Japanese Journal of Gastroenterology and Hepatology

Research Article

Recent Progresses in Predicting Protein Subcellular Localization with Artificial Intelligence (AI) Tools Developed *Via* the 5-Steps Rule

Chou CK[1,2*]

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

[2]Department of Physics,Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu, 610054, China

Received: 13 Oct 2019
Accepted: 02 Nov 2019
Published: 07 Nov 2019

***Corresponding to:**
Kuo-Chen Chou, Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu, 610054, China, E-mail: kchou@san.rr.com

1. Abstract

With the avalanche of protein sequences generated in the post-genomic age, it is highly desired to develop AI tools for rapidly and effectively identifying the subcellular locations of uncharacterized proteins based on their sequences information alone. Actually, considerable progresses have been achieved in this regard. This review is focused on those methods, which have the capacity to deal with multi-label proteins that may simultaneously exist in two or more subcellular location sites. Protein molecules with this kind of characteristic are vitally important for finding multi-target drugs, a current hot trend in drug development. Focused in this review are also those methods that have use-friendly web-servers established so that the majority of experimental scientists can use them to get the desired results without the need to go through the detailed mathematics involved.

2. Keywords: Artificial intelligence (AI) tools;5-step rules; Multi-label proteins; Multi-target drugs; Global accuracy and metrics; Local accuracy and metrics; Absolute true rate; Web-server

3. Introduction

As demonstrated by a series of recent publications [1-34] and summarized in two comprehensive review papers [26,35], to develop a really useful Artificial Intelligence (AI) tool, one needs to follow Chou's 5-steps rule 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 AI tool established by observing the guidelines of Chou's 5-step rules have the following notable merits: (**1**) crystal clear in logic development, (**2**) completely transparent in operation, (**3**) easily to repeat the reported results by other investigators, (**4**) with high potential in stimulating other new AI tools, and (**5**) 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.

The present minireview was focused on some recent development about AI tools in protein subcellular prediction or proteomics. But note that the 5-steps rule can also be used to deal with many different areas, such as material science [36] and even commercial science (e.g., analyzing the effect of bank credit card versus mobile payment). The only difference between the biological science and other science 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.

4. 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 AI 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 [37]. Actually, it is these multiplex proteins that are of significance for in-depth understanding the biological processes in a living cell.

4.1.Four Series of AI Tools

In the last decade or so, a number of AI tools 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**) $\mathbb{X}$-mPLoc, (**2**) iLoc-$\mathbb{X}$, (**3**) pLoc-m$\mathbb{X}$, and (4) pLoc_bal-m$\mathbb{X}$, where the wildcard $\mathbb{X}$ may denote "Euk" (eukaryotic), "Hum" (human), "Animal", "Plant", "Virus", "Gneg" (Gram-negative bacterial), "Gpos" (Gram-positive bacterial) proteins, respectively, as formulated by

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

The protein samples in the $\mathbb{X}$-mPLoc series [38-43] were formulated by hybridizing the GO (Gene Ontology) information, FunD (Functional Domain) information, and PSSM (Sequential Evolutionary) information into the general PseAAC[35], which was extended from pseudo amino acid composition [44,45].

The protein samples in the iLoc-$\mathbb{X}$ series [46-52] were formulated by incorporating the GO information and PSSM information into the general PseAAC.

The protein samples in the pLoc-m$\mathbb{X}$ series [53-59] 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 [23,26,27,60-62] were formulated by further balancing out the protein samples used in pLoc-mX series.

As for the justification of using the GO information for predicting the subcellular localization of proteins, see Section 4 of a review paper [63], where an insightful analysis has been elaborated and there is no need to repeat here.

4.2.Benchmark Dataset

All the AI tools 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 [43].

4.3. Sample Formulation

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

$$\mathbf{P} = R_1R_2R_3R_4R_5R_6R_7\cdots R_L \quad (2)$$

Citation: Chou CK, Recent Progresses in Predicting Protein Subcellular Localization with Artificial Intelligence (AI) Tools Developed Via the 5-Steps Rule. Japanese Journal of Gastroenterology and Hepatology. 2019; v2(5):1-21.

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 [1,2], "Covariance Discriminant" or CD algorithm [64-66], "Nearest Neighbor" or NN algorithm [67,68], and "Random Forest" or RF algorithm [69,70]) can only handle vectors as elaborated in [71], convert the sequential expression of Eq.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 [44,45]. 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., [17,18,25,72-167] as well as a long list of references cited in [168]). Because it has been widely and increasingly used, four powerful open access soft-wares, called 'PseAAC' [169], 'PseAAC-Builder' [85], 'propy' [95], and 'PseAAC-General' [106], were established: the former three are for generating various modes of special PseAAC[170]; while the fourth one for those of general PseAAC[35], 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" [171] mode. Encouraged by the successes of using PseAACto 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 [172-175] that have proved very successful as well [11,176-187]. According to the concept of general PseAAC[35], any protein sequence can be formulated as a PseAAC vector given by

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

where T is a transpose operator, while the integer Ω is a parameter and its value as well as the components $\Psi_u (u = 1, 2, \cdots, \Omega)$ will depend on how to extract the desired information from the amino acid sequence ofP.

In the last decade or so, a number of AI tools 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 roughly classified into four series: (1) $\mathbb{X}$-mPLoc, (2) iLoc-$\mathbb{X}$, (3) pLoc-m$\mathbb{X}$, and (4) pLoc_bal-m$\mathbb{X}$, where the wildcard $\mathbb{X}$ may denote "Euk" (eukaryotic), "Hum" (human), "Animal", "Plant", "Virus", "Gneg" (Gram-negative bacterial), "Gpos" (Gram-positive bacterial) proteins, respectively, as formulated by

$$\mathbb{X} \in \begin{cases} \text{Euk} \\ \text{Hum} \\ \text{Animal} \\ \text{Plant} \\ \text{Virus} \\ \text{Gneg} \\ \text{Gpos} \end{cases} \tag{4}$$

The protein samples in the $\mathbb{X}$-mPLoc series [38-43] were formulated by hybridizing the GO (Gene Ontology) information, FunD (Functional Domain) information, and PSSM (Sequential Evolutionary) information into the general PseAAC of Eq.3.

The protein samples in the iLoc-$\mathbb{X}$ series [46-52] were formulated by incorporating the GO information and PSSM information into the general PseAAC.

The protein samples in the pLoc-m$\mathbb{X}$ series [53-59] 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 [23,26,27,60-62] 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 [63], where an insightful analysis has been elaborated and there is no need to repeat here.

4.4. Operation Engine

The operation engine for $\mathbb{X}$-mPLoc series was constructed by fusing an array of OET-KNN (Optimized Evidence-Theoretic K-Nearest Neighbor) classifiers [188-190].

The operation engine for iLoc-$\mathbb{X}$ series was the multi-labeled KNN (K-Nearest Neighbor) classifier [46].

The operation engine for the pLoc-m$\mathbb{X}$ and pLoc_bal-m$\mathbb{X}$ series was the ML-GKR (Multi-Label Gaussian kernel Regression) classifier [53].

4.5. Metrics and Cross-Validation

In order to objectively evaluate the prediction quality

of an AI tool, one needs to consider the following two issues. (**1**) What metrics should be used to quantitatively reflect the AI's quality? (**2**) What test approach should be adopted to score the metrics?

Different from the metrics used to measure the prediction quality of a single-label AI tool, the metrics for a multi-label AI are much more complicated. To quantitatively evaluate the power of a multi-label AI tool, 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 [63].

$$\begin{cases} \text{Aiming}\uparrow = \dfrac{1}{N^q}\displaystyle\sum_{k=1}^{N^q}\left(\frac{\|\mathbb{L}_k \cap \mathbb{L}_k^*\|}{\|\mathbb{L}_k^*\|}\right), & [0,1] \\ \text{Coverage}\uparrow = \dfrac{1}{N^q}\displaystyle\sum_{k=1}^{N^q}\left(\frac{\|\mathbb{L}_k \cap \mathbb{L}_k^*\|}{\|\mathbb{L}_k\|}\right), & [0,1] \\ \text{Accuracy}\uparrow = \dfrac{1}{N^q}\displaystyle\sum_{k=1}^{N^q}\left(\frac{\|\mathbb{L}_k \cap \mathbb{L}_k^*\|}{\|\mathbb{L}_k \cup \mathbb{L}_k^*\|}\right), & [0,1] \\ \text{Absolute true}\uparrow = \dfrac{1}{N^q}\displaystyle\sum_{k=1}^{N^q}\Delta(\mathbb{L}_k, \mathbb{L}_k^*), & [0,1] \\ \text{Absolute false}\downarrow = \dfrac{1}{N^q}\displaystyle\sum_{k=1}^{N^q}\left(\frac{\|\mathbb{L}_k \cup \mathbb{L}_k^*\| - \|\mathbb{L}_k \cap \mathbb{L}_k^*\|}{M}\right), & [1,0] \end{cases} \quad (5)$$

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, $\cup$ means the symbol for the "union" in the set theory, $\cap$ denotes the symbol for the "intersection", $\mathbb{L}_k$ denotes the subset that contains all the labels observed by experiments for the k-th tested sample, $\mathbb{L}_k^*$ represents the subset that contains all the labels predicted for the k-th sample, and

$$\Delta(\mathbb{L}_k, \mathbb{L}_k^*) = \begin{cases} 1, & \text{if all the labels in } \mathbb{L}_k^* \text{ are identical to those in } \mathbb{L}_k \\ 0, & \text{otherwise} \end{cases} \quad (6)$$

In Eq.5, the first four metrics with an upper arrow $\uparrow$ are called positive metrics, meaning that the larger the rate is thebetter the prediction quality will be; the 5th metrics with a down arrow $\downarrow$ is called negative metrics, implying just the opposite meaning. As we can see from Eq.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 Eq.5 reflect the quality of a multi-label AI tool 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 [63]. 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 Eq.5 are used to evaluate the prediction quality of a multi-label AI tool 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 four intuitive metrics that were derived [1,66] based on the symbols introduced for studying the cleavage sites of signal peptides [191-193]. The set of metrics are given below

$$\begin{cases} \text{Sn}(i) = 1 - \dfrac{N_+^+(i)}{N^+(i)} & 0 \le \text{Sn}(i) \le 1 \\ \text{Sp}(i) = 1 - \dfrac{N_+^-(i)}{N^-(i)} & 0 \le \text{Sp}(i) \le 1 \\ \text{Acc}(i) = 1 - \dfrac{N_-^+(i) + N_+^-(i)}{N^+(i) + N^-(i)} & 0 \le \text{Acc}(i) \le 1 \\ \text{MCC}(i) = \dfrac{1 - \left(\frac{N_-^+(i)}{N^+(i)} + \frac{N_+^-(i)}{N^-(i)}\right)}{\sqrt{\left(1 + \frac{N_+^-(i) - N_-^+(i)}{N^+(i)}\right)\left(1 + \frac{N_-^+(i) - N_+^-(i)}{N^-(i)}\right)}} - 1 & \le \text{MCC}(i) \le 1 \\ (i = 1, 2, \cdots, M) \end{cases} \quad (7)$$

where Sn, Sp, Acc, and MCC represent the sensitivity, specificity, accuracy, and Mathew's correlation coefficient, respectively (Chen et al., 2007), *i* denotes the *i*-th subcellular location (or subset) in the benchmark dataset, and M has exactly the same meaning as in Eq.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., [3,5,7,10,12,31,34,177,178,182,194-200]), the set of metrics in Eq.7 can be used to evaluate the prediction quality of a multi-label AI tool for the proteins in each of subcellular locations concerned (see, e.g., [55,59]), and hence is called the "set of metrics for local accuracy" or the "set of local metrics".

4.6. 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 [201]. Of these three, however, the jackknife test was deemed the least arbitrary that can always yield a unique result for a given benchmark dataset [202,203], as clearly elucidated in a comprehensive review paper [35] 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., [77,79,98,107,204-207]).

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

5. Web Servers

The last but not least important guideline in the 5-step rules is about the web-server. As pointed out in [208] and demonstrated in a series of recent publications (see, e.g., [2-12,14-16,171,177-182,194-198,200,209-243]), user-friendly and publicly accessible web-servers represent the future direction for developing practically more useful AI's tools. Actually, many practically useful web-servers have significantly increased the impacts of AI tools on medicinal chemistry [71], driving medicinal chemistry into an unprecedented revolution [168].

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

5.1. -mPLoc Series

This series contains six publically accessible web-servers:(**1**) "Euk-mPLoc" at http://www.csbio.sjtu.edu.cn/bioinf/euk-multi-2/[40] for predicting the subcellular localization of eukaryotic proteins. (**2**) "Hum-mPLoc" at http://www.csbio.sjtu.edu.cn/bioinf/hum-multi-2/[38] for predicting the subcellular localization of human proteins. (**3**) "Plant-mPLoc" at http://www.csbio.sjtu.edu.cn/bioinf/plant-multi/[41] for predicting the subcellular localization of plant proteins. (**4**) "Virus-mPLoc" at http://www.csbio.sjtu.edu.cn/bioinf/virus-multi/[43] for predicting the subcellular localization of virus proteins. (**5**) "Gneg-mPLoc" at http://www.csbio.sjtu.edu.cn/bioinf/Gneg-multi /[42] for predicting the subcellular localization of Gram-negative bacterial proteins. (**6**) "Gpos-mPLoc" at http://www.csbio.sjtu.edu.cn/bioinf/Gpos-multi/[39] for predicting subcellular localization of Gram-positive bacterial proteins.

The aforementioned six AI tools have also been integrated into a package called "Cell-PLoc" at http://www.csbio.sjtu.edu.cn/bioinf/Cell-PLoc/[202] and its updated version "Cell-PLoc 2.0" at http://www.csbio.sjtu.edu.cn/bioinf/Cell-PLoc-2/[203].

5.2. iLoc- Series

It contains the following seven AI tools. (**1**) "iLoc-Euk" at http://www.jci-bioinfo.cn/iLoc-Euk[46] for predicting the subcellular localization of eukaryotic proteins. (**2**) "iLoc-Hum" at http://www.jci-bioinfo.cn/iLoc-Hum[49] for predicting the subcellular localization of human proteins. (**3**) "iLoc-Animal" at http://www.jci-bioinfo.cn/iLoc-Animal[52] for predicting the subcellular localization of animal proteins. (**4**) "iLoc-Plant" at http://www.jci-bioinfo.cn/iLoc-Plant[47] for predicting the subcellular localization of plant proteins. (**5**) "iLoc-Virus" at http://www.jci-bioinfo.cn/iLoc-Virus[48] for predicting the subcellular localization of virus proteins. (**6**) "iLoc-Gneg" at http://www.jci-bioinfo.cn/iLoc-Gneg[50] for predicting the subcellular localization of Gram-negative proteins. (**7**) "iLoc-Gpos" at http://www.jci-bioinfo.cn/iLoc-Gpos[51].

5.3. pLoc-m Series

There are seven AI tools in this series as listed below. (**1**) "pLoc-mEuk" at http://www.jci-bioinfo.cn/pLoc-mEuk/[57] for predicting the subcellular localization of eukaryotic proteins. (**2**) "pLoc-mHum" at http://www.jci-bioinfo.cn/pLoc-mHum/[59] for predicting the subcellular localization of human proteins. (**3**) "pLoc-

mAnimal" at http://www.jci-bioinfo.cn/pLoc-mAnimal/ [55] for predicting the subcellular localization of animal proteins. (**4**) "pLoc-mPlant" at http://www.jci-bioinfo.cn/pLoc-mPlant/[53] for predicting the subcellular localization of plant proteins. (**5**) "pLoc-mVirus" at http://www.jci-bioinfo.cn/pLoc-mVirus/[54] for predicting the subcellular localization of virus proteins. (**6**) "pLoc-mGneg" at http://www.jci-bioinfo.cn/pLoc-mGneg/[58] for predicting the subcellular localization of Gram-negative proteins. (**7**) "pLoc-mGpos" at http://www.jci-bioinfo.cn/pLoc-mGpos/[56] for predicting the subcellular localization of Gram-positive proteins.

5.4. pLoc_bal-m$^{\mathbb{X}}$Series

There are seven AI tools in this series as listed below. (**1**) "pLoc_bal-mEuk" at [26]. (**2**) "pLoc_bal-mHum" [26]. (**3**) "pLoc_bal-mAnimal" [62]. (**4**) "pLoc_bal-mPlant" [23]. (**5**) "pLoc_bal-mVirus" [27]. (**6**) "pLoc_bal-mGneg" [60]. (**7**) "pLoc_bal-mGpos" [27].

Listed in (Table 1) are the global accuracy rates (cf. Eq.5) predicted with the aforementioned seven IA tools, while the corresponding the local accuracy rates (cf. Eq.7) are given in (**Table 2**). As shown from the rates in the two tables, all the seven AI tools have yielded very high prediction quality in both the global and local cases. By means of these AI tools 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 [26] as a showcase. (**1**) Click the link at http://www.jci-bioinfo.cn/pLoc_bal-mEuk/, you'll see the top page of the AI tool 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 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 AI tool 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 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."

Table 1: List of the five global metrics rates reported from each of the seven predictors in the pLoc_bal-m$^{\mathbb{X}}$ series.

No	Predictor[a]	Aiming[b]	Coverage[b]	Accuracy[b]	Absolute true[b]	Absolute false[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%

aSee Eq.1 of Section 2 for further explanation.
bSee Eq.5 for the definition of the global metrics.

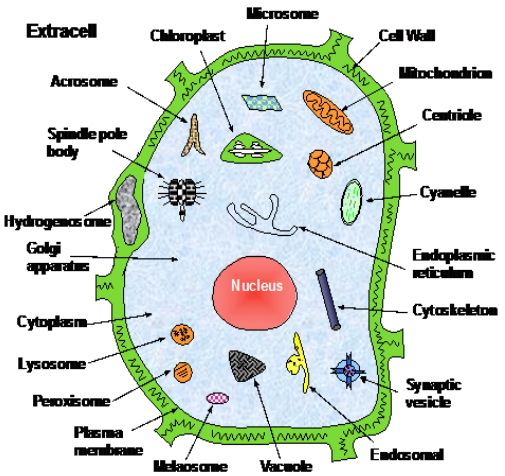

Figure 1: Schematic illustration to show the 22 organelles or subcellular locations in a eukaryotic cell. Adapted from Chou and Shen with permission [189].

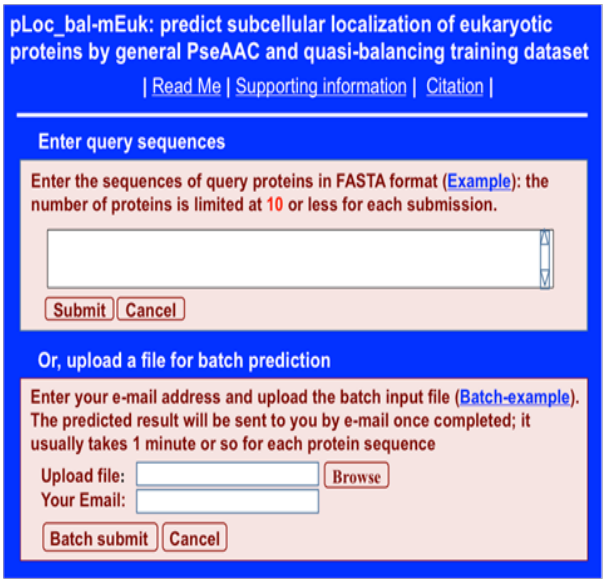

Figure 2: A semi screenshot for the top page of pLoc_bal-mEuk.

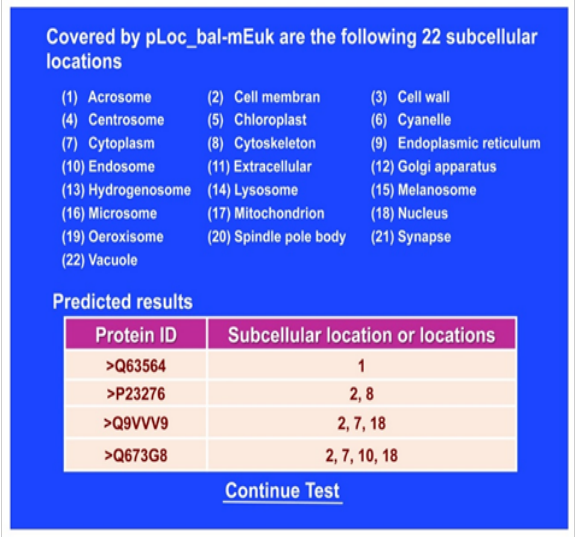

Figure 3: A semi screenshot for the webpage obtained by following Step 3 of Section 3.5.4.

Table 2: Performanceof pLoc_bal-mEukfor each of the 22 subcellular locations.

i	Location[a]	Sn(*i*)[b]	Sp(*i*)[b]	Acc(*i*)[b]	MCC(*i*)[b]
1	Acrosome	1	0.9997	0.9997	0.9353
2	Cell membrane	0.9986	0.9907	0.9914	0.9505
3	Cell wall	0.9796	0.999	0.9988	0.9158
4	Centrosome	1	0.9961	0.9961	0.8712
5	Chloroplast	0.9948	0.9988	0.9986	0.9851
6	Cyanelle	1	1	1	1
7	Cytoplasm	0.8477	0.9559	0.9254	0.8137
8	Cytoskeleton	1	0.9959	0.996	0.9024
9	Endoplasmic reticulum	0.9978	0.997	0.997	0.9741
10	Endosome	1	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	1	1	1
14	Lysosome	1	0.9999	0.9999	0.9913
15	Melanosome	1	1	1	1
16	Microsome	1	0.9995	0.9995	0.8742
17	Mitochondrion	1	0.994	0.9945	0.9636
18	Nucleus	0.8858	0.955	0.9343	0.8429
19	Peroxisome	1	0.9988	0.9988	0.9609
20	Spindle pole body	1	0.9991	0.9991	0.9518
21	Synapse	1	0.9994	0.9994	0.9504
22	Vacuole	1	0.9984	0.9985	0.9657

aSee Table 1 and the relevant context for further explanation.
bSee Eq.7 for the metrics definition.

6. 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 [63] 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 IA tools introduced in this review paper have been all established by following the 5-steps rule [35], 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[35,44,45,170,244], 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 [71,202,203,245], 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 [246-250], which represents a very hot trend

currently in drug development [251].

It is instructive to point out that, in comparison with their counterparts, the benchmark datasets in Section 2.2 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 [252-254], graphical rules in molecular biology [255-258], and low-frequency internal motion in biomacromolecules (such as protein and DNA) [259]. Particularly, what happened is that this kind of insightful implication has also been demonstrated in [260] and many follow-up publications [261-284].

References

1. Chen W, Feng PM, Lin H. iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. Nucleic Acids Res. 2013; 41: 68.

2. Feng PM, Chen W, Lin H, Chou KC.iHSP-PseRAAAC: Identifying the heat shock protein families using pseudo reduced amino acid alphabet composition. Anal Biochem. 2013; 442: 118-25.

3. Lin H, Deng EZ, Ding H, Chen W, Chou KC. iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition. Nucleic Acids Res. 2014; 42: 12961-72.

4. ChenW, Feng PM, Deng EZ, Lin H, Chou KC. iTIS-PseTNC: a sequence-based predictor for identifying translation initiation site in human genes using pseudo trinucleotide composition. Anal Biochem. 2014; 462: 76-83.

5. Ding H, Deng EZ, Yuan LF, Liu L, Lin H, Chen W et al. iCTX-Type: A sequence-based predictor for identifying the types of conotoxins in targeting ion channels. BioMed Research International (BMRI). 2014; 2014: 286419.

6. Liu B, Fang L, Wang S, Wang X, Li H, Chou KC. Identification of microRNA precursor with the degenerate K-tuple or Kmer strategy.Journal of Theoretical Biology. 2015; 385: 153-9.

7. Liu Z, Xiao X, QiuWR, Chou KC. iDNA-Methyl: Identifying DNA methylation sites via pseudo trinucleotide composition. Anal. Biochem. 2015; 474: 69-77.

8. Xiao X, Min JL, Lin WZ, Liu Z, Cheng X, Chou KC. iDrug-Target: predicting the interactions between drug compounds and target proteins in cellular networking via the benchmark dataset optimization approach. J Biomol Struct Dyn. 2015; 33:2221-33.

9. Jia J, Liu Z, Xiao X, Liu B, Chou KC. iSuc-PseOpt: Identifying lysine succinylation sites in proteins by incorporating sequence-coupling effects into pseudo components and optimizing imbalanced training dataset. Anal Biochem. 2016; 497: 48-56.

10. Jia J, Zhang L, Liu Z, Xiao X, Chou KC. pSumo-CD: Predicting sumoylation sites in proteins with covariance discriminant algorithm by incorporating sequence-coupled effects into general PseAAC. Bioinformatics. 2016; 32: 3133-41.

11. Liu B, Fang L, Long R, Lan X, Chou KC. iEnhancer-2L: a two-layer predictor for identifying enhancers and their strength by pseudo k-tuple nucleotide composition. Bioinformatics. 2016; 32: 362-9.

12. Chen W, Feng P, Yang H, Ding H, Lin H, Chou KC. iRNA-AI: identifying the adenosine to inosine editing sites in RNA sequences. Oncotarget. 2017; 8: 4208-17.

13. Chen W, Ding H, Zhou X, Lin H, Chou KC. iRNA(m6A)-PseDNC: Identifying N6-methyladenosine sites using pseudo dinucleotide composition. Anal Biochem. 2018; 561-2, 59-65.

14. Chen W, Feng P, Yang H, Ding H, Lin H, Chou KC.iRNA-3typeA: identifying 3-types of modification at RNA's adenosine sites. Molecular Therapy: Nucleic Acid. 2018; 11: 468-74.

15. Qiu WR, Sun BQ, Xiao X, Xu ZC, Jia JH, Chou KC. iKcr-PseEns: Identify lysine crotonylation sites in histone proteins with pseudo components and ensemble classifier. Genomics. 2018; 110: 239-46.

16. Feng P, Yang H, Ding H, Lin H, Chen W, Chou KC. iDNA6mA-PseKNC: Identifying DNA N(6)-methyladenosine sites by incorporating nucleotide physicochemical properties into PseKNC. Genomics. 2019; 111: 96-102.

17. Hussain W, Khan, SD, Rasool N, Khan SA, Chou KC. SPalmitoylC-PseAAC: A sequence-based model developed via Chou's 5-steps rule and general PseAAC for identifying S-palmitoylation sites in proteins. Anal Biochem. 2019; 568: 14-23.

18. Hussain W, Khan YD, Rasool N, Khan SA, Chou KC. SPrenylC-PseAAC: A sequence-based model developed

via Chou's 5-steps rule and general PseAAC for identifying S-prenylation sites in proteins. J Theor Biol. 2019; 468: 1-11.

19. Jia J, Li X, Qiu W, Xiao X, Chou KC. iPPI-PseAAC(CGR): Identify protein-protein interactions by incorporating chaos game representation into PseAAC. Journal of TheoreticalBiology. 2019; 460: 195-203.

20. Khan YD, Jamil M, Hussain W, Rasool N, Khan SA, Chou KC.pSSbond-PseAAC: Prediction of disulfide bonding sites by integration of PseAAC and statistical moments.J Theor Biol. 2019; 463: 47-55.

21. Lu Y, Wang S, Wang J, Zhou G, Zhang Q, Zhou X et al.An Epidemic Avian Influenza Prediction Model Based on Google Trends.Letters in Organic Chemistry. 2019; 16: 303-10.

22. Khan YD, Batool A, Rasool N, Khan A.Prediction of nitrosocysteine sites using position and composition variant features.Letters in Organic Chemistry. 2019; 16: 283-93.

23. Cheng X, Xiao X, Chou KC.pLoc_bal-mPlant: predict subcellular localization of plant proteins by general PseAAC and balancing training dataset. Curr Pharm Des. 2018; 24: 4013-22.

24. Li JX, Wang SQ, Du QS, Wei H, Li XM, Meng JZ et al.Simulated protein thermal detection (SPTD) for enzyme thermostability study and an application example for pullulanase from Bacillus deramificans.Curr Pharm Des. 2018; 24: 4023-33.

25. Ghauri AW, Khan YD, Rasool N, Khan SA, Chou KC.pNitro-Tyr-PseAAC: Predict nitrotyrosine sites in proteins by incorporating five features into Chou's general PseAAC.Curr Pharm Des. 2018; 24: 4034-43.

26. Chou KC.Advance in predicting subcellular localization of multi-label proteins and its implication for developing multi-target drugs.Current Medicinal Chemistry. 2019.

27. Xiao X, Cheng X, Chen G, Mao Q, Chou KC.pLoc_bal-mGpos: predict subcellular localization of Gram-positive bacterial proteins by quasi-balancing training dataset and PseAAC.Genomics. 2019; 111: 886-92.

28. Zhang M, Li F, Marquez-Lago TT, Leier A, Fan C, Kwoh CK et al.MULTiPly: a novel multi-layer predictor for discovering general and specific types of promoters.Bioinformatics. 2019; 35(17):2957-65.

29. Chen Z, Zhao P, Li F, Marquez-Lago TT, Leier A, Revote J et al.iLearn: an integrated platform and meta-learner for eature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data.Brief in Bioinform; 2019.

30. Zhang Y, Xie R, Wang J, Leier A, Marquez-Lago TT, Akutsu T et al.Computational analysis and prediction of lysine malonylation sites by exploiting informative features in an integrative machine-learning framework.Brief in Bioinform. 2018.

31. Song J, Wang Y, Li F, Akutsu T, Rawlings ND, Webb GI, et al.iProt-Sub: a comprehensive package for accurately mapping and predicting protease-specific substrates and cleavage sites. Brief in Bioinform. 2018; 20: 638-58.

32. Song J, Li F, Takemoto K, Haffari G, Akutsu T, Webb GI et al.PREvaIL, an integrative approach for inferring catalytic residues using sequence, structural and network features in a machine learning framework.Journal of TheoreticalBiology. 2018; 443: 125-37.

33. Li F, Wang Y, Li C, Marquez-Lago TT, Leier A, Rawlings ND et al.Twenty years of bioinformatics research for protease-specific substrate and cleavage site prediction: a comprehensive revisit and benchmarking of existing methods.Brief in Bioinform. 2018.

34. Li F, Li C, Marquez-Lago TT, Leier A, Akutsu T, Purcell AW et al.Quokka: a comprehensive tool for rapid and accurate prediction of kinase family-specific phosphorylation sites in the human proteome.Bioinformatics. 2018; 34: 4223-31.

35. Chou KC.Some remarks on protein attribute prediction and pseudo amino acid composition (50th Anniversary Year Review, 5-steps rule).J Theor Biol. 2011; 273: 236-47.

36. Zhai X, Chen M, Lu W.Accelerated search for perovskite materials with higher Curie temperature based on the machine learning methods.Computational Materials Science. 2018; 151: 41-8.

37. Chou KC, Shen HB.Recent progresses in protein subcellular location prediction.Anal Biochem. 2007; 370: 1-16.

38. Shen HB, Chou KC.A top-down approach to enhance the power of predicting human protein subcellular localization: Hum-mPLoc 2.0.Anal Biochem. 2009; 394: 269-74.

39. Shen HB, Chou KC..Gpos-mPLoc: A top-down approach to improve the quality of predicting subcellular localization of Gram-positive bacterial proteins.Protein & Peptide Letters. 2009; 16: 1478-84.

40. Chou KC, Shen HB.A new method for predicting the subcellular localization of eukaryotic proteins with both single

and multiple sites: Euk-mPLoc 2.0 PLoS ONE. 2010; 5(4): e9931.

41. Chou KC, Shen HB.Plant-mPLoc: A Top-Down Strategy to Augment the Power for Predicting Plant Protein Subcellular Localization.PLoS ONE. 2010; 5: e11335.

42. Shen HB.Gneg-mPLoc: A top-down strategy to enhance the quality of predicting subcellular localization of Gram-negative bacterial proteins.Journal of TheoreticalBiology. 2010; 264: 326-33.

43. Shen HB, Chou KC.Virus-mPLoc: A Fusion Classifier for Viral Protein Subcellular Location Prediction by Incorporating Multiple Sites.J Biomol Struct Dyn (JBSD). 2010; 28: 175-86.

44. Chou KC.Prediction of protein cellular attributes using pseudo amino acid composition.PROTEINS: Structure, Function, and Genetics. 2001; 43: 246-55.

45. Chou KC.Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes.Bioinformatics. 2005; 21: 10-9.

46. Chou KC, Wu ZC, Xiao X.iLoc-Euk: A Multi-Label Classifier for Predicting the Subcellular Localization of Singleplex and Multiplex Eukaryotic Proteins.PLoS One. 2011; 6: e18258.

47. Wu ZC, Xiao X, Chou KC.iLoc-Plant: a multi-label classifier for predicting the subcellular localization of plant proteins with both single and multiple sites.Molecular BioSystems. 2011; 7: 3287-97.

48. Xiao X, Wu ZC, Chou KC.iLoc-Virus: A multi-label learning classifier for identifying the subcellular localization of virus proteins with both single and multiple sites.J Theor Biol. 2011; 284: 42-51.

49. Chou KC, Wu ZC, Xiao X.iLoc-Hum: Using accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites.Molecular Biosystems. 2012; 8: 629-41.

50. Xiao X, Wu ZC, Chou KC.A multi-label classifier for predicting the subcellular localization of gram-negative bacterial proteins with both single and multiple sites.PLoS ONE. 2011; 6: e20592.

51. Wu ZC, Xiao X, Chou KC.iLoc-Gpos: A Multi-Layer Classifier for Predicting the Subcellular Localization of Singleplex and Multiplex Gram-Positive Bacterial Proteins.Protein & Peptide Letters. 2012; 19: 4-14.

52. Lin WZ, Fang JA, Xiao X, Chou KC.iLoc-Animal: A multi-label learning classifier for predicting subcellular localization of animal proteins Molecular BioSystems. 2013; 9: 634-44.

53. Cheng X, Xiao X, Chou KC.pLoc-mPlant: predict subcellular localization of multi-location plant proteins via incorporating the optimal GO information into general PseAAC.Molecular BioSystems. 2017; 13: 1722-7.

54. Cheng X, Xiao X, Chou KC.pLoc-mVirus: predict subcellular localization of multi-location virus proteins via incorporating the optimal GO information into general PseAAC.Gene (Erratum: ibid., 2018, Vol.644, 156-156). 2017; 628: 315-21.

55. Cheng X, Zhao SG, Lin WZ, Xiao X, Chou KC.pLoc-mAnimal: predict subcellular localization of animal proteins with both single and multiple sites.Bioinformatics. 2017; 33: 3524-31.

56. Xiao X, Cheng X, Su S, Nao Q, Chou KC.pLoc-mGpos: Incorporate key gene ontology information into general PseAAC for predicting subcellular localization of Gram-positive bacterial proteins.Natural Science. 2017; 9: 331-49.

57. Cheng X, Xiao X, Chou KC.pLoc-mEuk: Predict subcellular localization of multi-label eukaryotic proteins by extracting the key GO information into general PseAAC.Genomics. 2018; 110: 50-8.

58. Cheng X, Xiao X, Chou KC.pLoc-mGneg: Predict subcellular localization of Gram-negative bacterial proteins by deep gene ontology learning via general PseAAC.Genomics. 2018; 110: 231-9.

59. Cheng X, Xiao X, Chou KC.pLoc-mHum: predict subcellular localization of multi-location human proteins via general PseAAC to winnow out the crucial GO information. Bioinformatics. 2018; 34: 1448-56.

60. Cheng X, Xiao X, Chou KC.pLoc_bal-mGneg: predict subcellular localization of Gram-negative bacterial proteins by quasi-balancing training dataset and general PseAAC.Journal of TheoreticalBiology. 2018; 458: 92-102.

61. Chou KC, Cheng X, Xiao X.pLoc_bal-mHum: predict subcellular localization of human proteins by PseAAC and quasi-balancing training dataset Genomics. 2018; S0888-7543(18): 30276-3.

62. Cheng X, Lin WZ, Xiao X, Chou KC.pLoc_bal-mAnimal: predict subcellular localization of animal proteins by balancing training dataset and PseAAC.Bioinformatics. 2019; 35: 398-406.

63. Chou KC.Some remarks on predicting multi-label attributes

in molecular biosystems.Molecular Biosystems. 2013; 9: 1092-100.

64. Chou KC, Elrod DW.Bioinformatical analysis of G-protein-coupledreceptors.Journal of Proteome Research. 2002; 1: 429-33.

65. Chen W, Lin H, Feng PM, Ding C, Zuo YC, Chou KC.iNuc-PhysChem: A Sequence-Based Predictor for Identifying Nucleosomes via Physicochemical Properties.PLoS ONE. 2012; 7: e47843.

66. Xu Y, Ding J, Wu LY, Chou KC.iSNO-PseAAC: Predict cysteine S-nitrosylation sites in proteins by incorporating position specific amino acid propensity into pseudo amino acid composition PLoS ONE. 2013; 8: e55844.

67. Cai YD, Chou KC.Predicting subcellular localization of proteins in a hybridization space.Bioinformatics. 2004; 20: 1151-6.

68. Chou KC, Cai YD.Prediction of protease types in a hybridization space.Biochem Biophys ResComm (BBRC). 2006; 339: 1015-20.

69. Lin WZ, Fang JA, Xiao X, Chou KC.iDNA-Prot: Identification of DNA Binding Proteins Using Random Forest with Grey Model.PLoS ONE. 2011; 6: e24756.

70. Kandaswamy KK, Martinetz T, Moller S, Suganthan PN, Sridharan S, Pugalenthi G.AFP-Pred: A random forest approach for predicting antifreeze proteins from sequence-derived properties.J. Theor. Biol. 2011; 270: 56-62.

71. Chou KC.Impacts of bioinformatics to medicinal chemistry. Medicinal Chemistry. 2015; 11: 218-34.

72. Fang Y, Guo Y, Feng Y, Li M.Predicting DNA-binding proteins: approached from Chou's pseudo amino acid composition and other specific sequence features.Amino Acids. 2008; 34: 103-9.

73. Zhang SW, Chen W, Yang F, Pan Q.Using Chou's pseudo amino acid composition to predict protein quaternary structure: a sequence-segmented PseAAC approach.Amino Acids. 2008; 35: 591-8.

74. Chen C, Chen L, Zou X, Cai P.Prediction of protein secondary structure content by using the concept of Chou's pseudo amino acid composition and support vector machine.Protein & Peptide Letters. 2009; 16: 27-31.

75. Lin H, Wang H, Ding H, Chen YL, Li QZ.Prediction of Subcellular Localization of Apoptosis Protein Using Chou's Pseudo Amino Acid Composition.Acta Biotheoretica. 2009; 57: 321-30.

76. Esmaeili M, Mohabatkar H, Mohsenzadeh S.Using the concept of Chou's pseudo amino acid composition for risk type prediction of human papillomaviruses.J Theor Biol. 2010; 263: 203-9.

77. Mohabatkar H.Prediction of cyclin proteins using Chou's pseudo amino acid composition.Protein & Peptide Letters. 2010; 17: 1207-14.

78. Qiu JD, Huang JH, Shi SP, Liang RP.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. 2010; 17: 715-22.

80. Sahu SS, Panda G.A novel feature representation method based on Chou's pseudo amino acid composition for protein structural class prediction.Computational Biology and Chemistry. 2010; 34: 320-7.

81. Yu L, Guo Y, Li Y, Li G, Li M, Luo J et al.SecretP: Identifying bacterial secreted proteins by fusing new features into Chou's pseudo amino acid composition.J Theor Biol. 2010; 267: 1-6.

82. Guo J, Rao N, Liu G, Yang Y, Wang G.Predicting protein folding rates using the concept of Chou's pseudo amino acid composition.Journal ofComputational Chemistry. 2011; 32: 1612-7.

83. Lin J, Wang Y.Using a novel AdaBoost algorithm and Chou's pseudo amino acid composition for predicting protein subcellular localization.Protein & Peptide Letters. 2011; 18: 1219-25.

84. Mohammad BM, Behjati M, Mohabatkar H.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. 2011; 12: 191-7.

85. Zou D, He Z, He J, Xia Y.Supersecondary structure prediction using Chou's pseudo amino acid composition. JComput Chem. 2011; 32: 271-8.

86. Du P, Wang X, Xu C, Gao Y.PseAAC-Builder: A cross-platform stand-alone program for generating various special Chou's pseudo amino acid compositions.Anal Biochem. 2012; 425: 117-9.

87. Hayat M, Khan A.Discriminating Outer Membrane Proteins with Fuzzy K-Nearest Neighbor Algorithms Based on the General Form of Chou's PseAAC.Protein & Peptide Letters. 2012; 19: 411-21.

88. Li LQ, Zhang Y, Zou LY, Zhou Y, Zheng XQ.Prediction of Protein Subcellular Multi-Localization Based on the General form of Chou's Pseudo Amino Acid Composition.Protein & Peptide Letters. 2012; 19: 375-87.

89. Liao B, Xiang Q, Li D.Incorporating Secondary Features into the General form of Chou's PseAAC for Predicting Protein Structural Class.Protein & Peptide Letters. 2012; 19: 1133-8.

90. Mei S.Multi-kernel transfer learning based on Chou's PseAAC formulation for protein submitochondria localization.J Theor Biol. 2012; 293: 121-30.

91. Mei S.Predicting plant protein subcellular multi-localization by Chou's PseAAC formulation based multi-label homolog knowledge transfer learning.J Theor Biol. 2012; 310: 80-7.

92. Qin YF, Wang CH, Yu XQ, Zhu J, Liu TG, Zheng XQ.Predicting Protein Structural Class by Incorporating Patterns of Over-Represented k-mers into the General form of Chou's PseAAC. Protein & Peptide Letters, 2012; 19: 388-97.

93. Sun XY, Shi SP, Qiu JD, Suo SB, Huang SY, Liang RP.Identifying protein quaternary structural attributes by incorporating physicochemical properties into the general form of Chou's PseAAC via discrete wavelet transform.Molecular BioSystems. 2012; 8: 3178-84.

94. Zhao XW, Li XT, Ma ZQ, Yin MH.Identify DNA-Binding Proteins with Optimal Chou's Amino Acid Composition.Protein & Peptide Letters. 2012; 19: 398-405.

95. Zhao XW, Ma ZQ, Yin MH.Predicting protein-protein interactions by combing various sequence- derived features into the general form of Chou's Pseudo amino acid composition. Protein & Peptide Letters. 2012; 19: 492-500.

96. Cao DS, Xu QS, Liang YZ.propy: a tool to generate various modes of Chou's PseAAC.Bioinformatics. 2013; 29: 960-2.

97. Chang TH, Wu LC, Lee TY, Chen SP, Huang HD, Horng JT.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. 2013; 27: 91-103.

98. Fan GL, Li QZ, Zuo YC.Predicting acidic and alkaline enzymes by incorporating the average chemical shift and gene ontology informations into the general form of Chou's PseAAC. Pocess Biochemistry. 2013; 48: 1048-53.

99. Fan GL, Li QZ.Discriminating bioluminescent proteins by incorporating average chemical shift and evolutionary information into the general form of Chou's pseudo amino acid composition.J Theor Biol.2013; 334: 45-51.

100. Khosravian M, Faramarzi FK, Beigi MM, Behbahani M, Mohabatkar H.Predicting Antibacterial Peptides by the Concept of Chou's Pseudo amino Acid Composition and Machine Learning Methods.Protein & Peptide Letters. 2013; 20: 180-6.

101. Mohabatkar H, Beigi MM, Abdolahi K, Mohsenzadeh S.Prediction of Allergenic Proteins by Means of the Concept of Chou's Pseudo Amino Acid Composition and a Machine Learning Approach.Medicinal Chemistry. 2013; 9: 133-7.

102. Pacharawongsakda E, Theeramunkong T. 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. 2013; 12: 311-20.

103. Sarangi AN, Lohani M. Aggarwal R. Prediction of Essential Proteins in Prokaryotes by Incorporating Various Physico-chemical Features into the General form of Chou's Pseudo Amino Acid Composition. Protein Pept Lett. 2013; 20: 781-95.

104. Wang X, Li GZ, Lu WC. 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. 2013; 20: 309-17.

105. Xiaohui N, Nana L, Jingbo X, Dingyan C, Yuehua P, Yang X, et al. Using the concept of Chou's pseudo amino acid composition to predict protein solubility: An approach with entropies in information theory. J. Theor. Biol. 2013; 332: 211-7.

106. Xie HL, Fu L, Nie XD. Using ensemble SVM to identify human GPCRs N-linked glycosylation sites based on the general form of Chou's PseAAC. Protein Eng Des Sel. 2013; 26: 735-42.

107. Du P, Gu S, Jiao Y. 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. 2014; 15: 3495-506.

108. Hajisharifi Z, Piryaiee M, Mohammad Beigi M, Behbahani

M, Mohabatkar H. Predicting anticancer peptides with Chou's pseudo amino acid composition and investigating their mutagenicity via Ames test. J. Theor. Biol. 2014; 341: 34-40.

109. Han GS, Yu ZG, Anh V. 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. J. Theor. Biol. 2014; 344: 31-9.

110. Hayat M, Iqbal N. 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. 2014; 116: 184-92.

111. Jia C, Lin X, Wang Z. Prediction of Protein S-Nitrosylation Sites Based on Adapted Normal Distribution Bi-Profile Bayes and Chou's Pseudo Amino Acid Composition. Int J Mol Sci. 2014; 15: 10410-23.

112. Li L, Yu S, Xiao W, Li Y, Li M, Huang L, et al. Prediction of bacterial protein subcellular localization by incorporating various features into Chou's PseAAC and a backward feature selection approach. Biochimie. 2014; 104: 100-7.

113. Mondal S, Pai PP. Chou's pseudo amino acid composition improves sequence-based antifreeze protein prediction. J. Theor. Biol. 2014; 356: 30-5.

114. Zhang J, Zhao X, Sun P, Ma Z. PSNO: Predicting Cysteine S-Nitrosylation Sites by Incorporating Various Sequence-Derived Features into the General Form of Chou's PseAAC. Int J Mol Sci. 2014; 15: 11204-19.

115. Ahmad S, Kabir M, Hayat M. 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. 2015; 122: 165-74.

116. Dehzangi A, Heffernan R, Sharma A, Lyons J, Paliwal K, Sattar A. Gram-positive and Gram-negative protein subcellular localization by incorporating evolutionary-based descriptors into Chou's general PseAAC. J. Theor. Biol. 2015; 364: 284-94.

117. Khan ZU, Hayat M, Khan MA. Discrimination of acidic and alkaline enzyme using Chou's pseudo amino acid composition in conjunction with probabilistic neural network model. J. Theor. Biol. 2015; 365: 197-203.

118. Liu B, Chen J, Wang X. Protein remote homology detection by combining Chou's distance-pair pseudo amino acid composition and principal component analysis. Molecular genetics and genomics : MGG. 2015; 290: 1919-31.

119. Liu B, Xu J, Fan S, Xu R, Jiyun Zhou J, Wang X. PseDNA-Pro: DNA-binding protein identification by combining Chou's PseAAC and physicochemical distance transformation. Molecular Informatics. 2015; 34: 8-17.

120. Mandal M, Mukhopadhyay A, Maulik U. 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. 2015; 53: 331-44.

121. Sanchez V, Peinado AM, Perez-Cordoba JL, Gomez AM. A new signal characterization and signal-based Chou's PseAAC representation of protein sequences. Journal of bioinformatics and computational biology. 2015; 13: 1550024.

122. Sharma R, Dehzangi A, Lyons J, Paliwal K, Tsunoda T, Sharma A. Predict Gram-Positive and Gram-Negative Subcellular Localization via Incorporating Evolutionary Information and Physicochemical Features Into Chou's General PseAAC. IEEE Trans Nanobioscience. 2015; 14: 915-26.

123. Zhang M, Zhao B, Liu X. Predicting industrial polymer melt index via incorporating chaotic characters into Chou's general PseAAC. Chemometrics and Intelligent Laboratory Systems (CHEMOLAB). 2015; 146: 232-40.

124. Zhang SL. Accurate prediction of protein structural classes by incorporating PSSS and PSSM into Chou's general PseAAC. Chemometrics and Intelligent Laboratory Systems (CHEMOLAB). 2015; 142: 28-35.

125. Behbahani M, Mohabatkar H, Nosrati M. Analysis and comparison of lignin peroxidases between fungi and bacteria using three different modes of Chou's general pseudo amino acid composition. J. Theor. Biol. 2016; 411: 1-5.

126. Jiao YS, Du PF. Prediction of Golgi-resident protein types using general form of Chou's pseudo amino acid compositions: Approaches with minimal redundancy maximal relevance feature selection. J. Theor. Biol. 2016; 402: 38-44.

127. Ju Z, Cao JZ, Gu H. Predicting lysine phosphoglycerylation with fuzzy SVM by incorporating k-spaced amino acid pairs into Chou's general PseAAC. J. Theor. Biol. 2016; 397: 145-50.

128. Kabir M, Hayat M. iRSpot-GAEnsC: identifing recombination spots via ensemble classifier and extending the concept of Chou's PseAAC to formulate DNA samples. Molecular Genetics and Genomics. 2015; 291: 285-96.

129. Tahir M, Hayat M. iNuc-STNC: a sequence-based predictor

for identification of nucleosome positioning in genomes by extending the concept of SAAC and Chou's PseAAC. MolBiosyst. 2016; 12: 2587-93.

130. Tiwari AK. 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. 2016; 134: 197-213.

131. Ju Z, He JJ. Prediction of lysine propionylation sites using biased SVM and incorporating four different sequence features into Chou's PseAAC. J Mol Graph Model. 2017; 76: 356-63.

132. Ju Z, He JJ. Prediction of lysine crotonylation sites by incorporating the composition of k-spaced amino acid pairs into Chou's general PseAAC. J Mol Graph Model. 2017; 77: 200-4.

133. Khan M, Hayat M, Khan SA, Iqbal N. Unb-DPC: Identify mycobacterial membrane protein types by incorporating un-biased dipeptide composition into Chou's general PseAAC. J. Theor. Biol. 2017; 415: 13-19.

134. Liang Y, Zhang S. Predict protein structural class by incorporating two different modes of evolutionary information into Chou's general pseudo amino acid composition. J Mol Graph Model. 2017; 78: 110-7.

135. Meher PK, Sahu TK, Saini V, Rao AR. Predicting antimicrobial peptides with improved accuracy by incorporating the compositional, physico-chemical and structural features into Chou's general PseAAC. Sci Rep. 2017; 7: 42362.

136. Qiu WR, Zheng QS, Sun,BQ, Xiao X. Multi-iPPseEvo: A Multi-label Classifier for Identifying Human Phosphorylated Proteins by Incorporating Evolutionary Information into Chou's General PseAAC via Grey System Theory. Mol Inform. 2016; 36: UNSP 1600085.

137. Tripathi P, Pandey PN. A novel alignment-free method to classify protein folding types by combining spectral graph clustering with Chou's pseudo amino acid composition. J. Theor. Biol. 2017; 424: 49-54.

138. Xu C, Ge L, Zhang Y, Dehmer M, Gutman I. Prediction of therapeutic peptides by incorporating q-Wiener index into Chou's general PseAAC. J Biomed Inform. 2017; doi:10.1016/j.jbi.2017.09.011.

139. Yu B, Li S, Qiu WY, Chen C, Chen RX, Wang L, et al. Accurate prediction of subcellular location of apoptosis proteins combining Chou's PseAAC and PsePSSM based on wavelet denoising. Oncotarget. 2017; 8: 107640-65.

140. Yu B, Lou L, Li S, Zhang Y, Qiu W, Wu X. et al. Prediction of protein structural class for low-similarity sequences using Chou's pseudo amino acid composition and wavelet denoising. J Mol Graph Model. 2017; 76: 260-73.

141. Ahmad J, Hayat M. MFSC: Multi-voting based Feature Selection for Classification of Golgi Proteins by Adopting the General form of Chou's PseAAC components. J. Theor. Biol. 2018; 463: 99-109.

142. Akbar S, Hayat M. iMethyl-STTNC: Identification of N(6)-methyladenosine sites by extending the Idea of SAAC into Chou's PseAAC to formulate RNA sequences. J. Theor. Biol. 2018; 455: 205-11.

143. Arif M, Hayat M, Jan Z. 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. J. Theor. Biol. 2018; 442; 11-21.

144. Butt AH, Rasool N, Khan YD. Predicting membrane proteins and their types by extracting various sequence features into Chou's general PseAAC. Molecular biology reports. 2018; 45: 2295-306.

145. Contreras-Torres, E. Predicting structural classes of proteins by incorporating their global and local physicochemical and conformational properties into general Chou's PseAAC. J. Theor. Biol. 2018; 454: 139-45.

146. Fu X, Zhu W, Liso B, Cai L, Peng L, Yang J. Improved DNA-binding protein identification by incorporating evolutionary information into the Chou's PseAAC. IEEE Access. 2018; 20: 2018.2876656.

147. Javed F, Hayat M. Predicting subcellular localizations of multi-label proteins by incorporating the sequence features into Chou's PseAAC. Genomics. 2018; 7543: 30519-6.

148. Krishnan MS. Using Chou's general PseAAC to analyze the evolutionary relationship of receptor associated proteins (RAP) with various folding patterns of protein domains. J. Theor. Biol. 2018; 445: 62-74.

149. Liang Y, Zhang S. Identify Gram-negative bacterial secreted protein types by incorporating different modes of PSSM into Chou's general PseAAC via Kullback-Leibler divergence. J. Theor. Biol. 2018; 454: 22-9.

150. Mousavizadegan M, Mohabatkar H. Computational

prediction of antifungal peptides via Chou's PseAAC and SVM. Journal of bioinformatics and computational biology. 2018; 1850016.

151. Qiu W, Li S, Cui X, Yu Z, Wang M, Du J, et al. Predicting protein submitochondrial locations by incorporating the pseudo-position specific scoring matrix into the general Chou's pseudo-amino acid composition. J. Theor. Biol. 2018; 450: 86-103.

152. Rahman SM, Shatabda S, Saha S, Kaykobad M, Sohel Rahman. M. DPP-PseAAC: A DNA-binding Protein Prediction model using Chou's general PseAAC. J. Theor. Biol. 2018; 452: 22-34.

153. Sankari ES, Manimegalai DD. Predicting membrane protein types by incorporating a novel feature set into Chou's general PseAAC. J. Theor. Biol. 2018; 455: 319-28.

154. Srivastava A, Kumar R, Kumar M. BlaPred: predicting and classifying beta-lactamase using a 3-tier prediction system via Chou's general PseAAC. J. Theor. Biol. 2018; 457: 29-36.

155. Zhang S, Duan X. Prediction of protein subcellular localization with oversampling approach and Chou's general PseAAC. J. Theor. Biol. 2018; 437: 239-50.

156. Zhang S, Liang Y. Predicting apoptosis protein subcellular localization by integrating auto-cross correlation and PSSM into Chou's PseAAC. J. Theor. Biol. 2018; 457: 163-9.

157. Adilina S, Farid DM, Shatabda S. Effective DNA binding protein prediction by using key features via Chou's general PseAAC. J. Theor. Biol. 2019; 460: 64-78.

158. Ahmad J, Hayat M. MFSC: Multi-voting based feature selection for classification of Golgi proteins by adopting the general form of Chou's PseAAC components. J. Theor. Biol. 2019; 463: 99-109.

159. Awais M, Hussain W, Khan YD, Rasool N, Khan SA. 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 Trans ComputBiolBioinform. 2019; doi:10.1109/TCBB.2019.2919025.

160. Butt AH, Rasool N, Khan YD. Prediction of antioxidant proteins by incorporating statistical moments based features into Chou's PseAAC. J. Theor. Biol. 2019; 473: 1-8.

161. Chen G, Cao M, Yu J, Guo X, Shi S. Prediction and functional analysis of prokaryote lysine acetylation site by incorporating six types of features into Chou's general PseAAC. J. Theor. Biol. 2019; 461: 92-101.

162. Ehsan A, Mahmood MK, Khan YD, Barukab OM, Khan SA. 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. 2019; 20: 124-33.

163. Kabir M, Ahmad S, Iqbal M, Hayat M. 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. 2019; 7543: 30694-3.

164. Ning Q, Ma Z, Zhao X. dForml(KNN)-PseAAC: Detecting formylation sites from protein sequences using K-nearest neighbor algorithm via Chou's 5-step rule and pseudo components. J. Theor. Biol. 2019; 470: 43-9.

165. Shen Y, Tang J, Guo F. Identification of protein subcellular localization via integrating evolutionary and physicochemical information into Chou's general PseAAC. J. Theor. Biol. 2019; 462: 230-9.

166. Tahir M, Hayat M, Khan S.A. 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: MGG. 2019; 294: 199-210.

167. Tian B, Wu X, Chen C, Qiu W, Ma Q, Yu B. Predicting protein-protein interactions by fusing various Chou's pseudo components and using wavelet denoising approach. J. Theor. Biol. 2019; 462: 329-46.

168. Wang L, Zhang R, Mu Y. Fu-SulfPred: Identification of Protein S-sulfenylation Sites by Fusing Forests via Chou's General PseAAC. J. Theor. Biol. 2019; 461: 51-8.

169. Chou KC. An unprecedented revolution in medicinal chemistry driven by the progress of biological science. Current Topics in Medicinal Chemistry. 2017; 17: 2337-58.

170. Shen HB, Chouc KC, PseAAC: a flexible web-server for generating various kinds of protein pseudo amino acid composition. Anal. Biochem. 2008; 373: 386-8.

171. Chou KC. Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology. Current Proteomics. 2009; 6: 262-74.

172. Wang J, Yang B, Revote J, Leier A, Marquez-Lago TT. Webb G. POSSUM: a bioinformatics toolkit for generating numerical sequence feature descriptors based on PSSM profiles. Bioinformatics. 2017; 33: 2756-8.

173. Chen W, Lei TY, Jin DC, Lin H, Chou KC. PseKNC: a flexible web-server for generating pseudo K-tuple nucleotide composition. Anal. Biochem. 2014; 456: 53-60.

174. Liu B, Liu F, Fang L, Wang X. 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. 2015; 31: 1307-9.

175. Liu B, Liu F, Fang L, Wang X. repRNA: a web server for generating various feature vectors of RNA sequences. Molecular Genetics and Genomics. 2016; 291: 473-81.

176. Liu B, Liu F, Wang X, Chen J, Fang L. Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. Nucleic Acids Res. 2015; 43: W65-W71.

177. Chen W, Lin H, Chou KC. Pseudo nucleotide composition or PseKNC: an effective formulation for analyzing genomic sequences. MolBioSyst. 2015; 11: 2620-34.

178. Chen W, Feng PM, Lin H. iSS-PseDNC: identifying splicing sites using pseudo dinucleotide composition. Biomed Research International (BMRI). 2014; 2014: 623149.

179. Chen W, Tang H, Ye J, Lin H. iRNA-PseU: Identifying RNA pseudouridine sites Molecular Therapy - Nucleic Acids. 2016; 5: e332.

180. Liu B, Long R, Chou KC. iDHS-EL: Identifying DNase I hypersensi-tivesites by fusing three different modes of pseudo nucleotide composition into an ensemble learning framework. Bioinformatics. 2016; 32: 2411-18.

181. Feng P, Ding H, Yang H, Chen W, Lin H. iRNA-PseColl: Identifying the occurrence sites of different RNA modifications by incorporating collective effects of nucleotides into PseKNC. Molecular Therapy - Nucleic Acids. 2017; 7: 155-63.

182. Liu B, Wang S, Long R, Chuo KC. iRSpot-EL: identify recombination spots with an ensemble learning approach. Bioinformatics. 2017; 33: 35-41.

183. Liu B, Yang F. 2L-piRNA: A two-layer ensemble classifier for identifying piwi-interacting RNAs and their function. Molecular Therapy - Nucleic Acids. 2017; 7: 267-77.

184. Al Maru MA, Shatabda S. iRSpot-SF: Prediction of recombination hotspots by incorporating sequence based features into Chou's Pseudo components. Genomics. 2018; 111: 966-72.

185. Sabooh MF, Iqbal N, Khan M, Khan M, Maqbool HF. Identifying 5-methylcytosine sites in RNA sequence using composite encoding feature into Chou's PseKNC. J. Theor. Biol. 2018; 452: 1-9.

186. Zhang L, Kong L. iRSpot-ADPM: Identify recombination spots by incorporating the associated dinucleotide product model into Chou's pseudo components. J. Theor. Biol. 2018; 441: 1-8.

187. Zhang L, Kong L. iRSpot-PDI: Identification of recombination spots by incorporating dinucleotide property diversity information into Chou's pseudo components. Genomics. 2019; 111: 457-64.

188. Liu B, Wu H. 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 Science. 2017; 9: 67-91.

189. Shen HB. Using optimized evidence-theoretic K-nearest neighbor classifier and pseudo amino acid composition to predict membrane protein types. Biochemical & Biophysical Research Communications (BBRC). 2005; 334: 288-92.

190. Chou KC, Shen HB. Euk-mPLoc: a fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites. Journal of Proteome Research. 2007; 6: 1728-34.

191. Shen HB, Chou KC. QuatIdent: A web server for identifying protein quaternary structural attribute by fusing functional domain and sequential evolution information. Journal of Proteome Research. 2009; 8: 1577-84.

192. Chou KC. Prediction of protein signal sequences and their cleavage sites. Proteins: Struct., Funct., Genet. 2001; 42: 136-9.

193. Chou KC. Using subsite coupling to predict signal peptides. Protein Eng. 2001; 14: 75-9.

194. Chou KC. Prediction of signal peptides using scaled window. Peptides. 2001; 22: 1973-9.

195. Qiu WR, Xiao X. iRSpot-TNCPseAAC: Identify recombination spots with trinucleotide composition and pseudo amino acid components. Int J MolSci (IJMS). 2014;15: 1746-66.

196. Xu Y, Wen X, Wen LS, Wu LY, Deng NY. iNitro-Tyr:

Prediction of nitrotyrosine sites in proteins with general pseudo amino acid composition. PLoS ONE. 2014; 9: e105018.

197. Chen W, Feng P, Ding H, Lin H. iRNA-Methyl: Identifying N6-methyladenosine sites using pseudo nucleotide composition. Anal. Biochem. 2015; 490: 26-33.

198. Jia J, Liu Z, Xiao X, Liu B. iPPI-Esml: an ensemble classifier for identifying the interactions of proteins by incorporating their physicochemical properties and wavelet transforms into PseAAC. J. Theor. Biol. 2015; 377: 47-56.

199. Chen W, Ding H, Feng P, Lin H. iACP: a sequence-based tool for identifying anticancer peptides. Oncotarget. 2016; 7: 16895-909.

200. Chen W, Feng P, Ding H, Lin H. Using deformation energy to analyze nucleosome positioning in genomes. Genomics. 2016; 107: 69-75.

201. Jia J, Liu Z, Xiao X, Liu B. pSuc-Lys: Predict lysine succinylation sites in proteins with PseAAC and ensemble random forest approach. J. Theor. Biol. 2016; 394: 223-30.

202. Chou, KC, Zhang CT. Review: Prediction of protein structural classes. Crit. Rev. Biochem. Mol. Biol. 1995; 30: 275-349.

203. Chou KC. Shen HB. Cell-PLoc: A package of Web servers for predicting subcellular localization of proteins in various organisms. Nature Protocols. 2008; 3: 153-162.

204. Chou KC, Shen HB. Cell-PLoc 2.0: An improved package of web-servers for predicting subcellular localization of proteins in various organisms. Natural Science, 2010; 2: 1090-1103.

205. Zhou GP, Assa-Munt N. Some insights into protein structural class prediction. Proteins: StructFunct Genet. 2001; 44: 57-59.

206. Zhou GP, Doctor K. Subcellular location prediction of apoptosis proteins. Proteins: StructFunct Genet. 2003: 50: 44-8.

207. Zia-ur-Rehman, Khan A. 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. 2012; 19: 890-903.

208. Huang C, Yuan JQ. Predicting protein sub chloroplast locations with both single and multiple sites via three different modes of Chou's pseudo amino acid compositions. J Theor Biol. 2013; 335: 205-12.

209. Chou KC, Shen HB. Recent advances in developing web-servers for predicting protein attributes. Natural Science. 2009; 1: 63-92.

210. Xu Y, Shao XJ. Wu LY, Deng NY. iSNO-AAPair: incorporating amino acid pairwise coupling into PseAAC for predicting cysteine S-nitrosylation sites in proteins. PeerJ. 2013; 1: e171.

211. Liu B, Xu J, Lan X, Xu R, Zhou J Wang X et al. iDNA-Prot|dis: identifying DNA-binding proteins by incorporating amino acid distance-pairs and reduced alphabet profile into the general pseudo amino acid composition. PLoS ONE. 2014; 9: e106691.

212. Xu Y, Wen X, Shao X J, Deng NY. iHyd-PseAAC: Predicting hydroxyproline and hydroxylysine in proteins by incorporating dipeptide position-specific propensity into pseudo amino acid composition. Int J Mol Sci. 2014; 15: 7594-610.

213. Qiu WR, Xiao X, Lin WZ. iMethyl-PseAAC: Identification of Protein Methylation Sites via a Pseudo Amino Acid Composition Approach. Biomed Res Int (BMRI). 2014; 2014: 947416.

214. Fan YN, Xiao X, Min JL. iNR-Drug: Predicting the interaction of drugs with nuclear receptors in cellular networking. Intenational Journal of Molecular Sciences (IJMS). 2014; 15: 4915-37.

215. Guo SH, Deng EZ, Xu, LQ, Ding H, Lin H, et al. iNuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition. Bioinformatics. 2014; 30: 1522-9.

216. Qiu WR, Xiao X, Lin WZ. 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 (JBSD). 2015; 33: 1731-42.

217. Liu B, Fang L, Liu F, Wang X, Chen J. Identification of real micro RNA precursors with a pseudo structure status composition approach. PLoS ONE. 2015; 10: e0121501.

218. Liu Z, Xiao X, Yu DJ, Jia J, Qiu WR, Chou KC et al. pRNAm-PC: Predicting N-methyladenosine sites in RNA sequences via physical-chemical properties. Anal Biochem. 2016; 497; 60-7.

219. Xiao X, Ye HX, Liu Z, Jia JH, Chou KC5. iROS-gPseKNC: predicting replication origin sites in DNA by incorporating dinucleotide position-specific propensity into general pseudo nucleotide composition. Oncotarget. 2016; 7: 34180-9.

220. Qiu WR, Sun BQ, Xiao X, Xu ZC, Chou KC4. iPTM-mLys:

identifying multiple lysine PTM sites and their different types. Bioinformatics. 2016; 32: 3116-23.

221. Jia J, Liu Z, Xiao X, Liu B. iPPBS-Opt: A Sequence-Based Ensemble Classifier for Identifying Protein-Protein Binding Sites by Optimizing Imbalanced Training Datasets. Molecules. 2016; 21: E95.

222. Qiu WR, Xiao X, Xu ZC. iPhos-PseEn: identifying phosphorylation sites in proteins by fusing different pseudo components into an ensemble classifier. Oncotarget. 2016; 7: 51270-83.

223. Zhang CJ, Tang H, Li WC, Lin H, Chen W. iOri-Human: identify human origin of replication by incorporating dinucleotide physicochemical properties into pseudo nucleotide composition. Oncotarget. 2016; 7: 69783-93.

224. Liu B, Fang L, Liu F, Wang X. iMiRNA-PseDPC: microRNA precursor identification with a pseudo distance-pair composition approach. J BiomolStructDyn (JBSD). 2016; 34: 223-35.

225. Qiu WR, Sun BQ, Xiao X, Xu ZC. iHyd-PseCp: Identify hydroxyproline and hydroxylysine in proteins by incorporating sequence-coupled effects into general PseAAC. Oncotarge. 2016; 7: 44310-21.

226. Jia J, Liu Z, Xiao X, Liu B. Identification of protein-protein binding sites by incorporating the physicochemical properties and stationary wavelet transforms into pseudo amino acid composition (iPPBS-PseAAC). J BiomolStructDyn (JBSD). 2016; 34: 1946-61.

227. Jia J, Liu Z, Xiao X, Liu B. iCar-PseCp: identify carbonylation sites in proteins by Monto Carlo sampling and incorporating sequence coupled effects into general PseAAC. Oncotarget. 2016; 7: 34558-70.

228.https://www.nature.com/articles/srep32333?proof=trueIn%EF%BB%BF&draft=journal

229. Liu B, Wu H, Zhang D, Wang X. Pse-Analysis: a python package for DNA/RNA and protein/peptide sequence analysis based on pseudo components and kernel methods. Oncotarget. 2017; 8: 13338-43.

230. Qiu WR, Jiang SY, Xu ZC, Xiao X. iRNAm5C-PseDNC: identifying RNA 5-methylcytosine sites by incorporating physical-chemical properties into pseudo dinucleotide composition. Oncotarget. 2017; 8: 41178-88.

231. Qiu WR, Jiang SY, Sun BQ, Xiao, X, Cheng X, Chou KC et al. iRNA-2methyl: identify RNA 2′-O-methylation sites by incorporating sequence-coupled effects into general PseKNC and ensemble classifier. Medicinal Chemistry. 2017; 13: 734-43.

232. Xu Y, Li C. iPreny-PseAAC: identify C-terminal cysteine prenylation sites in proteins by incorporating two tiers of sequence couplings into PseAAC. Med Chem. 2017; 13: 544-51.

234. Qiu WR, Sun BQ, Xiao X, Xu D. iPhos-PseEvo: Identifying human phosphorylated proteins by incorporating evolutionary information into general PseAAC via grey system theory. Molecular Informatics. 2017; 36: UNSP 1600010.

235. Liu LM, Xu Y. iPGK-PseAAC: identify lysine phosphoglycerylation sites in proteins by incorporating four different tiers of amino acid pairwise coupling information into the general PseAAC. Med Chem. 2017; 13: 552-9.

236. Cheng X, Zhao SG, Xiao X. iATC-mISF: a multi-label classifier for predicting the classes of anatomical therapeutic chemicals. Bioinformatics (Corrigendum, ibid. 2017; 33: 341-6.

237. Cheng X, Zhao SG, Xiao X. iATC-mHyb: a hybrid multi-label classifier for predicting the classification of anatomical therapeutic chemicals. Oncotarget. 2017; 8: 58494-503.

238. Wang J, Yang B, Leier A, Marquez-Lago TT, Hayashida M, Rocker A et al. Bastion6: a bioinformatics approach for accurate prediction of type VI secreted effectors. Bioinformatics. 2018; 34: 2546-55.

239. Liu B, Li K, Huang DS. iEnhancer-EL: Identifying enhancers and their strength with ensemble learning approach. Bioinformatics. 2018; 34: 3835-42.

240. Chen Z, Zhao PY, Li F, Leier A, Marquez-Lago TT, Wang Y et al. iFeature: a python package and web server for features extraction and selection from protein and peptide sequences. Bioinformatics. 2018; 34: 2499-502.

241. Su ZD, Huang Y, Zhang ZY, Zhao YW, Wang D, Chen W et al. iLoc-lncRNA: predict the subcellular location of lncRNAs by incorporating octamer composition into general PseKNC. Bioinformatics. 2018; 34: 4196-204.

242. Liu B, Yang F, Huang DS. iPromoter-2L: a two-layer predictor for identifying promoters and their types by multi-window-based PseKNC. Bioinformatics. 2018; 34: 33-40.

243. Liu B, Weng F, Huang DS. iRO-3wPseKNC: Identify DNA replication origins by three-window-based PseKNC. Bioinformatics. 2018; 34: 3086-93.

244. Yang H, Qiu WR, Liu G, Guo FB, Chen W, Lin, H et al. iRSpot-Pse6NC: Identifying recombination spots in Saccharomyces

cerevisiae by incorporating hexamer composition into general PseKNC International Journal of Biological Sciences. 2018; 14: 883-91.

245. Song J, Li F, Leier A, Marquez-Lago TT, Akutsu T, Haffari G et al. PROSPERous: high-throughput prediction of substrate cleavage sites for 90 proteases with improved accuracy. Bioinformatics. 2018; 34: 684-7.

246. Chou KC. Review: Progress in protein structural class prediction and its impact to bioinformatics and proteomics. Current Protein and Peptide Science. 2005; 6: 423-36.

247. Zhong WZ, Zhou SF. Molecular science for drug development and biomedicine. Intenational Journal of Molecular Sciences. 2014; 15: 20072-8.

248. Du QS, Huang RB, Wang SQ. Designing inhibitors of M2 proton channel against H1N1 swine influenza virus. PLoS ONE. 2010; 5: e9388.

249. Wang SQ, Cheng XC, Dong WL, Wang RL. Three new powerful Oseltamivir derivatives for inhibiting the neuraminidase of influenza virus. BiochemBiophys Res Commun (BBRC). 2010; 401: 188-91.

250. Li XB, Wang SQ, Xu WR, Wang RL. Novel Inhibitor Design for Hemagglutinin against H1N1 Influenza Virus by Core Hopping Method. PLoS One. 2011; 6: e28111.

251. Ma Y, Wang SQ, Xu WR, Wang RL. Design novel dual agonists for treating type-2 diabetes by targeting peroxisome proliferator-activated receptors with core hopping approach. PLoS One. 2012; 7: e38546.

252. Liu L, Ma Y, Wang RL, Xu WR, Wang SQ. Find novel dual-agonist drugs for treating type 2 diabetes by means of cheminformatics. Drug Design, Development and Therapy. 2013; 7: 279-87.

253. Lu JJ, Pan W, Hu YJ, Wang YT. Multi-target drugs: the trend of drug research and development. PLoS One. 2012; 7: e40262.

254. Chou KC, Forsen S. Diffusion-controlled effects in reversible enzymatic fast reaction system: Critical spherical shell and proximity rate constants. Biophysical Chemistry. 1980; 12: 255-63.

255. Chou KC, Li TT, Forsen S. The critical spherical shell in enzymatic fast reaction systems. Biophysical Chemistry. 1980; 12: 265-9.

256. Li TT, Forsen S. The flow of substrate molecules in fast enzyme-catalyzed reaction systems. ChemicaScripta. 1980; 16: 192-6.

257. Chou KC, Forsen S. Graphical rules for enzyme-catalyzed rate laws. Biochem J. 1980; 187: 829-35.

258. Chou KC, Forsen S, Zhou GQ. Three schematic rules for deriving apparent rate constants. ChemicaScripta. 1980; 16: 109-13.

259. Chou KC, Carter RE, Forsen S. A new graphical method for deriving rate equations for complicated mechanisms. ChemicaScripta. 1981; 18: 82-6.

260. Chou KC, Forsen S. Graphical rules of steady-state reaction systems. Can J Chem. 1981; 59: 737-55.

261. Chou KC, Chen NY, Forsen S. The biological functions of low-frequency phonons: 2. Cooperative effects. ChemicaScripta. 1981; 18: 126-32.

262. Chou KC, Jiang SP, Liu WM, Fee CH. Graph theory of enzyme kinetics: 1. Steady-state reaction system. ScientiaSinica. 1979; 22: 341-58.

263. Zhou GP, Deng MH. An extension of Chou's graphic rules for deriving enzyme kinetic equations to systems involving parallel reaction pathways. Biochem. J. 1984; 222: 169-76.

264. Chou KC. Graphic rules in steady and non-steady enzyme kinetics. J. Biol. Chem. 1989; 264: 12074-9.

265. Chou KC. Review: Applications of graph theory to enzyme kinetics and protein folding kinetics. Steady and non-steady state systems. Biophysical Chemistry. 1990; 35: 1-24.

266. Althaus IW, Chou JJ, Gonzales AJ, Diebel MR, Kezdy FJ, Romero DL et al. Steady-state kinetic studies with the non-nucleoside HIV-1 reverse transcriptase inhibitor U-87201E. J Biol Chem. 1993; 268: 6119-24.

267. Althaus IW, Gonzales AJ, Chou JJ, Diebel MR, Kezdy FJ, Romero DL et al. The quinoline U-78036 is a potent inhibitor of HIV-1 reverse transcriptase. J Biol Chem. 1993; 268: 14875-80.

268. Althaus IW, Chou JJ, Gonzales AJ, Diebel MR, Kezdy FJ, Romero DL et al. Kinetic studies with the nonnucleoside HIV-1 reverse transcriptase inhibitor U-88204E. Biochemistry. 1993; 32: 6548-54.

269. Althaus IW, Chou JJ, Gonzales AJ, Diebel MR, Kezdy FJ. Steady-state kinetic studies with the polysulfonate U-9843, an HIV reverse transcriptase inhibitor. Cellular and Molecular Life Science (Experientia). 1994; 50: 23-8.

270. Althaus IW, Chou JJ, Gonzales AJ, Diebel MR, Kezdy FJ, Romero DL et al. Kinetic studies with the non-nucleoside human immunodeficiency virus type-1 reverse transcriptase inhibitor U-90152e. Biochem. Pharmacol. 1994; 47: 2017-28.

271. Chou KC, Kezdy FJ, Reusser F. Review: Kinetics of processive nucleic acid polymerases and nucleases. Anal. Biochem. 1994; 221: 217-30.

272. Althaus IW, Franks KM, Diebel MR, Kezdy FJ, Romero DL, Thomas RC et al. The benzylthio-pyrididine U-31,355, a potent inhibitor of HIV-1 reverse transcriptase. Biochem. Pharmacol. 1996; 51: 743-50.

273. Andraos J. Kinetic plasticity and the determination of product ratios for kinetic schemes leading to multiple products without rate laws: new methods based on directed graphs. Can. J. Chem. 2008; 86: 342-57.

274. Chou KC, Shen HB. FoldRate: A web-server for predicting protein folding rates from primary sequence. The Open Bioinformatics Journal. 2009; 3: 31-50.

275. Shen HB, Song JN. Prediction of protein folding rates from primary sequence by fusing multiple sequential features Journal of Biomedical Science and Engineering (JBiSE). 2009; 2: 136-43.

276. Chou KC. Graphic rule for drug metabolism systems. Current Drug Metabolism. 2010; 11: 369-78.

277. Chou KC, Lin WZ, Xiao X. Wenxiang: a web-server for drawing wenxiang diagrams Natural Science. 2011; 3: 862-5.

278. Zhou GP. The disposition of the LZCC protein residues in wenxiang diagram provides new insights into the protein-protein interaction mechanism. J. Theor. Biol. 2011; 284: 142-8.

279. Chou KC. Proposing pseudo amino acid components are an important milestone for proteome and genome analyses. International Journal for Peptide Research and Therapeutics (IJPRT). 2019.

280. Chou KC. Impacts of pseudo amino acid components and 5-steps rule to proteomics and proteome analysis. Current Topics in Medicinak Chemistry (CTMC). 2019.

281. Chou KC. An insightful 10-year recollection since the emergence of the 5-steps rule. Current Pharmaceutical Design. 2019.

282. Chou KC. An insightful 20-year recollection since the birth of pseudo amino acid components. Applied Biochemistry and Biotechnology (ABAB). 2019.

283. Chou KC. An insightful recollection for predicting protein subcellular locations in multi-label systems. Genomics. 2019.

284. Chou KC. An insightful recollection since the birth of Gordon Life Science Institute about 17 years ago. International Journal of Peptide Research and Therapeutics (IJPRT). 2019.

285. Chou KC. An insightful recollection since the distorted key theory was born about 23 years ago. Genomics. 2019.

286. Chou KC. Two kinds of metrics for computational biology. Genomics. 2019.