http://www.scirp.org/journal/ns Natural Science, 2017, Vol. 9, (No. 9), pp: 330-349

pLoc-mGpos: Incorporate Key Gene Ontology Information into General PseAAC for Predicting Subcellular Localization of Gram-Positive Bacterial Proteins

Xuan Xiao[1,2], Xiang Cheng[1,2,3], Shengchao Su[3], Qi Mao[3], Kuo-Chen Chou[2,4,5]

[1]Computer Department, Jingdezhen Ceramic Institute, Jingdezhen, China; [2]The Gordon Life Science Institute, Boston, USA; [3]College of Information Science and Technology, Donghua University, Shanghai, China; [4]Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu, China; [5]Faculty of Computing and Information Technology in Rabigh, King Abdul Aziz University, Jeddah, Saudi Arabia

Correspondence to: Xuan Xiao, xxiao@gordonlifescience.org; Kuo-Chen Chou, kcchou@gordonlifescience.org
Keywords: Multi-Target Drugs, Gene Ontology, Chou's General PseAAC, ML-GKR, Chou's Metrics
Received: September 16, 2017 **Accepted:** September 19, 2017 **Published:** September 22, 2017

ABSTRACT

The basic unit in life is cell. It contains many protein molecules located at its different organelles. The growth and reproduction of a cell as well as most of its other biological functions are performed via these proteins. But proteins in different organelles or subcellular locations have different functions. Facing the avalanche of protein sequences generated in the postgenomic age, we are challenged to develop high throughput tools for identifying the subcellular localization of proteins based on their sequence information alone. Although considerable efforts have been made in this regard, the problem is far apart from being solved yet. Most existing methods can be used to deal with single-location proteins only. Actually, proteins with multi-locations may have some special biological functions that are particularly important for drug targets. Using the ML-GKR (Multi-Label Gaussian Kernel Regression) method, we developed a new predictor called "pLoc-mGpos" by in-depth extracting the key information from GO (Gene Ontology) into the Chou's general PseAAC (Pseudo Amino Acid Composition) for predicting the subcellular localization of Gram-positive bacterial proteins with both single and multiple location sites. Rigorous cross-validation on a same stringent benchmark dataset indicated that the proposed pLoc-mGpos predictor is remarkably superior to "iLoc-Gpos", the state-of-the-art predictor for the same purpose. To maximize the convenience of most experimental scientists, a user-friendly web-server for the new powerful predictor has been established at

http://www.jci-bioinfo.cn/pLoc-mGpos/, by which users can easily get their desired results without the need to go through the complicated mathematics involved.

1. INTRODUCTION

As the most basic unit of life, a cell must also undergo three most important processes of any living things: growth, reproduction, and death [1]. It is one of the fundamental problems in cellular and molecular biology to thoroughly understand these processes. The knowledge thus acquired is also closely associated with drug development. To realize it, however, the knowledge of proteins in different organelles of a cell or its subcellular localization is prerequisite.

During the last two decades or so, many computational methods were developed to address this problem (see [2, 3] as well as a long list of references cited in the two important review articles).

But most of the existing computational methods were designed to treat the single-label system in which each of the constituent proteins has one, and only one, subcellular location. With more experimental data emerging, however, the localization of proteins in a cell is actually a multi-label system, where some proteins may simultaneously occur in two or more different location sites. This kind of multiplex proteins often bears some exceptional biological functions [4-6], and should deserve our special attention [7-12], particularly from the viewpoint of selecting multiple targets [13-15] or key targets [16-19] for drug development.

About 10 years ago, some efforts have been made to explore this kind of multiplex protein systems [6, 7, 10, 12, 20-30]. In comparison with the single-label systems, it would be much more difficult and complicated to deal with the multi-label systems. Particularly, it is extremely difficult for a multi-label predictor to yield a descent result for the "absolute true" rate. The reason is as follows. Suppose a gram-positive bacterial protein is labeled with "1" and "2", meaning that it may simultaneously exist in subcellular locations 1 and 2 in the real world. If its predicted result is "1", or "2", or "1 and 3", or "2 and 3", no score at all will be added for the absolute true rate. When and only when the predicted result is also exactly "1 and 2" meaning perfectly identical to the actual labels, will one score be added in calculating the absolute true rate. Therefore, it is the harshest metrics in measuring the quality of a multi-label predictor [31]. And that was why in proposing their multi-label predictors, many authors even did not mention the term of "absolute true rate".

In this study, we used the multi-label theory [31] to develop a new predictor to identify the subcellular localization of Gram-positive bacterial proteins aimed at improving its absolute true and absolute false rates, the two most important and harshest metrics for a multi-label predictor [31].

2. MATERIALS AND METHODS

2.1. Benchmark Dataset

According to the Chou's 5-step rule [32] that has been widely used by many recent investigators (see, e.g., [33-47]) for developing a statistical predictor, the first important and foremost thing is to construct or select a valid benchmark dataset to train and test the model [1, 42, 48]. In literature, the benchmark dataset usually consists of a training dataset and a testing dataset: the former is for the purpose of training a proposed model, while the latter for the purpose of testing it. But as elucidated in [3], it would suffice with one good quality benchmark dataset if the model is tested by the jackknife or subsampling (K-fold cross-validation) test because the outcome thus obtained is actually from a combination of many different independent dataset tests. In this study, the benchmark dataset was taken from [21, 27]. The reasons to do so are as follows: 1) The dataset contains statistically significant number of Gram-positive bacterial proteins with both single location and multiple locations confirmed by experiments. Besides, none of the proteins included has ≥25% pairwise sequence identity to any other in a same subset, which is important for reducing homologous bias. 2) It is also the same benchmark dataset used to train and test iLoc-Gpos [27],

the state-of-the-art predictor in this area, and hence will make the comparison based on the same condition and same criteria. For readers' convenience, the benchmark dataset is given in Supporting Information S1. It contains $N(\text{seq})=519$ sequence-different Gram-positive bacterial proteins classified into 4 subsets according to their subcellular locations. An overall view of these proteins in the 4 subcellular locations is given in Supporting Information S2, from which we can see that, of the 519 different Gram-positive bacterial proteins, 515 belong to one location, and 4 to two locations.

A breakdown of the $N(\text{seq})=519$ Gram-positive bacterial proteins according to their occurrences in the 4 different subcellular locations is given in **Table 1**, where

$$N(\text{vir})=\sum_{k=1}^{N(\text{seq})} n^{\text{L}}(k) \tag{1}$$

is the total number of "virtual proteins" [22, 49] or "locative proteins" [28] in the benchmark dataset, and $n^{\text{L}}(k)$ is the number of different labels (or subcellular locations) marked on the k-th sequence-different Gram-positive bacterial protein. Accordingly, the multiplicity degree MD [31] of the current benchmark dataset is

$$\text{MD}=\frac{\sum_{k=1}^{N(\text{seq})} n^{\text{L}}(k)}{N(\text{seq})}=\frac{N(\text{vir})}{N(\text{seq})}=1.008 \tag{2}$$

As we can see from Equation (2), $\text{MD}=1$ means the system containing no protein with more than one location, while $\text{MD}>1$ means some proteins having more than one location. The higher the value of MD, the more protein samples that have multiple locations or labels.

For simplify the description later, the benchmark dataset is denoted by $\mathbb{S}$, which can be further formulated as

$$\mathbb{S}=\mathbb{S}_1\cup\mathbb{S}_2\cup\mathbb{S}_3\cup\mathbb{S}_4 \tag{3}$$

where $\mathbb{S}_1$ only contains the Gram-positive bacterial protein samples from the "Cell membrane" location (cf. **Table 1**), $\mathbb{S}_2$ only contains those from the "Cell wall" location, $\mathbb{S}_3$ only contains those from the "Cytoplasm" location, and $\mathbb{S}_4$ only contains those from the "Extracell" location; $\cup$ denotes the symbol for "union" in the set theory.

Table 1. Breakdown of the Gram-positive bacterial proteins in the benchmark dataset $\mathbb{S}$ into 4 subsets according to their different subcellular localizations (cf. Supporting Information S1 and Supporting Information S2).

Subset	Subcellular location name	Number of proteins
$\mathbb{S}_1$	Cell membrane	174
$\mathbb{S}_2$	Cell wall	18
$\mathbb{S}_3$	Cytoplasm	208
$\mathbb{S}_4$	Extracell	123
	Total number of virtual proteins $N(\text{vir})$[a]	523
	Total number of proteins with different sequences N	519
	The multiplicity degree MD[b]	1.008

[a]See Equation (1) and the relevant text for the definition of the number of virtual proteins; [b]See Equation (2) for the definition of multiplicity degree.

2.2. Proteins Sample Formulation

Now let us consider the 2nd step of the Chou's 5-step rule [32]; *i.e.*, how to formulate the biological sequence samples with an effective mathematical expression that can truly reflect their essential correlation with the target concerned. Given a Gram-positive bacterial protein sequence **P**, its most straightforward expression is

$$\mathbf{P} = \mathrm{R}_1\mathrm{R}_2\mathrm{R}_3\mathrm{R}_4\mathrm{R}_5\mathrm{R}_6\mathrm{R}_7\cdots\mathrm{R}_L \tag{4}$$

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, such as SVM (Support Vector Machine) [36], KNN (K-Nearest Neighbor) [50], and RF (Random Forest) [51], can only handle vectors [52], we have to convert the sequential expression of Equation (4) into a vector. But a vector defined in a discrete model might completely lose all the sequence-order information. To deal with this problem, the PseAAC (Pseudo Amino Acid Composition) was introduced [53-55]. Ever since the concept of pseudo amino acid composition or Chou's PseAAC [55-58] was proposed, it has been widely used in many biomedicine and drug development areas [59, 60] as well as nearly all the areas of computational proteomics(see, e.g., [39, 43, 45, 61-73] and a long list of references cited in two review papers [74, 75]). Encouraged by the successes of using PseAAC to deal with protein/peptide sequences, its idea and approach have been extended to deal with DNA/RNA sequences [76-82] in computational genomics via PseKNC (Pseudo K-tuple Nucleotide Composition) [83, 84]. Recently, a very powerful web-server called "Pse-in-One" [85] and its updated version "Pse-in-One 2.0" [86] were developed, by which users can generate any pseudo components for both protein/peptide and DNA/RNA sequences as they wish or define.

According to the concept of Chou's general PseAAC [32], any protein sequence can be formulated as a PseAAC vector given by

$$\mathbf{P} = \left[\Psi_1\ \Psi_2\ \cdots\ \Psi_u\ \cdots\ \Psi_\Omega\right]^{\mathrm{T}} \tag{5}$$

where $\mathbf{T}$ 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**, as elaborated below.

Being one type of general PseAAC [32], the GO (Gene Ontology) has been widely used to improve the prediction quality of protein subcellular localization (see, e.g., [23, 25, 26, 87-91]). The advantage of using the GO approach is that proteins mapped into the GO space (instead of Euclidean space or any other simple geometric space) would be better clustered according to their subcellular locations, as elaborated in [9, 92]. For the rationale of using the GO approach to predict the protein subcellular localization, and an incisive discussion/analysis to justify the GO approach, see Section VI in a comprehensive review paper [31].

However, the existing GO approaches (see, e.g., [10, 23, 25, 26, 87]) have the following shortcomings. 1) Only the digital numbers 0 and 1 (or their simple combination) were used to incorporate the GO information, and hence some important information may be missed. 2) The dimension of the protein vectors, namely Ω of Equation (5), in the previous GO approaches was very high; e.g., it is 1,930 in [88] and 9567 in [93], and hence may lead to the "curse of dimensionality" or "high-dimension disaster" problem [94].

Here, we are to introduce a novel GO approach, through which we can extract the key information by winnowing many trivial ones so as to significantly reduce the dimension of PseAAC vector of Equation (5). The detailed procedures are as follows.

Step 1. Use BLAST to search all the Gram-positive bacterial proteins in the Swiss-Prot database for those proteins that have high homology (*i.e.*, more than 60% pairwise sequence identity) with the protein **P** of Equation (4). The proteins thus obtained are collected into a subset, $\mathbb{S}_{\mathbf{P}}^{\mathrm{homo}}$, called the homology set of **P**. Subsequently, retrieve the GO codes of the protein in $\mathbb{S}_{\mathbf{P}}^{\mathrm{homo}}$ that has the highest homology with **P**.

Each of the GO codes is a numerical label containing 7-digit figure (see, e.g., [88]). If it has no GO code at all, do the same for the 2nd highest homologous protein in $\mathbb{S}_{\mathbf{P}}^{\text{homo}}$; if it has no GO gode again, do the same for the 3rd highest homologous one; go on like this until obtaining a GO code or a set of GO codes as given below

$$\left\{\text{GO}_1^{\mathbf{P}}\ \text{GO}_2^{\mathbf{P}}\ \cdots\ \text{GO}_k^{\mathbf{P}}\ \cdots\ \text{GO}_{n^{\text{g}}}^{\mathbf{P}}\right\} \tag{6}$$

where $\text{GO}_k^{\mathbf{P}}\left(k=1,2,\cdots,n^{\text{g}}\right)$ is the k-th GO code for the protein in $\mathbb{S}_{\mathbf{P}}^{\text{homo}}$ that has first been found with a set of GO codes according to the aforementioned order, and n^{g} is the total number of the GO codes it has. Suppose we find from the training dataset that the total number of proteins having exactly the same GO code as $\text{GO}_k^{\mathbf{P}}$ is $N(k)$, of which the number of proteins in the u-th subset is

$$n(k,u)\left(k=1,2,\cdots,n^{\text{g}};u=1,2,\cdots,L_{\text{cell}}\right) \tag{7}$$

where $L_{\text{cell}}=4$ is the total number of subcellular locations investigated (see Equation (2) or **Table 1**).

Step 2. Based on Equation (7), the general PseAAC vector in Equation (5) and its dimension can be uniquely defined as

$$\Psi_u=\underset{1\le k\le n^{\text{g}}}{\text{Max}}\left[\frac{n(u,k)}{N(k)}\right]\ \left(u=1,2,\cdots,\Omega=L_{\text{cell}}=4\right) \tag{8}$$

where $N(k)$ is the total number of Gram-positive bacterial proteins in the training dataset that have the same GO number as $\text{GO}_k^{\mathbf{P}}$ and the operator Max means taking the maximum value among those with respect to different k. It is through such optimization operation to extract the most important GO information for the current study and screen out many trivial GO codes to significantly reduce the PseAAC vector's dimension.

Listed in Supporting Information S3 are the PseAAC vectors defined by Equation (8) for the 519 sequence-different Gram-positive bacterial proteins in Supporting Information S1, respectively. As we can see there, the dimension of the current PseAAC vectors has been reduced to 4, about thousand times lower than those in the previous approaches [21, 27, 88, 93]. This is really a big breakthrough in using GO approach to predict protein subcellular localization.

2.3. Operation Algorithm

The 3rd step in the Chou's 5-step rule [32] is about the operation algorithm (or engine) to run the prediction. Here, we adopted the ML-GKR (multi-label Gaussian kernel regression) classifier, as described below.

According to Equation (8) or Supporting Information S3, the i-th Gram-positive bacterial protein $\mathbf{P}^i$ in the benchmark dataset $\mathbb{S}$ of Equation (3) can be formulated as

$$\mathbf{P}_{\text{GO}}^i=\left[\Psi_1^i\ \ \Psi_2^i\ \ \Psi_3^i\ \ \Psi_4^i\right]^{\text{T}},i=1,2,\cdots,N\left(\text{seq}\right) \tag{9}$$

Now let us use the 4-D vector $\mathbf{L}^i$ to describe its subcellular location(s) in the multi-label system; *i.e.*,

$$\mathbf{L}^i=\left[\ell_1^i\ \ \ell_2^i\ \ \ell_3^i\ \ \ell_4^i\right]^{\text{T}} \tag{10}$$

where

$$\ell_u^i=\begin{cases}+1 & \text{if } \mathbf{P}^i\in\mathbb{S}_u\\ -1 & \text{otherwise}\end{cases}\ \left(u=1,2,3,4\right) \tag{11}$$

Likewise, for a query Gram-positive bacterial protein $\mathbf{P}^{\text{q}}$ we have

$$\mathbf{P}^{\text{q}}=\left[\Psi_1^{\text{q}}\ \ \Psi_2^{\text{q}}\ \ \Psi_3^{\text{q}}\ \ \Psi_4^{\text{q}}\right]^{\text{T}} \tag{12}$$

Its subcellular location label (s) in the multi-label system should be accordingly given by

$$\mathbf{L}^{\mathrm{q}} = \left[\ell_1^{\mathrm{q}} \;\; \ell_2^{\mathrm{q}} \;\; \ell_3^{\mathrm{q}} \;\; \ell_4^{\mathrm{q}} \right]^{\mathrm{T}} \tag{13}$$

where

$$\ell_u^q = \begin{cases} +1 & \text{if } \Delta_u \geq 0 \\ -1 & \text{otherwise} \end{cases} \quad (u = 1,2,3,4) \tag{14}$$

The Δ_u in Equation (13) is given by

$$\Delta_u = \left[\sum_{i=1}^{N(\text{train})} \ell_u^i \cdot \exp\left(-\frac{\left\| \mathbf{P}^{\mathrm{q}} - \mathbf{P}^i \right\|^2}{2\theta^2} \right) \right] \left[\sum_{i=1}^{N(\text{train})} \exp\left(-\frac{\left\| \mathbf{P}^{\mathrm{q}} - \mathbf{P}^i \right\|^2}{2\theta^2} \right) \right]^{-1} \tag{15}$$

where N(train) is the number of proteins used to train the model, θ is a parameter whose optimal value will be determined later, and $\left\| \mathbf{P}^{\mathrm{q}} - \mathbf{P}^i \right\|^2$ is the Euclidean distance [95] between the query protein (Equation (12) and the i-th protein(Equation (9) in the benchmark dataset $\mathbb{S}$; *i.e.*,

$$\left\| \mathbf{P}_{\mathrm{GO}}^{\mathrm{q}} - \mathbf{P}_{\mathrm{GO}}^{\boldsymbol{i}} \right\|^2 = \sum_{u=1}^{4} \left(\Psi_u^{\mathrm{q}} - \Psi_u^i \right)^2 \tag{16}$$

Thus, the location label vector $\mathbf{L}^{\mathrm{q}}$ of Equation (13) for the query Gram-positive bacterial protein $\mathbf{P}^{\mathrm{q}}$ is well defined, and hence its subcellular location or locations can be explicitly predicted as well. For example: if $\ell_1^{\mathrm{q}} = \ell_2^{\mathrm{q}} = +1$ while all the other components in Equation (13) are equal to -1, this means that the query Gram-positive bacterial protein $\mathbf{P}^{\mathrm{q}}$ is located in the 1st and 2nd subcellular locations (cf. Table 1); if $\ell_3^{\mathrm{q}} = +1$ while all the others are equal to -1, meaning that the query Gram-positive bacterial protein is located in the 3rd subcellular location only; and so forth.

The predictor developed via the aforementioned procedures is called pLoc-mGpos, where "pLoc" stands for "predict subcellular localization", and "mGpos" for "multi-label Gram-positive bacterial proteins". Shown in Figure 1 is a flowchart to illustrate the process of how the pLoc-mGpos is working.

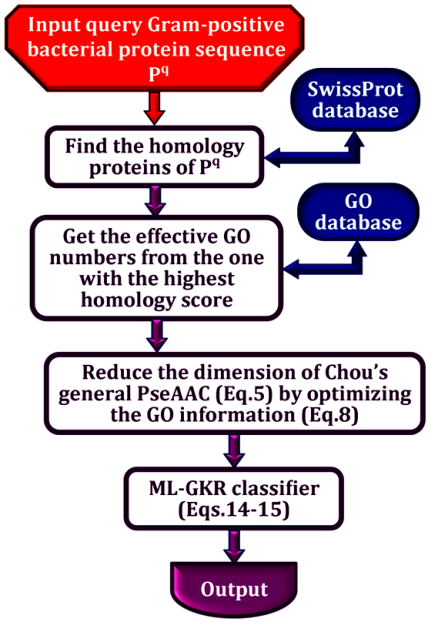

Figure 1. A flowchart to show the process of how the pLoc-mGpos predictor works.

3. RESULTS AND DISCUSSION

As mentioned in the Chou's 5-step rule [32], one of the important procedures in developing a new predictor is how to objectively evaluate its anticipated accuracy. To address this, two issues need to be considered. 1) What metrics should be used to quantitatively reflect the predictor's quality? 2) What test approach should be adopted to count the metrics scores?

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

Different from the metrics used to measure the prediction quality of single-label systems, the metrics for the multi-label systems are much more complicated. To make them more intuitive and easier to understand for most experimental scientists, here we adopt the following intuitive Chou's five metrics [31] that have recently been widely used for studying various multi-label systems (see, e.g., [30, 39, 44, 50, 96-101]):

$$\begin{cases} \text{Aiming} \uparrow = \dfrac{1}{N^{\mathrm{q}}}\sum_{k=1}^{N^{\mathrm{q}}}\left(\dfrac{\left\|\mathbb{L}_k \cap \mathbb{L}_k^*\right\|}{\left\|\mathbb{L}_k^*\right\|}\right), \quad [0,1] \\ \text{Coverage} \uparrow = \dfrac{1}{N^{\mathrm{q}}}\sum_{k=1}^{N^{\mathrm{q}}}\left(\dfrac{\left\|\mathbb{L}_k \cap \mathbb{L}_k^*\right\|}{\left\|\mathbb{L}_k\right\|}\right), \quad [0,1] \\ \text{Accuracy} \uparrow = \dfrac{1}{N^{\mathrm{q}}}\sum_{k=1}^{N^{\mathrm{q}}}\left(\dfrac{\left\|\mathbb{L}_k \cap \mathbb{L}_k^*\right\|}{\left\|\mathbb{L}_k \cup \mathbb{L}_k^*\right\|}\right), \quad [0,1] \\ \text{Absolute true} \uparrow = \dfrac{1}{N^{\mathrm{q}}}\sum_{k=1}^{N^{\mathrm{q}}}\Delta\left(\mathbb{L}_k, \mathbb{L}_k^*\right), \quad [0,1] \\ \text{Absolute false} \downarrow = \dfrac{1}{N^{\mathrm{q}}}\sum_{k=1}^{N^{\mathrm{q}}}\left(\dfrac{\left\|\mathbb{L}_k \cup \mathbb{L}_k^*\right\| - \left\|\mathbb{L}_k \cap \mathbb{L}_k^*\right\|}{M}\right), \quad [1,0] \end{cases} \tag{17}$$

where N^{q} is the total number of query proteins or tested proteins, M is the total number of different labels for the investigated system (for the current study it is $L_{\text{cell}} = 4$), $\|\ \|$ means the operator acting on the set therein to count the number of its elements, $\cup$ means the symbol for the "union" in the set theory, $\cap$ denotes the symbol for the "intersection", $\mathbb{L}_k$ denotes the subset that contains all the labels observed by experiments for the k-th tested sample, $\mathbb{L}_k^*$ represents the subset that contains all the labels predicted for the k-th sample, and

$$\Delta\left(\mathbb{L}_k, \mathbb{L}_k^*\right) = \begin{cases} 1, & \text{if all the labels in } \mathbb{L}_k^* \text{ are identical to those in } \mathbb{L}_k \\ 0, & \text{otherwise} \end{cases} \tag{18}$$

In Equation (17), the first four metrics with an upper arrow $\uparrow$ are called positive metrics, meaning that the larger the rate is the better the prediction quality will be; the 5th metrics with a down arrow $\downarrow$ is called negative metrics, implying just the opposite meaning.

From Equation (17) we can see the following: 1) the "Aiming" defined by the 1st sub-equation is for checking the rate or percentage of the correctly predicted labels over the practically predicted labels; 2) the "Coverage" defined in the 2nd sub-equation is for checking the rate of the correctly predicted labels over the actual labels in the system concerned; 3) the "Accuracy" in the 3rd sub-equation is for checking the average ratio of correctly predicted labels over the total labels including correctly and incorrectly predicted labels as well as those real labels but are missed in the prediction; 4) the "Absolute true" in the 4th sub-equation is for checking the ratio of the perfectly or completely correct prediction events over the total prediction events; 5) the "Absolute false" in the 5th sub-equation is for checking the ratio of the completely

wrong prediction over the total prediction events.

3.2. 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 [95]. Of these three, however, the jackknife test is deemed the least arbitrary that can always yield a unique outcome for a given benchmark dataset as elucidated in [32]. Accordingly, the jackknife test has been widely recognized and increasingly used by investigators to examine the quality of various predictors (see, e.g., [35, 39, 41, 63, 65, 102-105]). Accordingly, the jackknife test was also used in this study.

3.3. Parameter Determination

Since Equation (15) contains a parameter θ, the predicted results obtained by pLoc-mGpos will depend on the parameter's value. In this study, the optimal value for θ was determined by maximizing the absolute true rate (see the 4th sub-equation in Equation (17) by the jackknife validation on the benchmark dataset. As shown in **Figure 2**, when $\theta = 1/8$, the absolute true rate reached its highest score. And such a value would be used for further study.

3.4. Comparison with the State-of-the-Art Predictor

Listed in **Table 2** are the rates obtained by the current pLoc-Gpos predictor via the jackknife test on the benchmark dataset (Supporting Information S1). For facilitating comparison, listed in that table are also the corresponding results obtained by the iLoc-Gpos [27] and Gpos-mPLoc [21], the two existing most powerful predictors for identifying the subcellular localization of Gram-positive bacterial proteins with both single and multiple sites.

As shown in **Table 2**, among the five metrics in Equation (17) used to quantitatively measure the quality of a multi-label predictor [31], the rates for "Aiming", "Accuracy", and "Absolute false" by iLoc-Gpos [27] and Gpos-mPLoc [21] were missed, indicating lack of rigorousness in checking the prediction quality. In other words, the authors of the two previous predictors only reported the rates for "Coverage" and "Absolute true". But even though, their reported success rates are remarkably lower than the corresponding rates achieved by the current predictor pLoc-mGpos proposed in this paper.

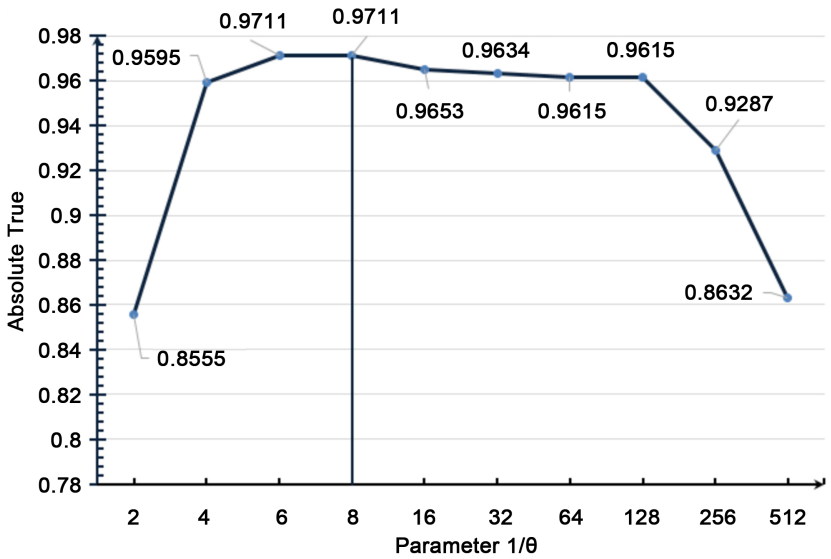

Figure 2. A plot to show the process of finding the optimal θ value in Equation (15). See the main text for further explanation.

Table 2. Comparison with the state-of-the-art methods in predicting the subcellular localization of Gram-positive bacterial proteins[a].

Predictor	Aiming (↑)[b]	Coverage (↑)[b]	Accuracy (↑)[b]	Absolute true (↑)[b]	Absolute false (↓)[b]
pLoc-mGpos[c]	97.69%	97.13 %	97.4 %	97.11%	0.14%
iLoc-Gpos[d]	N/A	93.12%	N/A	92.87%	N/A
Gpos-mPLoc[d]	N/A	82.20%	N/A	N/A	N/A

[a]The rates listed below were derived by the jackknife test on the benchmark dataset $\mathbb{S}$ (Supporting Information S1); [b]See Equation (17) for the definition of the metrics; [c]The predictor proposed in this paper with the parameter $\theta = 1/8$; [d]The predictor proposed in [27]; [e]The predictor proposed in [21].

As pointed out in a comprehensive review [31], among the aforementioned five metrics listed in **Table 2**, the most important are "absolute true" and "absolute false". It is extremely difficult for a multi-label predictor to enhance its absolute true rate and lower down its absolute false rate. Therefore, in developing methods for predicting subcellular localization of proteins with both single location site and multiple location sites, many investigators even did not mention the "absolute true" and "absolute false" rates. In contrast to that, it has been clearly reported in **Table 2** that the absolute true rate achieved by the current pLoc-mGpos predictor can reach as high as over 97%, while its absolute false rate is only 0.14% meaning that the error rate is extremely low.

Furthermore, in both the iLoc-Gpos paper [27] and the Gpos-mPLoc paper [21], no detailed scores whatsoever were given for the four metrics [106] widely used in studying various classifications. To make it up, let us introduce the following set of metrics:

$$\begin{cases} \mathrm{Sn}(i) = 1 - \dfrac{N_-^+(i)}{N^+(i)} & 0 \le \mathrm{Sn} \le 1 \\ \mathrm{Sp}(i) = 1 - \dfrac{N_+^-(i)}{N^-(i)} & 0 \le \mathrm{Sp} \le 1 \\ \mathrm{Acc}(i) = 1 - \dfrac{N_-^+(i) + N_+^-(i)}{N^+(i) + N^-(i)} & 0 \le \mathrm{Acc} \le 1 \\ \mathrm{MCC}(i) = \dfrac{1 - \left(\dfrac{N_-^+(i)}{N^+(i)} + \dfrac{N_+^-(i)}{N^-(i)} \right)}{\sqrt{\left(1 + \dfrac{N_+^-(i) - N_-^+(i)}{N^+(i)} \right) \left(1 + \dfrac{N_-^+(i) - N_+^-(i)}{N^-(i)} \right)}} & -1 \le \mathrm{MCC} \le 1 \\ (i = 1, 2, \cdots, 20) \end{cases} \tag{19}$$

where Sn, Sp, Acc, and MCC represent the sensitivity, specificity, accuracy, and Mathew's correlation coefficient, respectively [106], and i denotes the i-subcellular location in the benchmark dataset. $N^+(i)$ is the total number of the samples investigated in the i-th subset, whereas $N_-^+(i)$ is the number of the samples in $N^+(i)$ that are incorrectly predicted to be of other locations; $N^-(i)$ is the total number of samples in any 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. The metrics of Equation (19) have been widely used to examine the quality of predictors in genome/proteome analysis (see, e.g., [46, 47, 76-80, 107-109]) and computational biomedicine (see, e.g., [82, 110-112]).

Given in Table 3 are the corresponding results obtained by pLoc-mGpos for each of the four subcellular locations. As we can see from the table, all the scores are within the region of 0.8374 to 0.9924, fully consistent with its overall performance as reported in Table 2.

The above compelling facts have clearly demonstrated that the new iLoc-mGpos predictor is indeed very powerful for predicting the subcellular localization of multi-label Gram-positive bacterial proteins.

Table 3. Performance of pLoc-mGpos for each of the four subcellular locations.

i	Subcellular location[a]	Sn(i)[b]	Sp(i)[b]	Acc(i)[b]	MCC(i)[b]
1	Cell membrane	0.9598	0.9884	0.9788	0.9523
2	Cell wall	0.8889	0.992	0.9884	0.8374
3	Cytoplasm	0.9856	0.9871	0.9865	0.9719
4	Extracell	0.9512	0.9924	0.9827	0.9518

[a]See Table 1 and the relevant context for further explanation; [b]See Equation (19) for the metrics definition.

3.5. Web Server and User Guide

As pointed out in [113], user-friendly and publicly accessible web-servers represent the future direction for developing practically more useful predictors or any computational tools. Actually, user-friendly web-servers as shown in a series of recent publications [40, 46, 100, 107-112, 114-122] will significantly enhance the impacts of theoretical work because they can attract the broad experimental scientists [52]. In view of this, the web-server for the new predictor pLoc-mGpos has been established at http://www.jci-bioinfo.cn/pLoc-mGpos/. Moreover, to maximize the convenience of most experimental scientists, a step-by-step guide of how to use the web-server to get their desired results is given in given below.

Step 1. Opening the web-server at http://www.jci-bioinfo.cn/pLoc-mGpos/, you will see the top page of pLoc-mGposon your computer screen, as shown in **Figure 3**. Click on the Read Me button to see a brief introduction about the predictor.

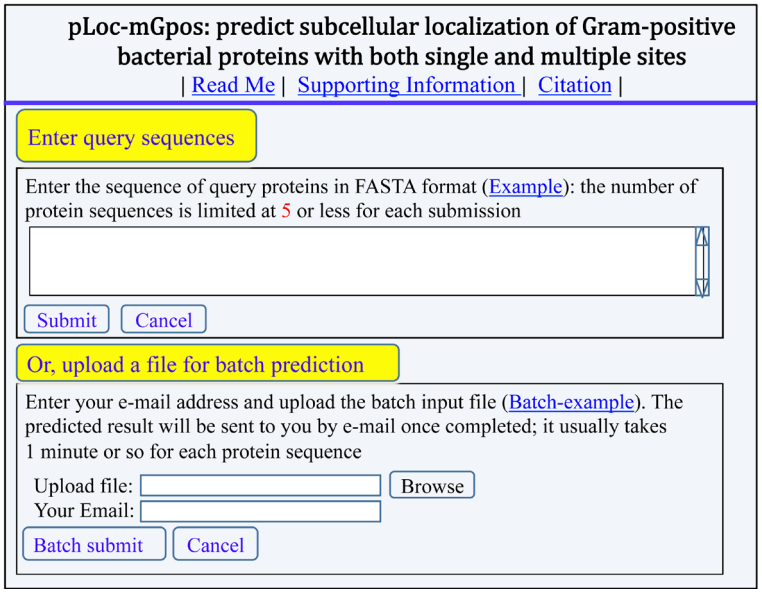

Figure 3. A semi screenshot for the top page of pLoc-mGpos web-server predictor.

Step 2. Either type or copy/paste the sequences of query Gram-positive bacterial proteins into the input box at the center of Figure 3. The input sequence should be in the FASTA format. For the examples of sequences in FASTA format, click the Example button right above the input box.

Step 3. Click on the Submit button to see the predicted result. For instance, if you use the three protein sequences in the Example window as the input, after 10 seconds or so, you will see the following on the screen of your computer (Figure 4). 1) The names of the subcellular locations numbered from1 to 4 covered by the current predictor are shown on the top. 2) The query protein Q93QY7 of example-1 corresponds to "1" meaning it belonging to "cell membrane" only; the query protein P60611 of example-2 corresponds to "3" meaning it belonging to "cytoplasm" only; the query protein P25959 of example-3 corresponds to "1, 4", meaning it belonging to "cell membrane" and "extracell; the query protein P34020 of example-4 corresponds to "3, 4", meaning it belonging to "cytoplasm" and "extracell". All these results are fully consistent with experimental observations.

Step 4. As shown on the lower panel of Figure 3, you may also choose the batch prediction by entering your e-mail address and your desired batch input file (in FASTA format of course) via the "Browse" button. To see the sample of batch input file, click on the button Batch-example. After clicking the button Batch-submit, you will see "Your batch job is under computation; once the results are available, you will be notified by e-mail."

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

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

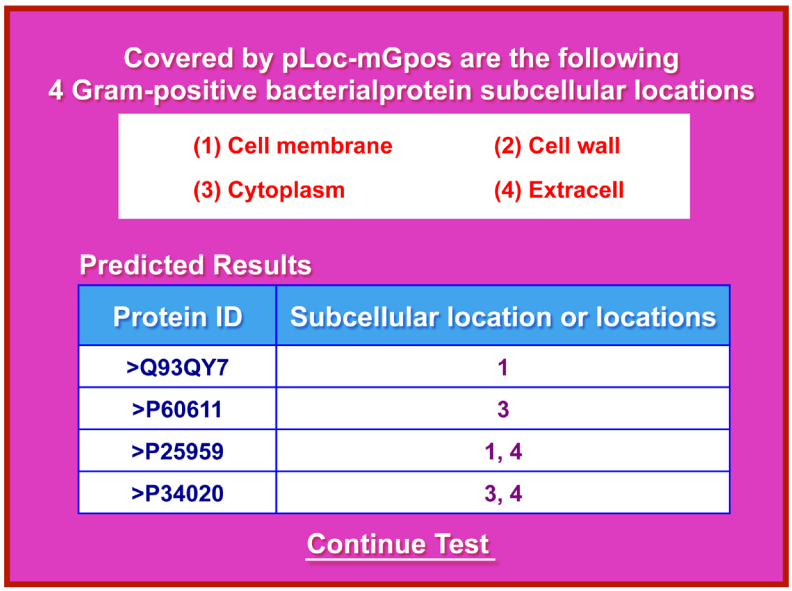

Figure 4. A semi screenshot for the webpage obtained by executing Step 3 of Section 3.5.

4. CONCLUSION

Gram-positive bacterial protein subcellular location prediction is a challenging problem, particularly when the query Gram-positive bacterial proteins have multi-label features meaning that they may occur at two or more different location sites. Here, we have developed a new predictor called pLoc-mGpos by incorporating the key GO information into Chou's general PseAAC [32]. Compared with iLoc-Gpos [27], the existing most powerful predictor that also has the capacity to deal with the multiple locations of Gram-positive bacterial proteins, the success scores achieved by the new predictor are overwhelmingly better according to the metrics widely used to measure the quality of multi-label predictors.

Why could the new predictor be so powerful? The key is that the PseAAC vectors used in the new predictor has been optimized via Equation (8) to substantially reduce their dimension but mean while significantly better reflect the correlation with the desired targets. The novel approach represents a revolutionary breakthrough in using the GO approach for predicting the subcellular localization of proteins with both single location site and multiple location sites.

It is anticipated that pLoc-mGpos will become a very useful high throughput tool for both basic research and drug development.

ACKNOWLEDGMENTS

This work was supported by the grants from the National Natural Science Foundation of China (No. 31560316, 61261027, 61262038, 61202313 and 31260273), the Province National Natural Science Foundation of JiangXi (No. 20132BAB201053), the Jiangxi Provincial Foreign Scientific and Technological Cooperation Project (No.20120BDH80023), the Department of Education of JiangXi Province (GJJ160866). This paper was partially supported by National Natural Science Foundation of China (No. 61271114 and No. 61203325) and Innovation Program of Shanghai Municipal Education Commission (No. 14ZZ068).

REFERENCES

1. Zhang, Z.D., Liang, K., Li, K., Wang, G.Q., Zhang, K.W. and Cai, L. (2017) Chlorella Vulgaris Induces Apoptosis of Human Non-Small Cell Lung Carcinoma (NSCLC) Cells. *Medicinal Chemistry*, **13**, 560-568. https://doi.org/10.2174/1573406413666170510102024
2. Nakai, K. (2000) Protein Sorting Signals and Prediction of Subcellular Localization. *Advances in Protein Chemistry*, **54**, 277-344. https://doi.org/10.1016/S0065-3233(00)54009-1
3. Chou, K.C. and Shen, H.B. (2007) Review: Recent Progresses in Protein Subcellular Location Prediction. *Analytical Biochemistry*, **370**, 1-16. https://doi.org/10.1016/j.ab.2007.07.006
4. Glory, E. and Murphy, R.F. (2007) Automated Subcellular Location Determination and High-Throughput Microscopy. *Developmental Cell*, **12**, 7-16. https://doi.org/10.1016/j.devcel.2006.12.007
5. Chou, K.C. and Shen, H.B. (2006) Addendum to "Hum-PLoc: A Novel Ensemble Classifier for Predicting Human Protein Subcellular Localization". *Biochemical and Biophysical Research Communications* (*BBRC*), **348**, 1479. https://doi.org/10.1016/j.bbrc.2006.08.030
6. 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
7. 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
8. 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.
9. Chou, K.C. and Shen, H.B. (2010) Cell-PLoc 2.0: An Improved Package of Web-Servers for Predicting Subcellular Localization of Proteins in Various Organisms. *Natural Science*, **2**, 1090-1103.
10. Shen, H.B. and Chou, K.C. (2007) Hum-mPLoc: An Ensemble Classifier for Large-Scale Human Protein Subcellular Location Prediction by Incorporating Samples with Multiple Sites. *Biochem Biophys Res Commun* (*BBRC*), **355**, 1006-1011.
11. 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

12. Shen, H.B. and Chou, K.C. (2010) Virus-mPLoc: A Fusion Classifier for Viral Protein Subcellular Location Prediction by Incorporating Multiple Sites. *Journal of Biomolecular Structure and Dynamics* (*JBSD*), **28**, 175-186. https://doi.org/10.1080/07391102.2010.10507351

13. Ma, Y., Wang, S.Q., Xu, W.R. and Wang, R.L. (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

14. 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.

15. Du, Q.S., Wang, S.Q., Xie, N.Z., Wang, Q.Y. and Huang, R.B. (2017) 2L-PCA: A Two-Level Principal Component Analyzer for Quantitative Drug Design and Its Applications. *Oncotarget*, **8**, 70564-70578. https://doi.org/10.18632/oncotarget.19757

16. 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

17. Du, Q.S., Huang, R.B., Wang, C.H. and Li, X.M. (2009) Energetic Analysis of the Two Controversial Drug Binding Sites of the M2 Proton Channel in Influenza A Virus. *Journal of Theoretical Biology*, **259**, 159-164. https://doi.org/10.1016/j.jtbi.2009.03.003

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

19. Li, X.B., Wang, S.Q., Xu, W.R. and Wang, R.L. (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

20. Chou, K.C. and Shen, H.B. (2006) Predicting Protein Subcellular Location by Fusing Multiple Classifiers. *Journal of Cellular Biochemistry*, **99**, 517-527. https://doi.org/10.1002/jcb.20879

21. 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

22. 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

23. 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

24. Wan, S.B., Hu, L.L., Niu, S., Wang, K. and Cai, Y.D. (2011) Identification of Multiple Subcellular Locations for Proteins in Budding Yeast. *Current Bioinformatics*, **6**, 71-80. https://doi.org/10.2174/157489311795222374

25. 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

26. 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

27. 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

28. 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

29. 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

30. 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

31. 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

32. Chou, K.C. (2011) Some Remarks on Protein Attribute Prediction and Pseudo Amino Acid Composition (50th Anniversary Year Review). *Journal of Theoretical Biology*, **273**, 236-247. https://doi.org/10.1016/j.jtbi.2010.12.024

33. Xu, Y., Ding, J. and Wu, L.Y. (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

34. Jia, J., Liu, Z. and Xiao, X. (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

35. 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

36. Chen, J., Long, R., Wang, X.L. and Liu, B. (2016) dRHP-PseRA: Detecting Remote Homology Proteins Using Profile-Based Pseudo Protein Sequence and Rank Aggregation. *Scientific Reports*, **6**, 32333. https://doi.org/10.1038/srep32333

37. Jia, J., Liu, Z., Xiao, X. and Liu, B. (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

38. Liu, B., Long, R. and Chou, K.C. (2016) iDHS-EL: Identifying DNase I Hypersensitive Sites 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

39. 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**, 42362. https://doi.org/10.1038/srep42362

40. Qiu, W.R., Sun, B.Q., Xiao, X. and Xu, Z.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

41. 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

42. Su, Q., Lu, W., Du, D., Chen, F. and Niu, B. (2017) Prediction of the Aquatic Toxicity of Aromatic Compounds to Tetrahymena Pyriformis through Support Vector Regression. *Oncotarget*, **8**, 49359-49369. https://doi.org/10.18632/oncotarget.17210

43. Rahimi, M., Bakhtiarizadeh, M.R. and Mohammadi-Sangcheshmeh, A. (2017) OOgenesis_Pred: A Sequence-Based Method for Predicting Oogenesis Proteins by Six Different Modes of Chou's Pseudo Amino Acid Composition. *Journal of Theoretical Biology*, **414**, 128-136. https://doi.org/10.1016/j.jtbi.2016.11.028

44. 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. https://doi.org/10.1093/bioinformatics/btx387

45. 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

46. Qiu, W.R., Jiang, S.Y. and Xu, Z.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

47. Liu, B., Yang, F., Huang, D.S. and Chou, K.C. (2017) iPromoter-2L: A Two-Layer Predictor for Identifying Promoters and Their Types by Multi-Window-Based PseKNC. *Bioinformatics*.

48. Niu, B., Zhang, M., Du, P., Jiang, L., Qin, R., Su, Q., Chen, F. and Du, D. (2017) Small Molecular Floribundiquinone B Derived from Medicinal Plants Inhibits Acetylcholinesterase Activity. *Oncotarget*, **8**, 57149-57162. https://doi.org/10.18632/oncotarget.19169

49. 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

50. Xiao, X., Wang, P., Lin, W.Z. and Jia, J.H. (2013) iAMP-2L: A Two-Level Multi-Label Classifier for Identifying Antimicrobial Peptides and Their Functional Types. *Analytical Biochemistry*, **436**, 168-177. https://doi.org/10.1016/j.ab.2013.01.019

51. Jia, J., Liu, Z., Xiao, X. and Liu, B. (2016) iPPBS-Opt: A Sequence-Based Ensemble Classifier for Identifying Protein-Protein Binding Sites by Optimizing Imbalanced Training Datasets. *Molecules*, **21**, 95. https://doi.org/10.3390/molecules21010095

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

53. Chou, K.C. (2000) Prediction of Protein Subcellular Locations by Incorporating Quasi-Sequence-Order Effect. *Biochemical and Biophysical Research Communications* (*BBRC*), **278**, 477-483. https://doi.org/10.1006/bbrc.2000.3815

54. Chou, K.C. (2001) Prediction of Protein Cellular Attributes Using Pseudo Amino Acid Composition. *PROTEINS: Structure, Function, and Genetics*, **43**, 246-255. https://doi.org/10.1002/prot.1035

55. 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

56. 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

57. 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

58. Lin, S.X. and Lapointe, J. (2013) Theoretical and Experimental Biology in One—A Symposium in Honour of Professor Kuo-Chen Chou's 50th Anniversary and Professor Richard Giegé's 40th Anniversary of Their Scientific Careers. *J. Biomedical Science and Engineering* (*JBiSE*), **6**, 435-442. https://doi.org/10.4236/jbise.2013.64054

59. 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

60. Zhou, G.P. and Zhong, W.Z. (2016) Perspectives in Medicinal Chemistry. *Current Topics in Medicinal Chemistry*, **16**, 381-382. https://doi.org/10.2174/1568026616041510141140 30

61. Zhou, X.B., Chen, C., Li, Z.C. and Zou, X.Y. (2007) Using Chou's Amphiphilic Pseudo Amino Acid Composition and Support Vector Machine for Prediction of Enzyme Subfamily Classes. *Journal of Theoretical Biology*, **248**, 546–551. https://doi.org/10.1016/j.jtbi.2007.06.001

62. Nanni, L. and Lumini, A. (2008) Genetic Programming for Creating Chou's Pseudo Amino Acid Based Features for Submitochondria Localization. *Amino Acids*, **34**, 653-660. https://doi.org/10.1007/s00726-007-0018-1

63. 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

64. 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

65. Mohabatkar, H., Mohammad Beigi, M. and Esmaeili, A. (2011) Prediction of GABA(A) Receptor Proteins Using the Concept of Chou's Pseudo Amino Acid Composition and Support Vector Machine. *Journal of Theoretical Biology*, **281**, 18-23. https://doi.org/10.1016/j.jtbi.2011.04.017

66. Mohammad Beigi, 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

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

68. 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

69. 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

70. 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

71. Jia, J., Liu, Z., Xiao, X. and Liu, B. (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 & Dynamics* (*JBSD*), **34**, 1946-1961. https://doi.org/10.1080/07391102.2015.1095116

72. 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 & Modelling*, **76**, 260-273. https://doi.org/10.1016/j.jmgm.2017.07.012

73. Huo, H., Li, T., Wang, S., Lv, Y., Zuo, Y. and Yang, L. (2017) Prediction of Presynaptic and Postsynaptic Neurotoxins by Combining Various Chou's Pseudo Components. *Scientific Reports*, **7**, 5827. https://doi.org/10.1038/s41598-017-06195-y

74. 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

75. 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

76. 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

77. Qiu, W.R. and Xiao, X. (2014) iRSpot-TNCPseAAC: Identify Recombination Spots with Trinucleotide Composition and Pseudo Amino Acid Components. *International Journal of Molecular Science* (*IJMS*), **15**, 1746-1766. https://doi.org/10.3390/ijms15021746

78. 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

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

80. 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

81. Feng, P., Ding, H., Yang, H. and Chen, W. (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

82. 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

83. 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

84. 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

85. Liu, B., Liu, F., 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

86. 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 Science*, **9**, 67-91. https://doi.org/10.4236/ns.2017.94007

87. Cai, Y.D. and Chou, K.C. (2003) Nearest Neighbour Algorithm for Predicting Protein Subcellular Location by Combining Functional Domain Composition and Pseudon Amino Acid Composition. *Biochemical and Biophysical Research Communications* (*BBRC*), **305**, 407-411. https://doi.org/10.1016/S0006-291X(03)00775-7

88. Chou, K.C. and Cai, Y.D. (2003) A New Hybrid Approach to Predict Subcellular Localization of Proteins by Incorporating Gene Ontology. *Biochemical and Biophysical Research Communications* (*BBRC*), **311**, 743-747. https://doi.org/10.1016/j.bbrc.2003.10.062

89. Chou, K.C. and Cai, Y.D. (2004) Prediction of Protein Subcellular Locations by GO-FunD-PseAA Predictor. *Biochemical and Biophysical Research Communications* (*BBRC*), **320**, 1236-1239. https://doi.org/10.1016/j.bbrc.2004.06.073

90. Li, L., Zhang, Y., Zou, L., Li, C., Yu, B., Zheng, X. and Zhou, Y. (2012) An Ensemble Classifier for Eukaryotic Protein Subcellular Location Prediction Using Gene Ontology Categories and Amino Acid Hydrophobicity. *PLoS ONE*, **7**, e31057. https://doi.org/10.1371/journal.pone.0031057

91. Wan, S., Mak, M.W. and Kung, S.Y. (2013) GOASVM: A Subcellular Location Predictor by Incorporating Term-Frequency Gene Ontology into the General Form of Chou's Pseudo Amino Acid Composition. *Journal of Theoretical Biology*, **323**, 40-48. https://doi.org/10.1016/j.jtbi.2013.01.012

92. 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

93. Chou, K.C. and Shen, H.B. (2006) Predicting Eukaryotic Protein Subcellular Location by Fusing Optimized Evidence-Theoretic K-Nearest Neighbor Classifiers. *Journal of Proteome Research*, **5**, 1888-1897. https://doi.org/10.1021/pr060167c

94. Wang, T., Yang, J. and Shen, H.B. (2008) Predicting Membrane Protein Types by the LLDA Algorithm. *Protein & Peptide Letters*, **15**, 915-921. https://doi.org/10.2174/092986608785849308

95. 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

96. 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*. https://doi.org/10.1016/j.ygeno.2017.08.005

97. 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*. https://doi.org/10.1093/bioinformatics/btx476

98. 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. https://doi.org/10.1016/j.gene.2017.07.036

99. 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

100. Cheng, X., Zhao, S.G. and Xiao, X. (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

101. 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

102. 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

103. Chou, K.C. and Elrod, D.W. (2003) Prediction of Enzyme Family Classes. *Journal of Proteome Research*, **2**, 183-190. https://doi.org/10.1021/pr0255710

104. Chou, K.C. and Shen, H.B. (2007) MemType-2L: A Web Server for Predicting Membrane Proteins and Their Types by Incorporating Evolution Information through Pse-PSSM. *Biochemical and Biophysical Research Communications* (*BBRC*), **360**, 339-345. https://doi.org/10.1016/j.bbrc.2007.06.027

105. Ali, F. and Hayat, M. (2015) Classification of Membrane Protein Types Using Voting Feature Interval in Combination with Chou's Pseudo Amino Acid Composition. *Journal of Theoretical Biology*, **384**, 78-83. https://doi.org/10.1016/j.jtbi.2015.07.034

106. Chen, J., Liu, H. and Yang, J. (2007) Prediction of Linear B-Cell Epitopes Using Amino Acid Pair Antigenicity Scale. *Amino Acids*, **33**, 423-428. https://doi.org/10.1007/s00726-006-0485-9

107. Xu, Y., Wen, X., Shao, X.J. and Deng, N.Y. (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 (IJMS)*, **15**, 7594-7610. https://doi.org/10.3390/ijms15057594

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

109. 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

110. Liu, L.M. and Xu, Y. (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

111. Qiu, W.R., Jiang, S.Y., Sun, B.Q., Xiao, X. and Cheng, X. (2017) iRNA-2methyl: Identify RNA 2'-O-Methylation Sites by Incorporating Sequence-Coupled Effects into General PseKNC and Ensemble Classifier. *Medicinal Chemistry*.

112. Xu, Y. and Li, 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

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

114. Xu, Y., Shao, X.J., Wu, L.Y. and Deng, N.Y. (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

115. Jia, J., Liu, Z., Xiao, X. and Liu, B. (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/10.7717/peerj.171

116. Liu, B., Wu, H., Zhang, D. and Wang, X. (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

117. Qiu, W.R., Xiao, X. and Xu, Z.H. (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

118. Xiao, X., Ye, H.X., Liu, Z. and Jia, J.H. (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

119. 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.

120. 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

121. 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

122. Wang, J., Yang, B., Revote, J., Leier, A., Marquez-Lago, T.T., Webb, G. and Song, J. (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