Nuc-

lear Receptors in Cellular Networking. *International Journal of Molecular Sciences*, **15**, 4915-4937. https://doi.org/10.3390/ijms15034915

215. Guo, S.H., Deng, E.Z., Xu, L.Q., Ding, H., Lin, H., Chen, W. and Chou, K.C. (2014) iNuc-PseKNC: A Sequence-Based Predictor for Predicting Nucleosome Positioning in Genomes with Pseudo k-Tuple Nucleotide Composition. *Bioinformatics*, **30**, 1522-1529. https://doi.org/10.1093/bioinformatics/btu083

216. Qiu, W.R., Xiao, X., Lin, W.Z. and Chou, K.C. (2015) iUbiq-Lys: Prediction of Lysine Ubiquitination Sites in Proteins by Extracting Sequence Evolution Information via a Grey System Model. *Journal of Biomolecular Structure and Dynamics*, **33**, 1731-1742. https://doi.org/10.1080/07391102.2014.968875

217. Liu, B., Fang, L., Li, F.U., Wang, X., Chen, J. and Chou, K.C. (2015) Identification of Real microRNA Precursors with a Pseudo Structure Status Composition Approach. *PLoS ONE*, **10**, e0121501. https://doi.org/10.1371/journal.pone.0121501

218. Liu, Z., Xiao, X., Yu, D.J., Jia, J., Qiu, W.R. and Chou, K.C. (2016) pRNAm-PC: Predicting N-Methyladenosine Sites in RNA Sequences via Physical-Chemical Properties. *Analytical Biochemistry*, **497**, 60-67. https://doi.org/10.1016/j.ab.2015.12.017

219. Xiao, X., Ye, H.X., Liu, Z., Jia, J.H. and Chou, K.C. (2016) iROS-gPseKNC: Predicting Replication Origin Sites in DNA by Incorporating Dinucleotide Position-Specific Propensity into General Pseudo Nucleotide Composition. *Oncotarget*, **7**, 34180-34189. https://doi.org/10.18632/oncotarget.9057

220. Qiu, W.R., Sun, B.Q., Xiao, X., Xu, Z.C. and Chou, K.C. (2016) iPTM-mLys: Identifying Multiple Lysine PTM Sites and Their Different Types. *Bioinformatics*, **32**, 3116-3123. https://doi.org/10.1093/bioinformatics/btw380

221. Jia, J., Liu, Z., Xiao, X., Liu, B. and Chou, K.C. (2016) iPPBS-Opt: A Sequence-Based Ensemble Classifier for Identifying Protein-Protein Binding Sites by Optimizing Imbalanced Training Datasets. *Molecules*, **21**, E95. https://doi.org/10.3390/molecules21010095

222. Qiu, W.R., Xiao, X., Xu, Z.C. and Chou, K.C. (2016) iPhos-PseEn: Identifying Phosphorylation Sites in Proteins by Fusing Different Pseudo Components into an Ensemble Classifier. *Oncotarget*, **7**, 51270-51283. https://doi.org/10.18632/oncotarget.9987

223. Zhang, C.J., Tang, H., Li, W.C., Lin, H., Chen, W. and Chou, K.C. (2016) iOri-Human: Identify Human Origin of Replication by Incorporating Dinucleotide Physicochemical Properties into Pseudo Nucleotide Composition. *Oncotarget*, **7**, 69783-69793. https://doi.org/10.18632/oncotarget.11975

224. Liu, B., Fang, L., Li, F.U., Wang, X. and Chou, K.C. (2016) iMiRNA-PseDPC: microRNA Precursor Identification with a Pseudo Distance-Pair Composition Approach. *Journal of Biomolecular Structure and Dynamics*, **34**, 223-235. https://doi.org/10.1080/07391102.2015.1014422

225. Qiu, W.R., Sun, B.Q., Xiao, X., Xu, Z.C. and Chou, K.C. (2016) iHyd-PseCp: Identify Hydroxyproline and Hydroxylysine in Proteins by Incorporating Sequence-Coupled Effects into General PseAAC. *Oncotarget*, **7**, 44310-44321. https://doi.org/10.18632/oncotarget.10027

226. Jia, J., Liu, Z., Xiao, X., Liu, B. and Chou, K.C. (2016) Identification of Protein-Protein Binding Sites by Incorporating the Physicochemical Properties and Stationary Wavelet Transforms into Pseudo Amino Acid Composition (iPPBS-PseAAC). *Journal of Biomolecular Structure and Dynamics*, **34**, 1946-1961. https://doi.org/10.1080/07391102.2015.1095116

227. Jia, J., Liu, Z., Xiao, X., Liu, B. and Chou, K.C. (2016) iCar-PseCp: Identify Carbonylation Sites in Proteins by Monto Carlo Sampling and Incorporating Sequence Coupled Effects into General PseAAC. *Oncotarget*, **7**, 34558-34570. https://doi.org/10.18632/oncotarget.9148

228. Chen, J., Long, R., Wang, X.L., Liu, B. and Chou, K.C. (2016) dRHP-PseRA: Detecting Remote Homology Proteins Using Profile-Based Pseudo Protein Sequence and Rank Aggregation. *Scientific Reports*, **6**, Article No.

32333. https://doi.org/10.1038/srep32333

229. Liu, B., Wu, H., Zhang, D., Wang, X. and Chou, K.C. (2017) Pse-Analysis: A Python Package for DNA/RNA and Protein/Peptide Sequence Analysis Based on Pseudo Components and Kernel Methods. *Oncotarget*, **8**, 13338-13343. https://doi.org/10.18632/oncotarget.14524

230. Qiu, W.R., Jiang, S.Y., Xu, Z.C., Xiao, X. and Chou, K.C. (2017) iRNAm5C-PseDNC: Identifying RNA 5-Methylcytosine Sites by Incorporating Physical-Chemical Properties into Pseudo Dinucleotide Composition. *Oncotarget*, **8**, 41178-41188. https://doi.org/10.18632/oncotarget.17104

231. Qiu, W.R., Jiang, S.Y., Sun, B.Q., Xiao, X., Cheng, X. and Chou, K.C. (2017) iRNA-2methyl: Identify RNA 2'-O-Methylation Sites by Incorporating Sequence-Coupled Effects into General PseKNC and Ensemble Classifier. *Medicinal Chemistry*, **13**, 734-743. https://doi.org/10.2174/1573406413666170623082245

232. Xu, Y., Li, C. and Chou, K.C. (2017) iPreny-PseAAC: Identify C-Terminal Cysteine Prenylation Sites in Proteins by Incorporating Two Tiers of Sequence Couplings into PseAAC. *Medicinal Chemistry*, **13**, 544-551. https://doi.org/10.2174/1573406413666170419150052

233. Qiu, W.R., Sun, B.Q., Xiao, X., Xu, D. and Chou, K.C. (2017) iPhos-PseEvo: Identifying Human Phosphorylated Proteins by Incorporating Evolutionary Information into General PseAAC via Grey System Theory. *Molecular Informatics*, **36**, UNSP 1600010. https://doi.org/10.1002/minf.201600010

234. Li, L., Xu, Y. and Chou, K.C. (2017) iPGK-PseAAC: Identify Lysine Phosphoglycerylation Sites in Proteins by Incorporating Four Different Tiers of Amino Acid Pairwise Coupling Information into the General PseAAC. *Medicinal Chemistry*, **13**, 552-559. https://doi.org/10.2174/1573406413666170515120507

235. Cheng, X., Zhao, S.G., Xiao, X. and Chou, K.C. (2017) iATC-mISF: A Multi-Label Classifier for Predicting the Classes of Anatomical Therapeutic Chemicals. *Bioinformatics*, **33**, 341-346. (Corrigendum, ibid., 2017, Vol. 33, 2610) https://doi.org/10.1093/bioinformatics/btx387

236. Cheng, X., Zhao, S.G., Xiao, X. and Chou, K.C. (2017) iATC-mHyb: A Hybrid Multi-Label Classifier for Predicting the Classification of Anatomical Therapeutic Chemicals. *Oncotarget*, **8**, 58494-58503. https://doi.org/10.18632/oncotarget.17028

237. Wang, J., Yang, B., Leier, A., Marquez-Lago, T.T., Hayashida, M., Rocker, A., Yanju, Z., Akutsu, T., Chou, K.C., Strugnell, R.A., Song, J. and Lithgow, T. (2018) Bastion6: A Bioinformatics, Approach for Accurate Prediction of Type VI Secreted Effectors. *Bioinformatics*, **34**, 2546-2555. https://doi.org/10.1093/bioinformatics/bty155

238. Liu, B., Li, K., Huang, D.S. and Chou, K.C. (2018) iEnhancer-EL: Identifying Enhancers and Their Strength with Ensemble Learning Approach. *Bioinformatics*, **34**, 3835-3842. https://doi.org/10.1093/bioinformatics/bty458

239. Chen, Z., Zhao, P.Y., Li, F., Leier, A., Marquez-Lago, T.T., Wang, Y., Webb, G.I., Smith, A.I., Daly, R.J., Chou, K.C. and Song, J. (2018) iFeature: A Python Package and Web Server for Features Extraction and Selection from Protein and Peptide Sequences. *Bioinformatics*, **34**, 2499-2502. https://doi.org/10.1093/bioinformatics/bty140

240. Su, Z.D., Huang, Y., Zhang, Z.Y., Zhao, Y.W., Wang, D., Chen, W., Chou, K.C. and Lin, H. (2018) iLoc-lncRNA: Predict the Subcellular Location of lncRNAs by Incorporating Octamer Composition into General PseKNC. *Bioinformatics*, **34**, 4196-4204.

241. Liu, B., Yang, F., Huang, D.S. and Chou, K.C. (2018) iPromoter-2L: A Two-Layer Predictor for Identifying Promoters and Their Types by Multi-Window-Based PseKNC. *Bioinformatics*, **34**, 33-40. https://doi.org/10.1093/bioinformatics/btx579

242. Liu, B., Weng, F., Huang, D.S. and Chou, K.C. (2018) iRO-3wPseKNC: Identify DNA Replication Origins by Three-Window-Based PseKNC. *Bioinformatics*, **34**, 3086-3093. https://doi.org/10.1093/bioinformatics/bty312

243. Yang, H., Qiu, W.R., Liu, G., Guo, F.B., Chen, W., Chou, K.C. and Lin, H. (2018) iRSpot-Pse6NC: Identifying Recombination Spots in *Saccharomyces cerevisiae* by Incorporating Hexamer Composition into General

PseKNC. *International Journal of Biological Sciences*, **14**, 883-891. https://doi.org/10.7150/ijbs.24616

244. Song, J., Li, F., Leier, A., Marquez-Lago, T.T., Akutsu, T., Haffari, G., Chou, K.C., Webb, G.I. and Pike, R.N. (2018) PROSPERous: High-Throughput Prediction of Substrate Cleavage Sites for 90 Proteases with Improved Accuracy. *Bioinformatics*, **34**, 684-687. https://doi.org/10.1093/bioinformatics/btx670

245. Chou, K.C. (2005) Review: Progress in Protein Structural Class Prediction and Its Impact to Bioinformatics, and Proteomics. *Current Protein & Peptide Science*, **6**, 423-436. https://doi.org/10.2174/138920305774329368

246. Zhong, W.Z. and Zhou, S.F. (2014) Molecular Science for Drug Development and Biomedicine. *International Journal of Molecular Sciences*, **15**, 20072-20078. https://doi.org/10.3390/ijms151120072

247. Du, Q.S., Huang, R.B., Wang, S.Q. and Chou, K.C. (2010) Designing Inhibitors of M2 Proton Channel against H1N1 Swine Influenza Virus. *PLoS ONE*, **5**, e9388. https://doi.org/10.1371/journal.pone.0009388

248. Wang, S.Q., Cheng, X.C., Dong, W.L., Wang, R.L. and Chou, K.C. (2010) Three New Powerful Oseltamivir Derivatives for Inhibiting the Neuraminidase of Influenza Virus. *Biochemical and Biophysical Research Communications*, **401**, 188-191. https://doi.org/10.1016/j.bbrc.2010.09.020

249. Li, X.B., Wang, S.Q., Xu, W.R., Wang, R.L. and Chou, K.C. (2011) Novel Inhibitor Design for Hemagglutinin against H1N1 Influenza Virus by Core Hopping Method. *PLoS ONE*, **6**, e28111. https://doi.org/10.1371/journal.pone.0028111

250. Ma, Y., Wang, S.Q., Xu, W.R., Wang, R.L. and Chou, K.C. (2012) Design Novel Dual Agonists for Treating Type-2 Diabetes by Targeting Peroxisome Proliferator-Activated Receptors with Core Hopping Approach. *PLoS ONE*, **7**, e38546. https://doi.org/10.1371/journal.pone.0038546

251. Liu, L., Ma, Y., Wang, R.L., Xu, W.R., Wang, S.Q. and Chou, K.C. (2013) Find Novel Dual-Agonist Drugs for Treating Type 2 Diabetes by Means of Cheminformatics. *Drug Design*, *Development and Therapy*, **7**, 279-287. https://doi.org/10.2147/DDDT.S42113

252. Lu, J.J., Pan, W., Hu, Y.J. and Wang, Y.T. (2012) Multi-Target Drugs: The Trend of Drug Research and Development. *PLoS ONE*, **7**, e40262. https://doi.org/10.1371/journal.pone.0040262

253. Chou, K.C. and Forsen, S. (1980) Diffusion-Controlled Effects in Reversible Enzymatic Fast Reaction System: Critical Spherical Shell and Proximity Rate Constants. *Biophysical Chemistry*, **12**, 255-263. https://doi.org/10.1016/0301-4622(80)80002-0

254. Chou, K.C., Li, T.T. and Forsen, S. (1980) The Critical Spherical Shell in Enzymatic Fast Reaction Systems. *Biophysical Chemistry*, **12**, 265-269. https://doi.org/10.1016/0301-4622(80)80003-2

255. Li, T.T., Chou, K.C. and Forsen, S. (1980) The Flow of Substrate Molecules in Fast Enzyme-Catalyzed Reaction Systems. *Chemica Scripta*, **16**, 192-196.

256. Chou, K.C. and Forsen, S. (1980) Graphical Rules for Enzyme-Catalyzed Rate Laws. *Biochemical Journal*, **187**, 829-835. https://doi.org/10.1042/bj1870829

257. Chou, K.C., Forsen, S. and Zhou, G.Q. (1980) Three Schematic Rules for Deriving Apparent Rate Constants. *Chemica Scripta*, **16**, 109-113.

258. Chou, K.C., Carter, R.E. and Forsen, S. (1981) A New Graphical Method for Deriving Rate Equations for Complicated Mechanisms. *Chemica Scripta*, **18**, 82-86.

259. Chou, K.C. and Forsen, S. (1981) Graphical Rules of Steady-State Reaction Systems. *Canadian Journal of Chemistry*, **59**, 737-755. https://doi.org/10.1139/v81-107

260. Chou, K.C., Chen, N.Y. and Forsen, S. (1981) The Biological Functions of Low-Frequency Phonons: 2. Cooperative Effects. *Chemica Scripta*, **18**, 126-132.

261. Chou, K.C., Jiang, S.P., Li, W. and Fee, C.H. (1979) Graph Theory of Enzyme Kinetics: 1. Steady-State Reaction

System. *Scientia Sinica*, **22**, 341-358.

262. Zhou, G.P. and Deng, M.H. (1984) An Extension of Chou's Graphic Rules for Deriving Enzyme Kinetic Equations to Systems Involving Parallel Reaction Pathways. *Biochemical Journal*, **222**, 169-176. https://doi.org/10.1042/bj2220169

263. Chou, K.C. (1989) Graphic Rules in Steady and Non-Steady Enzyme Kinetics. *The Journal of Biological Chemistry*, **264**, 12074-12079.

264. Chou, K.C. (1990) Review: Applications of Graph Theory to Enzyme Kinetics and Protein Folding Kinetics. Steady and Non-Steady State Systems. *Biophysical Chemistry*, **35**, 1-24. https://doi.org/10.1016/0301-4622(90)80056-D

265. Althaus, I.W., Chou, J.J., Gonzales, A.J., Diebel, M.R., Chou, K.C., Kezdy, F.J., Romero, D.L., Aristoff, P.A., Tarpley, W.G. and Reusser, F. (1993) Steady-State Kinetic Studies with the Non-Nucleoside HIV-1 Reverse Transcriptase Inhibitor U-87201E. *The Journal of Biological Chemistry*, **268**, 6119-6124.

266. Althaus, I.W., Gonzales, A.J., Chou, J.J., Diebel, M.R., Chou, K.C., Kezdy, F.J., Romero, D.L., Aristoff, P.A., Tarpley, W.G. and Reusser, F. (1993) The Quinoline U-78036 Is a Potent Inhibitor of HIV-1 Reverse Transcriptase. *The Journal of Biological Chemistry*, **268**, 14875-14880.

267. Althaus, I.W., Chou, J.J., Gonzales, A.J., Diebel, M.R., Chou, K.C., Kezdy, F.J., Romero, D.L., Aristoff, P.A., Tarpley, W.G. and Reusser, F. (1993) Kinetic Studies with the Nonnucleoside HIV-1 Reverse Transcriptase Inhibitor U-88204E. *Biochemistry*, **32**, 6548-6554. https://doi.org/10.1021/bi00077a008

268. Althaus, I.W., Chou, J.J., Gonzales, A.J., Diebel, M.R., Chou, K.C., Kezdy, F.J., Romero, D.L., Aristoff, P.A., Tarpley, W.G. and Reusser, F. (1994) Steady-State Kinetic Studies with the Polysulfonate U-9843, an HIV Reverse Transcriptase Inhibitor. *Cellular and Molecular Life Sciences* (*Experientia*), **50**, 23-28. https://doi.org/10.1007/BF01992044

269. Althaus, I.W., Chou, J.J., Gonzales, A.J., Diebel, M.R., Chou, K.C., Kezdy, F.J., Romero, D.L., Thomas, R.C., Aristoff, P.A., Tarpley, W.G. and Reusser, F. (1994) Kinetic Studies with the Non-Nucleoside Human Immunodeficiency Virus Type-1 Reverse Transcriptase Inhibitor U-90152e. *Biochemical Pharmacology*, **47**, 2017-2028. https://doi.org/10.1016/0006-2952(94)90077-9

270. Chou, K.C., Kezdy, F.J. and Reusser, F. (1994) Review: Kinetics of Processive Nucleic Acid Polymerases and Nucleases. *Analytical Biochemistry*, **221**, 217-230. https://doi.org/10.1006/abio.1994.1405

271. Althaus, I.W., Chou, K.C., Franks, K.M., Diebel, M.R., Kezdy, F.J., Romero, D.L., Thomas, R.C., Aristoff, P.A., Tarpley, W.G. and Reusser, F. (1996) The Benzylthio-Pyrididine U-31,355, a Potent Inhibitor of HIV-1 Reverse Transcriptase. *Biochemical Pharmacology*, **51**, 743-750. https://doi.org/10.1016/0006-2952(95)02390-9

272. Andraos, J. (2008) Kinetic Plasticity and the Determination of Product Ratios for Kinetic Schemes Leading to Multiple Products without Rate Laws: New Methods Based on Directed Graphs. *Canadian Journal of Chemistry*, **86**, 342-357. https://doi.org/10.1139/v08-020

273. Chou, K.C. and Shen, H.B. (2009) FoldRate: A Web-Server for Predicting Protein Folding Rates from Primary Sequence. *The Open Bioinformatics Journal*, **3**, 31-50. https://doi.org/10.2174/1875036200903010031

274. Shen, H.B., Song, J.N. and Chou, K.C. (2009) Prediction of Protein Folding Rates from Primary Sequence by Fusing Multiple Sequential Features. *Journal of Biomedical Science and Engineering*, **2**, 136-143. https://doi.org/10.4236/jbise.2009.23024

275. Chou, K.C. (2010) Graphic Rule for Drug Metabolism Systems. *Current Drug Metabolism*, **11**, 369-378. https://doi.org/10.2174/138920010791514261

276. Chou, K.C., Lin, W.Z. and Xiao, X. (2011) Wenxiang: A Web-Server for Drawing Wenxiang Diagrams. *Natural Sciences*, **3**, 862-865. https://doi.org/10.4236/ns.2011.310111

277. Zhou, G.P. (2011) The Disposition of the LZCC Protein Residues in Wenxiang Diagram Provides New Insights into the Protein-Protein Interaction Mechanism. *Journal of Theoretical Biology*, **284**, 142-148. https://doi.org/10.1016/j.jtbi.2011.06.006

278. Chou, K.C. (2019) Proposing Pseudo Amino Acid Components Is an Important Milestone for Proteome and Genome Analyses. *International Journal of Peptide Research and Therapeutics* (*IJPRT*), No. 2, 1085-1098. https://doi.org/10.1007/s10989-019-09910-7

279. Chou, K.C. (2019) Impacts of Pseudo Amino Acid Components and 5-Steps Rule to Proteomics and Proteome Analysis. *Current Topics Medicinal Chemistry*, **19**, 2283-2300. https://doi.org/10.2174/1568026619666191018100141

280. Chou, K.C. (2019) An Insightful 10-Year Recollection since the Emergence of the 5-Steps Rule. *Current Pharmaceutical Design*, **25**, 4223-4234. https://doi.org/10.2174/1381612825666191129164042

281. Chou, K.C. (2019) An Insightful 20-Year Recollection since the Birth of Pseudo Amino Acid Components. *Applied Biochemistry and Biotechnology* (*ABAB*). (In Press) https://doi.org/10.1007/s00726-020-02828-1

282. Chou, K.C. (2019) An Insightful Recollection since the Birth of Gordon Life Science Institute about 17 Years Ago. *Advancement in Scientific and Engineering Research*, **4**, 17-30. https://doi.org/10.33495/aser_v4i2.19.105

283. Chou, K.C. (2019) An Insightful Recollection since the Distorted Key Theory Was Born about 23 Years Ago. *International Journal of Peptide Research and Therapeutics* (*IJPRT*). (In Press) https://doi.org/10.1016/j.ygeno.2019.09.001

284. Chou, K.C. (2019) Recent Progresses in Predicting Protein Subcellular Localization with Artificial Intelligence Tools Developed via the 5-Steps Rule. *Genomics*. (In Press)

285. Chou, K.C. (2019) Two Kinds of Metrics for Computational Biology. *Genomics*. (In Press)