Genomics xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Methods Paper

pLoc_bal-mHum: Predict subcellular localization of human proteins by PseAAC and quasi-balancing training dataset

Kuo-Chen Chou[a,c,*], Xiang Cheng[a,b], Xuan Xiao[a,b,*]

[a] Gordon Life Science Institute, Boston, MA 02478, USA
[b] Computer Science, Jingdezhen Ceramic Institute, Jingdezhen, China
[c] Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu, China

ARTICLE INFO

Keywords:
Multi-label system
Human proteins
Quasi-balance
IHTS treatment
5-step rules
A set of 5 intuitive metrics for multi-label predictors

ABSTRACT

A cell contains numerous protein molecules. One of the fundamental goals in molecular cell biology is to determine their subcellular locations since this information is extremely important to both basic research and drug development. In this paper, we report a novel and very powerful predictor called "pLoc_bal-mHum" for predicting the subcellular localization of human proteins based on their sequence information alone. Cross-validation tests on exactly the same experiment-confirmed dataset have indicated that the new predictor is remarkably superior to the existing state-of-the-art predictor in identifying the subcellular localization of human proteins. To maximize the convenience for the majority of experimental scientists, a user-friendly web-server for the new predictor has been established at http://www.jci-bioinfo.cn/pLoc_bal-mHum/, by which users can easily get their desired results without the need to go through the detailed mathematics.

1. I. introduction

The smallest unit of life is a cell, which contains many protein molecules. Most of the functions critical to the cell's survival are performed by the proteins located in many different organelles, usually termed "subcellular locations". Information of the subcellular locations of proteins can provide useful clues about their functions. Accordingly, knowledge of the subcellular localization of proteins is crucially important for both basic research and drug development [1–4].

With the avalanche of gene products in the post-genomic era, the gap between the newly discovered protein sequences and the knowledge of their subcellular localization is becoming increasingly large. In order to timely use these new proteins for basic research and drug design, it is highly demanded to develop powerful computational tools to annotate their subcellular locations according to their sequence information alone.

Actually, many efforts have been made in this regard that can be can be roughly separated into two parts: one is to deal with the single-label system in which it is assumed that each protein has one and only one subcellular location (the corresponding publications (see, e.g., [5–28] and a long list of references cited in the two review papers [3,29]); the other is to deal with the multi-label system in which each of the constituent proteins may have two or more subcellular locations. Actually, the latter is more close to the real world since, with more experimental data emerging, the localization of proteins in a cell is not a single-label but a multi-label system. Moreover, it is these kinds of multiplex proteins that usually have some unique functions important for really understanding the biological process in a cell and for drug design as well [2,30,31].

From 2007, a series of computational methods for predicting the subcellular localization of proteins from various organisms that have both single and multiple location sites were developed (see, e.g., [32–52]. But all these methods were still examined mainly by the metrics for the single-label system rather than those special for the multi-label system [53]. Therefore, the reported rates there cannot be used to completely reflect their prediction quality. For instance, none of the aforementioned papers reported the "absolute true" rate or "perfectly completely correct" rate, the most important metrics used to measure the power of a multi-label predictor [53].

In 2017–2018, seven powerful multi-label predictors have been established [54–60] for predicting the subcellular localization of multi-label plant, virus, Gram-negative bacterial, Gram-positive bacterial, animal, eukaryotic, and human proteins, respectively. For example, the multi-label predictor "pLoc-mHum" [60] is overwhelmingly superior to its counterparts [33,34,45] in predicting the subcellular localization of human proteins with both single and multiple location sites. The absolute true rate achieved by pLoc-mHum [60] is over 79%, which is > 10% higher than that by the iLoc-Hum predictor [45], and hence it

* Corresponding author at: Gordon Life Science Institute, Boston, MA 02478, USA.
E-mail addresses: kcchou@gordonlifescience.org (K.-C. Chou), xcheng@gordonlifescience.org (X. Cheng), xxiao@gordonlifescience.org (X. Xiao).

https://doi.org/10.1016/j.ygeno.2018.08.007
Received 3 May 2018; Received in revised form 14 August 2018; Accepted 16 August 2018

Table 1
The benchmark dataset $\mathbb{S}$ used in [60] for training and testing the predictor pLoc-mHum[a].

Subset	Subcellular location name	Number of proteins
$\mathbb{S}_1$	Centrosome	77
$\mathbb{S}_2$	Cytoplasm	817
$\mathbb{S}_3$	Cytoskeleton	79
$\mathbb{S}_4$	Endoplasmic Reticulum	229
$\mathbb{S}_5$	Endosome	24
$\mathbb{S}_6$	Extracellular	385
$\mathbb{S}_7$	Golgi Apparatus	161
$\mathbb{S}_8$	Lysosome	77
$\mathbb{S}_9$	Microsome	24
$\mathbb{S}_{10}$	Mitochondrion	364
$\mathbb{S}_{11}$	Nucleus	1021
$\mathbb{S}_{12}$	Peroxisome	47
$\mathbb{S}_{13}$	Plasma Membrane	354
$\mathbb{S}_{14}$	Synapse	22

[a] See Eq. 1 and relevant context as well as the Supporting Information S1 for further explanation.

represents the state-of-the-art computational prediction method in this regard.

However, the pLoc-mHum predictor [60] has the following problem: it was trained by an extremely skewed benchmark dataset. As we can see from column 3 of Table 1, the number of protein samples in "Synapse" is only 22, but that in "Nucleus" is 1021. The later is over 46 times the size of the former. Similar situation also exist for the other subcellular locations such as "Endosome", and "Microsome". As pointed out by a series of recent publications [61–65], most statistical predictors constructed via the machine-learning approach could not avoid biased consequence if the model was trained by the dataset formed by very uneven size subsets. Therefore, we are facing a challenge of how to further improve the state-of-the-art predictor by alleviate this kind of biased consequence.

The present study was devoted to address such a problem. As conducted in pLoc-mHum [60] and a series of recent reports in constructing various statistical predictors (see, e.g., [65–80], we should observe the 5-step rules [81], which request clear descriptions for the following five procedures: benchmark dataset, sample formulation, operation engine or algorithm, cross-validation, and web-server. But here our attentions are focused on the procedures that differ significantly from those in pLoc-mHum [60].

2. Materials and methods

2.1. Benchmark dataset

The benchmark dataset used in this study was taken from [60]. It can be formulated as

$$\mathbb{S} = \mathbb{S}_1 \bigcup \mathbb{S}_2 \bigcup \cdots \bigcup \mathbb{S}_u \bigcup \cdots \bigcup \mathbb{S}_{13} \bigcup \mathbb{S}_{14} \tag{1}$$

where $\mathbb{S}_1$ only contains the protein samples from the "Centrosome" location (cf. Table 1), $\mathbb{S}_2$ only contains those from the "Cytoplasm" location, and so forth; $\bigcup$ denotes the symbol for "union" in the set theory [3] . The detailed sequences of these proteins and their accession numbers (or ID codes) are given in Supporting Information S1 that is also available at http://www.jci-bioinfo.cn/pLoc_bal-mHum/Supp1.pdf where none of proteins included has ≥25% sequence identity to any other in the same subset (subcellular location).

2.2. Proteins sample formulation

The 2nd step of the 5-step rules [81] is how to formulate the biological sequence samples with an effective mathematical expression that can effectively reflect their essential correlation with the target concerned. For a protein sequence **P**, the most straightforward manner to express it is the sequential model as defined by

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

where L is the protein's length or the number of its constituent amino acid residues, $\mathrm{R}_1$is the 1st residue, R_2 the 2nd residue, R_3 the 3rd residue, and so forth. Because all the existing machine-learning algorithms can only handle vectors [4], one has to transfer a protein sample from the sequential mode (Eq. 2) to a discrete mode or vector. But a vector defined in a discrete mode might totally lose the sequence-order information. In order to approximately keep this kind of information, the PseAAC (Pseudo Amino Acid Composition) was proposed [12,82]. Ever since then, the concept of PseAAC has swiftly penetrated into nearly all the areas of computational proteomics with the purpose to formulate various different sequence patterns that are critical to the targets investigated (see, e.g., [13,23,27,51,83–124] and a long list of references cited in a recent review papers [125]. Encouraged by the successes of using PseAAC to deal with protein/peptide sequences, its concept was extended to cover DNA/RNA sequences via PseKNC (Pseudo K-tuple Nucleotide Composition) [126] and has been proved very successful as well [65–68,70,74,76–78,127–135]. Particularly, recently a very powerful web-server called "Pse-in-One" [136] and its updated version "Pse-in-One2.0" [137] have been established that can be used to generate any desired feature vectors for protein/peptide and DNA/RNA sequences according to the need of users' studies or their own definitions. According to [81], the general PseAAC of a protein sequence can be formulated as

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

where **T** is a transpose operator, while the integer Ω is a parameter and its value as well as the components Ψ_u $(u = 1, 2, \cdots, \Omega)$ will depend on how to extract the desired information from the amino acid sequence of **P**, as described in [60]. Thus, by means of the same procedures as elaborated in the Section 2.2 of [60], each of the protein samples in the benchmark dataset can be uniquely defined as a 14-D numerical vector as shown in columns 16–29 of Supporting Information S2, which can also be directly downloaded at http://www.jci-bioinfo.cn/pLoc_bal-mHum/Supp2.pdf.

2.3. Quasi-balancing training dataset

As indicated in Table 1, the benchmark dataset used to train the pLoc-mHum predictor [60] was extremely imbalanced and hence the model trained by such a skewed dataset cannot avoid the biased consequence. Actually, there exist some established methods to deal with this problem, the "Monte Carlo samples expanding" approach [138,139], "seed-propagation" approach [140], "LogiBoost" [141], "SMOTE" (synthetic minority over-sampling technique) approach [61,62,142], and "Bootstrap" approach [143]. Here we adopted the IHTS (Inserting Hypothetical Training Samples) treatment [63–65] to add some theoretical or hypothetical samples into the smaller subsets, so as to increase their sizes and construct a quasi-balanced dataset $\widetilde{\mathbb{S}}$ as formulated below.

$$\widetilde{\mathbb{S}} = \widetilde{\mathbb{S}_1} \bigcup \widetilde{\mathbb{S}_2} \bigcup \cdots \widetilde{\mathbb{S}_u} \bigcup \cdots \widetilde{\mathbb{S}_{13}} \bigcup \widetilde{\mathbb{S}_{14}} \tag{4}$$

where $\widetilde{\mathbb{S}_u}$ $(u = 1, 2, \cdots, 14)$ is derived from $\mathbb{S}_u$ of Eq. 1 via the IHTS approach [63,64] and it contains $\widetilde{n_u}$ samples with each corresponding to a 14-D vector. Its value is given by

$$\widetilde{n_u} = \begin{cases} 2n_u & \text{if } n_{\max} > 2n_u \\ n_{\max} & \text{if } n_{\max} \le 2n_{\mathrm{u}} \end{cases} \qquad (u = 1, 2, 3, \cdots, 14) \tag{5}$$

where n_u $(u = 1, 2, 3, \cdots, 14)$ is the number of protein samples in the u-th subset of the benchmark dataset $\mathbb{S}$ (see Eq.1), and

$$n_{\max} = \mathbf{Max}_{i=1}^{14}\{n_i\} \tag{6}$$

where the operator **Max** means taking the maximum one among the

Table 2
The original dataset $\mathbb{S}$ used to train and test the pLoc-mHum predictor [60] and the quasi-balanced dataset $\widetilde{\mathbb{S}}$ used to train the proposed predictor.

Location	Before quasi-balancing		After quasi-balancing	
	Subset	Size	Subset	Size
Centrosome	$\mathbb{S}_1$	77	$\widetilde{\mathbb{S}}_1$	154
Cytoplasm	$\mathbb{S}_2$	817	$\widetilde{\mathbb{S}}_2$	1021
Cytoskeleton	$\mathbb{S}_3$	79	$\widetilde{\mathbb{S}}_3$	158
Endoplasmic reticulum	$\mathbb{S}_4$	229	$\widetilde{\mathbb{S}}_4$	458
Endosome	$\mathbb{S}_5$	24	$\widetilde{\mathbb{S}}_5$	48
Extracellular	$\mathbb{S}_6$	385	$\widetilde{\mathbb{S}}_6$	770
Golgi apparatus	$\mathbb{S}_7$	161	$\widetilde{\mathbb{S}}_7$	322
Lysosome	$\mathbb{S}_8$	77	$\widetilde{\mathbb{S}}_8$	154
Microsome	$\mathbb{S}_9$	24	$\widetilde{\mathbb{S}}_9$	48
Mitochondrion	$\mathbb{S}_{10}$	364	$\widetilde{\mathbb{S}}_{10}$	728
Nucleus	$\mathbb{S}_{11}$	1021	$\widetilde{\mathbb{S}}_{11}$	1021
Peroxisome	$\mathbb{S}_{12}$	47	$\widetilde{\mathbb{S}}_{12}$	94
Plasma membrane	$\mathbb{S}_{13}$	354	$\widetilde{\mathbb{S}}_{13}$	708
Synapse	$\mathbb{S}_{14}$	22	$\widetilde{\mathbb{S}}_{14}$	44

numbers in the followed-up brackets. For the current benchmark dataset, the largest subset is of "Nucleus" (cf. Table 1) and hence $n_{\max} = 1{,}021$.

Thus, according to Eqs.5–6, $\widetilde{\mathbb{S}}_1$ contains 28 vectors of which 14 for the real proteins and the remaining 14 for the pseudo or theoretical proteins in "Acrosome"; $\widetilde{\mathbb{S}}_2$ contains 1394 vectors of which 697 for the real proteins while the remaining 697 for the pseudo or theoretical proteins in "Cell membrane"; $\widetilde{\mathbb{S}}_3$ contains 98 vectors of which only 49 for the real proteins while the remaining 97 for the pseudo proteins in "Cell wall"; and so forth (Table 2).

All those 14-D vectors for the theoretical or hypothetical proteins in Table 2 are also included in Supporting Information S2 but the first character of their codes is "X" and their nature is marked as "Theo", implying that they are not corresponding to real proteins but to theoretical or pseudo proteins.

Note that the theoretical samples added into the small subsets via the aforementioned IHTS treatment can only be expressed in the form of feature vectors as shown in Supporting Information S2 but not the form of amino acid sequences as given in Supporting Information S1. Nevertheless, doing so is perfectly fine; this is because when using "machine-learning" to train a prediction model, what we directly need are the feature-vectors of protein samples not their amino acid sequences [4]. Keep this in mind that is the key for quasi-balancing the training dataset, and that will be further justified later.

2.4. Operation algorithm

Exactly the same as in developing the pLoc-mHum predictor [60], the ML-GKR (multi-label Gaussian kernel regression) classifier has been adopted. Its details have been clearly elaborated in Section 2.3 of [60], and hence there is no need to repeat here. Note that the ML-GKR classifier contains an uncertain parameter θ. Its value can be determined via the optimization procedure as will be mentioned later.

The new predictor developed via the above procedures is called pLoc_bal-mHum, where "pLoc_bal" stands for "predict subcellular localization by quasi-balancing training dataset", and "mHum" for "multi-label human proteins". Shown in Fig. 1 is a flowchart to illustrate the process of how the pLoc_bal-mHum predictor is working.

3. Results and discussion

According to the 5-step rule [81], one of the important procedures in developing a new predictor is how to properly evaluate its

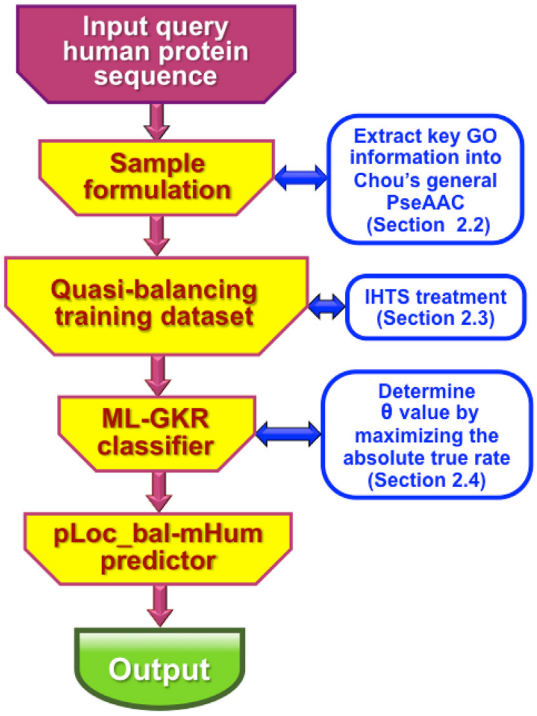

Fig. 1. A flowchart to show the process of how the pLoc_bal-mHum predictor works. For the detailed procedures of how to "extract key GO information", see [60].

anticipated accuracy. To deal with that, two issues need to be considered. (**1**) What metrics should be used to quantitatively reflect the predictor's quality? (**2**) What test method should be applied to score the metrics?

3.1. A Set of five metrics for multi-label systems

Different from the metrics used to measure the prediction quality of single-label systems, the metrics for the multi-label systems are much more complicated. To make them more intuitive and easier to understand for most experimental scientists, here we use the following intuitive Chou's five metrics [53] that have recently been widely used for studying various multi-label systems (see, e.g., [47,57,60,69,144,145]):

$$\begin{cases} \text{Aiming}\uparrow = \dfrac{1}{\mathrm{N}^{\mathrm{q}}}\sum_{k=1}^{\mathrm{N}^{\mathrm{q}}}\left(\dfrac{\|\mathbb{L}_k \bigcap \mathbb{L}_k^*\|}{\|\mathbb{L}_k^*\|}\right), & [0,1] \\ \text{Coverage}\uparrow = \dfrac{1}{\mathrm{N}^{\mathrm{q}}}\sum_{k=1}^{\mathrm{N}^{\mathrm{q}}}\left(\dfrac{\|\mathbb{L}_k \bigcap \mathbb{L}_k^*\|}{\|\mathbb{L}_k\|}\right), & [0,1] \\ \text{Accuracy}\uparrow = \dfrac{1}{\mathrm{N}^{\mathrm{q}}}\sum_{k=1}^{\mathrm{N}^{\mathrm{q}}}\left(\dfrac{\|\mathbb{L}_k \bigcap \mathbb{L}_k^*\|}{\|\mathbb{L}_k \bigcup \mathbb{L}_k^*\|}\right), & [0,1] \\ \text{Absolute true}\uparrow = \dfrac{1}{\mathrm{N}^{\mathrm{q}}}\sum_{k=1}^{\mathrm{N}^{\mathrm{q}})}\Delta(\mathbb{L}_k, \mathbb{L}_k^*), & [0,1] \\ \text{Absolute false}\downarrow = \dfrac{1}{\mathrm{N}^{\mathrm{q}}}\sum_{k=1}^{\mathrm{N}^{\mathrm{q}}}\left(\dfrac{\|\mathbb{L}_k \bigcup \mathbb{L}_k^*\| - \|\mathbb{L}_k \bigcap \mathbb{L}_k^*\|}{M}\right), & [1,0] \end{cases} \tag{7}$$

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 $L_{\mathrm{cell}} = 14$), $\|\|$ means the operator acting on the set

therein to count the number of its elements, $\bigcup$ means the symbol for the "union" in the set theory, $\bigcap$ 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

$$\Delta(\mathbb{L}_k, \mathbb{L}_k^*) = \begin{cases} 1, & \text{if all the labels in } \mathbb{L}_k^* \text{ are identical to those in } \mathbb{L}_k \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

In Eq. 7, the first four metrics with an upper arrow ↑ are called positive metrics, meaning that the larger the rate is the better the prediction quality will be; the 5th metrics with a down arrow ↓ is called negative metrics, implying just the opposite meaning.

From Eq. 7 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. Target 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 [146]. 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 [81]. Accordingly, the jackknife test has been widely recognized and increasingly used by investigators to examine the quality of various predictors (see, e.g., [98,100,110,147–152]). During the jackknifing process, each of the samples in the benchmark dataset will be in turn singled out and tested by the predictor trained by the remaining samples.

However, when carrying out the jackknife test on the optimized dataset described in Section 2.3, some special consideration is needed. This is because the quasi-balanced dataset contains many theoretical proteins. Validation should be performed strictly based on the experimental data only. To realize that, a special jackknife test, the so-called "target-jackknife" test [61–63], is needed.

During the target-jackknife process, only the experiment-confirmed samples listed in column 3 of Table 2, namely those marked with "Real" in Supporting Information S2, will be in turn singled out as a target (or test sample) for the cross validation, but not those marked with "Theo" generated for training the model only.

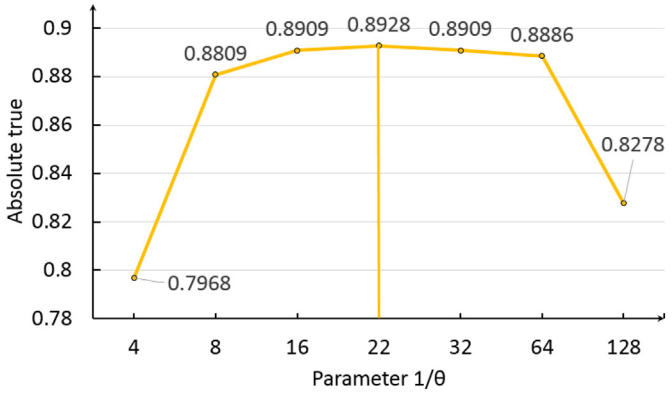

Fig. 2. A plot to show the process of finding the optimal θ value that is 1/22. See the text in Section 2.4 for further explanation.

Table 3
Comparison with the state-of-the-art method in predicting human protein subcellular localization[a].

Predictor	Aiming (↑)[a]	Coverage (↑)[a]	Accuracy (↑)[a]	Absolute true (↑)[a]	Absolute false (↓)[a]
pLoc-mHum[b]	90.57%	82.75%	84.39%	79.14%	1.2%
pLoc_bal-Hum[c]	92.06%	94. 54%	92.29%	89. 28%	0.48%

[a] See Eq. 5 for the definition of the metrics.
[b] See [60], where the reported metrics rates were obtained by the jackknife test on the benchmark dataset of Supporting Information S1 that contains experiment-confirmed proteins only.
[c] The proposed predictor; to assure that the test was performed on exactly the same experimental data as reported in [60] for pLoc-mHum, the target jackknife test was adopted. See Section 3.2 for further explanation.

For the concrete steps of how to perform the target-jackknife test, see the Section of "Jackknife and target–jackknife cross-validation" in [61], where a step-by-step guide had been clearly described and hence there in no need to repeat here. But it is instructive to point out that in order for reducing computational time the target-jackknife test carried out in [61] was actually a kind of quasi-jackknife.

3.3. Parameter determination

Since the ML-GKR classifier (cf. Section 2.4) contains an uncertain parameter θ, the predicted results obtained by pLoc_bal-mHum 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 Eq.7) by the target-jackknife validation on the optimized benchmark dataset of Eq.4 (Supporting Information S2). As shown in Fig. 2, when θ = 1/22, 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 3 are the rates achieved by the current pLoc_bal-mHum predictor via the target-jackknife test on the same experiment-confirmed dataset as used in [60]}. For facilitating comparison, listed there are also the corresponding results obtained by the pLoc-mHum [60], the existing most powerful predictor for identifying the subcellular localization of human proteins with both single and multiple

Table 4
Performance of pLoc_bal-mHum for each of the 14 subcellular locations.

i	Location[a]	Sn(i)[b]	Sp(i)[b]	Acc(i)[b]	MCC(i)[b]
1	Centrosome	0.9870	0.9944	0.9942	0.8954
2	Cytoplasm	0.8690	0.9672	0.9414	0.8472
3	Cytoskeleton	0.9873	0.9983	0.9981	0.9623
4	Endoplasmic reticulum	1.0000	0.9948	0.9952	0.9662
5	Endosome	1.0000	0.9994	0.9994	0.9605
6	Extracellular	0.9974	0.9923	0.9929	0.9685
7	Golgi apparatus	1.0000	0.9952	0.9955	0.9569
8	Lysosome	1.0000	0.9990	0.9990	0.9806
9	Microsomeb	1.0000	0.9994	0.9994	0.9605
10	Mitochondrion	1.0000	0.9905	0.9916	0.9615
11	Nucleus	0.9109	0.9559	0.9411	0.8665
12	Peroxisome	1.0000	0.9993	0.9994	0.9791
13	Plasma membrane	1.0000	0.9909	0.9920	0.9621
14	Synapse	1.0000	0.9997	0.9997	0.9779

[a] See Table 1 and the relevant context for further explanation.
[b] See Eq. 9 for the metrics definition.

location sites. As shown in Table 3, the newly proposed predictor pLoc_bal-mHum is remarkably superior to the existing state-of-the-art predictor pLoc-mHum in all the five metrics. Particularly, it can be seen from the table that the absolute true rate achieved by the new predictor is over 89%, which is far beyond the reach of any other existing methods [33,34,40,41,45,60]. This is because it is extremely difficult to enhance the absolute true rate of a prediction method for a multi-label system as clearly elucidated in [60]. Actually, to avoid embarrassment, many investigators even chose not to mention the metrics of absolute true rate in dealing with multi-label systems (see, e.g., [48,51,153–157]).

Moreover, to in-depth examine the prediction quality of the new predictor for the proteins in each of the subcellular locations concerned (cf. Table 1), we used a set of four intuitive metrics that were derived in [158,159] based on the Chou's symbols introduced for studying protein signal peptides [160,161] and that have ever since been widely concurred or justified (see, e.g., [61,62,65,70,74,76,77,143,158,162–173]. For the current study, the set of metrics can be formulated as [57,58,60,65,174]:

$$\begin{cases} \mathrm{Sn}(i) = 1 - \dfrac{N_-^+(i)}{N^+(i)} & 0 \le \mathrm{Sn(i)} \le 1 \\ \mathrm{Sp}(i) = 1 - \dfrac{N_+^-(i)}{N^-(i)} & 0 \le \mathrm{Sp(i)} \le 1 \\ \mathrm{Acc}(i) = 1 - \dfrac{N_-^+(i) + N_+^-(i)}{N^+(i) + N^-(i)} & 0 \le \mathrm{Acc(i)} \le 1 \\ \mathrm{MCC}(i) = \dfrac{1 - \left(\frac{N_-^+(i)}{N^+(i)} + \frac{N_+^-(i)}{N^-(i)}\right)}{\sqrt{\left(1 + \frac{N_+^-(i) - N_-^+(i)}{N^+(i)}\right)\left(1 + \frac{N_-^+(i) - N_+^-(i)}{N^-(i)}\right)}} & -1 \le \mathrm{MCC} \le 1 \\ (\, i = 1, 2, \cdots, 14 \,) \end{cases} \tag{9}$$

where Sn, Sp, Acc, and MCC represent the sensitivity, specificity, accuracy, and Mathew's correlation coefficient, respectively (Chen et al., 2007), and i denotes the i-th subcellular location (or subset) in the benchmark dataset. $N^+(i)$ is the total number of the samples investigated in the i-th subset, whereas $N_-^+(i)$ is the number of the samples in $N^+(i)$ that are incorrectly predicted to be of other locations; $N^-(i)$ is the total number of samples in any 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.

Listed in Table 4 are the results achieved by pLoc_bal-mHum for the human proteins in each of 14 subcellular locations. As we can see from the table, nearly all the success rates achieved by the new predictor for

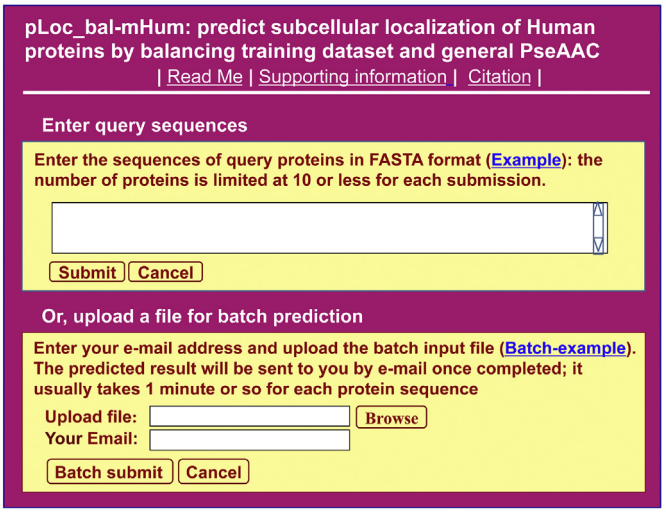

Fig. 3. A semi screenshot of the top page of pLoc_bal-mHum. It is accessible at the web-site of http://www.jci-bioinfo.cn/pLoc_bal-mHum/. See Step 1 of Section 3.5 for more explanation.

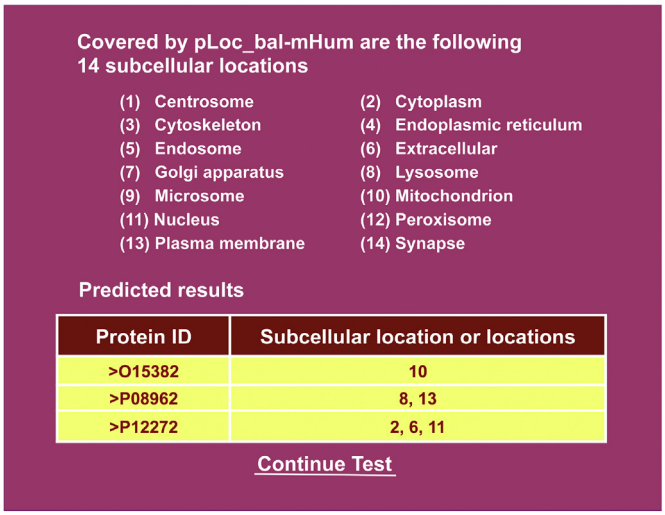

Fig. 4. A semi screenshot for the webpage obtained by following Step 3 of Section 3.5. It is available at http://www.jci-bioinfo.cn/pLoc_bal-mHum/iLocAnimalResultBySeqs. See Step 3 of Section 3.5 for more explanation.

the human proteins in each of the 14 subcellular locations are within the range of 95–100%, which is once again far beyond the reach of any of its counterparts.

The following question might be raised. There are many prediction methods for protein subcellular localization, why was the comparison made only with the pLoc-mHum [60] predictor? The reasons are as follows. (**1**) Although there are many methods for predicting protein subcellular localization, most of them can be used to deal with single-label protein system only because it was assumed in their developing processes that each of the proteins concerned in a cell had one and only one subcellular location site. (**2**) Despite there are some predictors that can be used to deal with the multi-label proteins, most of them did not report their "absolute true" or "perfect correct" rate, the most important metrics to indicate the power of a multi-label predictor [53]. This is because their absolute true rates were too low to be worthy of reporting. (**3**) So far there are only two existing predictors, one is iLoc-Hum [45] and one is pLoc-mHum [60], which did report the absolute true rates in predicting the subcellular localization for multi-label human proteins. The absolute true rate reported by the former was 68%, while that by the latter was 79%, meaning that pLoc-mHum [60] represents the current state-of-the-art predictor in this regard. Accordingly, it is more than enough to show the power of the proposed multi-label predictor by just comparing it with pLoc-mHum [60].

3.5. Web server and user guide

As pointed out in [175], user-friendly and publicly accessible web-servers represent the future direction for developing practically more useful predictors. Actually, user-friendly web-servers as given in a series of recent publications [54–56,59,65,67,68,73,74,76,78,80,159,167,176–189] have increasing impacts on medical science [4], driving medicinal chemistry into an unprecedented revolution [125]. In view of this, the web-server of the current pLoc_bal-mHum predictor has also been established. Moreover, to maximize users' convenience, a step-by-step guide is given below.

Step 1. Click the link at http://www.jci-bioinfo.cn/pLoc_bal-mHum/, the top page of the pLoc_bal-mHum web-server will appear on your computer screen, as shown in Fig. 3. Click on the Read Me button to see a brief introduction about the predictor.

Step 2. Either type or copy/paste the sequences of query plant proteins into the input box at the center of Fig. 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 four protein sequences in the Example window as the input, after 10 s or so, you will see a new screen (Fig. 4) occurring. On its upper part are listed the names of the subcellular locations numbered from (1) to (14) covered by the current predictor. On its lower part are the predicted results: the query protein "O15382" of example-1 corresponds to "10," meaning it belonging to "Mitochondrion" only; the query protein "P08962" of example-2 corresponds to "8, 13" meaning it belonging to "Lysosome" and "Plasma membrane"; the query protein "P12272" of example-3 corresponds to "2, 6, 11", meaning it belonging to "Cytoplasm", "Extracellular", and "Nucleus". All these results are perfectly consistent with experimental observations.

Step 4. As shown on the lower panel of Fig. 4, 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_balmHum.

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

3.6. Some remarks

It is instructive to point out that, of the 20 native amino acids, some are quite similar to each other in physicochemical properties [190], and hence can be categorized into a same group since they play similarly structural or functional roles in protein [191,192]. Accordingly, the protein feature vectors can be reduced from the classical 20-D space to a space with fewer dimensions as done in a series of recent publications [174,190,192–194]. This is an interesting idea, which may lead to a new approach or strategy to improve the quality of predicting protein subcellular localization. We shall make our efforts in this regard in our future studies.

Acknowledgments

The authors wish to thank the editor as well as the four anonymous reviewers for their constructive comments, which were very useful for strengthening the presentation of this study. 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).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ygeno.2018.08.007.

References

[1] J.S. Ehrlich, M.D. Hansen, W.J. Nelson, Spatio-temporal regulation of Rac1 localization and lamellipodia dynamics during epithelial cell-cell adhesion, Dev. Cell 3 (2002) 259–270.
[2] E. Glory, R.F. Murphy, Automated subcellular location determination and high-throughput microscopy, Dev. Cell 12 (2007) 7–16.
[3] K.C. Chou, H.B. Shen, Recent progresses in protein subcellular location prediction, Anal. Biochem. 370 (2007) 1–16.
[4] K.C. Chou, Impacts of bioinformatics to medicinal chemistry, Med. Chem. 11 (2015) 218–234.
[5] J. Cedano, P. Aloy, J.A. Perez-Pons, E. Querol, Relation between amino acid composition and cellular location of proteins, J. Mol. Biol. 266 (1997) 594–600.
[6] K.C. Chou, D.W. Elrod, Using discriminant function for prediction of subcellular location of prokaryotic proteins, Biochem Biophys Res Commun (BBRC) 252 (1998) 63–68.
[7] A. Reinhardt, T. Hubbard, Using neural networks for prediction of the subcellular location of proteins, Nucleic Acids Res. 26 (1998) 2230–2236.
[8] K.C. Chou, D.W. Elrod, Protein subcellular location prediction, Protein Eng. 12 (1999) 107–118.
[9] D.W. Elrod, Prediction of membrane protein types and subcellular locations, Proteins Struct. Funct. Genet. 34 (1999) 137–153.
[10] O. Emanuelsson, H. Nielsen, S. Brunak, G. von Heijne, Predicting subcellular localization of proteins based on their N-terminal amino acid sequence, J. Mol. Biol. 300 (2000) 1005–1016.
[11] K.C. Chou, Prediction of protein subcellular locations by incorporating quasi-sequence-order effect, Biochem Biophys Res Comm (BBRC) 278 (2000) 477–483.
[12] K.C. Chou, Prediction of protein cellular attributes using pseudo amino acid composition, PROTEINS: Structure, Function, and Genetics, Vol.44 2001, pp. 246–255 Erratum: ibid., 2001.
[13] Y.D. Cai, X.J. Liu, X.B. Xu, K.C. Chou, Support vector machines for prediction of protein subcellular location by incorporating quasi-sequence-order effect, J. Cell. Biochem. 84 (2002) 343–348.
[14] K.C. Chou, Y.D. Cai, Using functional domain composition and support vector machines for prediction of protein subcellular location, J. Biol. Chem. 277 (2002) 45765–45769.
[15] K.J. Park, M. Kanehisa, Prediction of protein subcellular locations by support vector machines using compositions of amino acid and amino acid pairs, Bioinformatics 19 (2003) 1656–1663.
[16] J.L. Gardy, C. Spencer, K. Wang, M. Ester, G.E. Tusnady, I. Simon, S. Hua, K. Defays, C. Lambert, K. Nakai, F.S. Brinkman, PSORT-B: Improving protein subcellular localization prediction for Gram-negative bacteria, Nucleic Acids Research 31 (2003) 3613–3617.
[17] K.C. Chou, Y.D. Cai, A new hybrid approach to predict subcellular localization of proteins by incorporating gene ontology, Biochemical and Biophysical Research Communications (BBRC) 311 (2003) 743–747.
[18] S. Matsuda, J.P. Vert, H. Saigo, N. Ueda, H. Toh, T. Akutsu, A novel representation of protein sequences for prediction of subcellular location using support vector machines, Protein Sci. 14 (2005) 2804–2813.
[19] A. Pierleoni, P.L. Martelli, P. Fariselli, R. Casadio, BaCelLo: a balanced subcellular localization predictor, Bioinformatics 22 (2006) e408–e416.
[20] K.C. Chou, H.B. Shen, Predicting protein subcellular location by fusing multiple classifiers, J. Cell. Biochem. 99 (2006) 517–527.
[21] P. Horton, K.J. Park, T. Obayashi, N. Fujita, H. Harada, C.J. Adams-Collier, K. Nakai, WoLF PSORT: protein localization predictor, Nucleic Acids Res. 35 (2007) W585–W587.
[22] H.B. Shen, K.C. Chou, Gpos-PLoc: an ensemble classifier for predicting subcellular localization of Gram-positive bacterial proteins, Protein Eng. Des. Sel. 20 (2007) 39–46.
[23] Y.S. Ding, T.L. Zhang, Using Chou's pseudo amino acid composition to predict subcellular localization of apoptosis proteins: an approach with immune genetic algorithm-based ensemble classifier, Pattern Recogn. Lett. 29 (2008) 1887–1892.
[24] J. Lin, Y. Wang, Using a novel AdaBoost algorithm and Chou's pseudo amino acid composition for predicting protein subcellular localization, Protein Pept. Lett. 18 (2011) 1219–1225.
[25] L. Hu, T. Huang, X. Shi, W.C. Lu, Y.D. Cai, Predicting functions of proteins in mouse based on weighted protein-protein interaction network and protein hybrid properties, PLoS ONE 6 (2011) e14556.
[26] G.L. Fan, Q.Z. Li, Predict mycobacterial proteins subcellular locations by incorporating pseudo-average chemical shift into the general form of Chou's pseudo amino acid composition, J. Theor. Biol. 304 (2012) 88–95.
[27] A. Dehzangi, R. Heffernan, A. Sharma, J. Lyons, K. Paliwal, A. Sattar, Gram-positive and Gram-negative protein subcellular localization by incorporating evolutionary-based descriptors into Chou's general PseAAC, J. Theor. Biol. 364 (2015) 284–294.
[28] R. Sharma, A. Dehzangi, J. Lyons, K. Paliwal, T. Tsunoda, A. Sharma, Predict Gram-positive and Gram-negative Subcellular Localization via Incorporating Evolutionary Information and Physicochemical Features into Chou's General PseAAC, IEEE Trans Nanobioscience 14 (2015) 915–926.
[29] K. Nakai, Protein sorting signals and prediction of subcellular localization, Adv. Protein Chem. 54 (2000) 277–344.
[30] L. Liu, Y. Ma, R.L. Wang, W.R. Xu, S.Q. Wang, Find novel dual-agonist drugs for treating type 2 diabetes by means of cheminformatics. Drug Design, Development and Therapy 7 (2013) 279–287.
[31] Y. Ma, S.Q. Wang, W.R. Xu, R.L. Wang, Design novel dual agonists for treating type-2 diabetes by targeting peroxisome proliferator-activated receptors with core hopping approach, PLoS One 7 (2012) e38546.
[32] K.C. Chou, H.B. Shen, Euk-mPLoc: a fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites, J. Proteome Res. 6 (2007) 1728–1734.
[33] H.B. Shen, K.C. Chou, 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 (2007) 1006–1011.
[34] H.B. Shen, K.C. Chou, A top-down approach to enhance the power of predicting

human protein subcellular localization: Hum-mPLoc 2.0, Analytical Biochemistry 394 (2009) 269–274.
[35] H.B. Shen, K.C. Chou, Gpos-mPLoc: a top-down approach to improve the quality of predicting subcellular localization of Gram-positive bacterial proteins, Protein Pept. Lett. 16 (2009) 1478–1484.
[36] K.C. Chou, H.B. Shen, A new method for predicting the subcellular localization of eukaryotic proteins with both single and multiple sites: Euk-mPLoc 2.0, PLoS ONE 5 (2010) e9931.
[37] K.C. Chou, H.B. Shen, Plant-mPLoc: a Top-down strategy to Augment the Power for predicting Plant Protein Subcellular Localization, PLoS One 5 (2010) e11335.
[38] H.B. Shen, K.C. Chou, Gneg-mPLoc: a top-down strategy to enhance the quality of predicting subcellular localization of Gram-negative bacterial proteins, J. Theor. Biol. 264 (2010) 326–333.
[39] H.B. Shen, K.C. Chou, Virus-mPLoc: a Fusion Classifier for Viral Protein Subcellular Location Prediction by Incorporating Multiple Sites, J Biomol Struct Dyn (JBSD) 28 (2010) 175–186.
[40] K.C. Chou, H.B. Shen, Cell-PLoc: a package of Web servers for predicting subcellular localization of proteins in various organisms, Nat. Protoc. 3 (2008) 153–162.
[41] K.C. Chou, H.B. Shen, Cell-PLoc 2.0: An improved package of web-servers for predicting subcellular localization of proteins in various organisms, Natural Science 2 (2010) 1090–1103.
[42] K.C. Chou, Z.C. Wu, X. Xiao, iLoc-Euk: a Multi-Label Classifier for predicting the Subcellular Localization of Singleplex and Multiplex Eukaryotic Proteins, PLoS One 6 (2011) e18258.
[43] Z.C. Wu, X. Xiao, K.C. Chou, iLoc-plant: a multi-label classifier for predicting the subcellular localization of plant proteins with both single and multiple sites, Mol. BioSyst. 7 (2011) 3287–3297.
[44] X. Xiao, Z.C. Wu, K.C. Chou, iLoc-virus: a multi-label learning classifier for identifying the subcellular localization of virus proteins with both single and multiple sites, J. Theor. Biol. 284 (2011) 42–51.
[45] K.C. Chou, Z.C. Wu, X. Xiao, iLoc-hum: using accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites, Mol. BioSyst. 8 (2012) 629–641.
[46] Z.C. Wu, X. Xiao, iLoc-Gpos: a Multi-Layer Classifier for predicting the Subcellular Localization of Singleplex and Multiplex Gram-positive Bacterial Proteins, Protein Pept. Lett. 19 (2012) 4–14.
[47] W.Z. Lin, J.A. Fang, X. Xiao, iLoc-Animal: A multi-label learning classifier for predicting subcellular localization of animal proteins, Molecular BioSystems 9 (2013) 634–644.
[48] C. Huang, J. Yuan, Using radial basis function on the general form of Chou's pseudo amino acid composition and PSSM to predict subcellular locations of proteins with both single and multiple sites, Biosystems 113 (2013) 50–57.
[49] C. Huang, J.Q. Yuan, A multilabel model based on Chou's pseudo amino acid composition for identifying membrane proteins with both single and multiple functional types, J. Membr. Biol. 246 (2013) 327–334.
[50] C. Huang, J.Q. Yuan, Predicting protein subchloroplast locations with both single and multiple sites via three different modes of Chou's pseudo amino acid compositions, J. Theor. Biol. 335 (2013) 205–212.
[51] E. Pacharawongsakda, T. Theeramunkong, 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 (2013) 311–320.
[52] X. Wang, G.Z. Li, W.C. Lu, 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 Pept. Lett. 20 (2013) 309–317.
[53] K.C. Chou, Some Remarks on predicting Multi-Label Attributes in Molecular Biosystems, Molecular Biosystems 9 (2013) 1092–1100.
[54] X. Cheng, X. Xiao, pLoc-mPlant: predict subcellular localization of multi-location plant proteins via incorporating the optimal GO information into general PseAAC, Mol. BioSyst. 13 (2017) 1722–1727.
[55] X. Cheng, X. Xiao, pLoc-mVirus: predict subcellular localization of multi-location virus proteins via incorporating the optimal GO information into general PseAAC, Gene 628 (2017) 315–321 Erratum: ibid., 2018, Vol.644, 156–156.
[56] X. Cheng, X. Xiao, pLoc-mGneg: Predict subcellular localization of Gram-negative bacterial proteins by deep gene ontology learning via general PseAAC, Genomics 110 (2018) 231–239.
[57] X. Xiao, X. Cheng, S. Su, Q. Nao, pLoc-mGpos: Incorporate key gene ontology information into general PseAAC for predicting subcellular localization of Gram-positive bacterial proteins, Nat. Sci. 9 (2017) 331–349.
[58] X. Cheng, S.G. Zhao, W.Z. Lin, X. Xiao, pLoc-mAnimal: predict subcellular localization of animal proteins with both single and multiple sites, Bioinformatics 33 (2017) 3524–3531.
[59] X. Cheng, X. Xiao, K.C. Chou, pLoc-mEuk: Predict subcellular localization of multi-label eukaryotic proteins by extracting the key GO information into general PseAAC, Genomics 110 (2018) 50–58.
[60] X. Cheng, X. Xiao, K.C. Chou, pLoc-mHum: predict subcellular localization of multi-location human proteins via general PseAAC to winnow out the crucial GO information, Bioinformatics 34 (2018) 1448–1456.
[61] Z. Liu, X. Xiao, W.R. Qiu, iDNA-methyl: Identifying DNA methylation sites via pseudo trinucleotide composition, Anal. Biochem. 474 (2015) 69–77.
[62] X. Xiao, J.L. Min, W.Z. Lin, Z. Liu, X. Cheng, iDrug-target: predicting the interactions between drug compounds and target proteins in cellular networking via the benchmark dataset optimization approach, J Biomol Struct Dyn (JBSD) 33 (2015) 2221–2233.
[63] J. Jia, Z. Liu, X. Xiao, B. Liu, iSuc-PseOpt: Identifying lysine succinylation sites in proteins by incorporating sequence-coupling effects into pseudo components and optimizing imbalanced training dataset, Anal. Biochem. 497 (2016) 48–56.
[64] J. Jia, Z. Liu, X. Xiao, B. Liu, iPPBS-opt: a Sequence-based Ensemble Classifier for Identifying Protein-Protein Binding Sites by Optimizing Imbalanced Training Datasets, Molecules 21 (2016) E95.
[65] B. Liu, F. Yang, D.S. Huang, iPromoter-2L: a two-layer predictor for identifying promoters and their types by multi-window-based PseKNC, Bioinformatics 34 (2018) 33–40.
[66] W. Chen, H. Tang, J. Ye, H. Lin, iRNA-PseU: Identifying RNA pseudouridine sites Molecular Therapy, Nucleic Acids 5 (2016) e332.
[67] P. Feng, H. Ding, H. Yang, W. Chen, H. Lin, iRNA-PseColl: Identifying the occurrence sites of different RNA modifications by incorporating collective effects of nucleotides into PseKNC, Molecular Therapy - Nucleic Acids 7 (2017) 155–163.
[68] B. Liu, F. Yang, 2L-piRNA: a two-layer ensemble classifier for identifying piwi-interacting RNAs and their function, Molecular Therapy - Nucleic Acids 7 (2017) 267–277.
[69] X. Cheng, S.G. Zhao, X. Xiao, iATC-mISF: a multi-label classifier for predicting the classes of anatomical therapeutic chemicals, Bioinformatics 33 (2017) 341–346 Corrigendum, ibid., 2017, Vol.33, 2610.
[70] B. Liu, S. Wang, R. Long, iRSpot-EL: identify recombination spots with an ensemble learning approach, Bioinformatics 33 (2017) 35–41.
[71] W.R. Qiu, S.Y. Jiang, Z.C. Xu, X. Xiao, iRNAm5C-PseDNC: identifying RNA 5-methylcytosine sites by incorporating physical-chemical properties into pseudo dinucleotide composition, Oncotarget 8 (2017) 41178–41188.
[72] W.R. Qiu, B.Q. Sun, X. Xiao, Z.C. Xu, J.H. Jia, iKcr-PseEns: Identify lysine crotonylation sites in histone proteins with pseudo components and ensemble classifier, Genomics 110 (2018) 239–246.
[73] J. Song, F. Li, K. Takemoto, G. Haffari, T. Akutsu, G.I. Webb, PREvaIL, an integrative approach for inferring catalytic residues using sequence, structural and network features in a machine learning framework, J. Theor. Biol. 443 (2018) 125–137.
[74] W. Chen, P. Feng, H. Yang, H. Ding, H. Lin, iRNA-3typeA: identifying 3-types of modification at RNA's adenosine sites, Molecular Therapy: Nucleic Acid 11 (2018) 468–474.
[75] J. Song, Y. Wang, F. Li, T. Akutsu, N.D. Rawlings, G.I. Webb, iProt-sub: a comprehensive package for accurately mapping and predicting protease-specific substrates and cleavage sites, Brief. Bioinform. (2018), https://doi.org/10.1093/bib/bby028.
[76] P. Feng, H. Yang, H. Ding, H. Lin, W. Chen, iDNA6mA-PseKNC: Identifying DNA N6-methyladenosine sites by incorporating nucleotide physicochemical properties into PseKNC, Genomics (2018), https://doi.org/10.1016/j.ygeno.2018.01.005.
[77] B. Liu, F. Weng, D.S. Huang, iRO-3wPseKNC: Identify DNA replication origins by three-window-based PseKNC, Bioinformatics (2018), https://doi.org/10.1093/bioinformatics/bty312/4978052.
[78] H. Yang, W.R. Qiu, G. Liu, F.B. Guo, W. Chen, H. Lin, iRSpot-Pse6NC: Identifying recombination spots in *Saccharomyces cerevisiae* by incorporating hexamer composition into general PseKNC International Journal of Biological Sciences 14 (2018) 883–891.
[79] Y.D. Khan, N. Rasool, W. Hussain, S.A. Khan, iPhosT-PseAAC: Identify phosphothreonine sites by incorporating sequence statistical moments into PseAAC, Anal. Biochem. 550 (2018) 109–116.
[80] Z.D. Su, Y. Huang, Z.Y. Zhang, Y.W. Zhao, D. Wang, W. Chen, H. Lin, iLoc-lncRNA: predict the subcellular location of lncRNAs by incorporating octamer composition into general PseKNC, Bioinformatics (2018), https://doi.org/10.1093/bioinformatics/bty508.
[81] K.C. Chou, Some remarks on protein attribute prediction and pseudo amino acid composition (50th anniversary Year Review), J. Theor. Biol. 273 (2011) 236–247.
[82] K.C. Chou, Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes, Bioinformatics 21 (2005) 10–19.
[83] Y.X. Pan, Z.Z. Zhang, Z.M. Guo, G.Y. Feng, Z.D. Huang, L. He, Application of pseudo amino acid composition for predicting protein subcellular location: stochastic signal processing approach, J. Protein Chem. 22 (2003) 395–402.
[84] K.C. Chou, Y.D. Cai, Predicting subcellular localization of proteins by hybridizing functional domain composition and pseudo amino acid composition, J. Cell. Biochem. 91 (2004) 1197–1203.
[85] M. Wang, J. Yang, G.P. Liu, Z.J. Xu, Weighted-support vector machines for predicting membrane protein types based on pseudo amino acid composition, Protein Eng. Des. Sel. 17 (2004) 509–516.
[86] Y.D. Cai, Predicting enzyme subclass by functional domain composition and pseudo amino acid composition, J. Proteome Res. 4 (2005) 967–971.
[87] Y.D. Cai, G.P. Zhou, Predicting enzyme family classes by hybridizing gene product composition and pseudo amino acid composition, J. Theor. Biol. 234 (2005) 145–149.
[88] Y. Gao, S.H. Shao, X. Xiao, Y.S. Ding, Y.S. Huang, Z.D. Huang, Using pseudo amino acid composition to predict protein subcellular location: approached with Lyapunov index, Bessel function, and Chebyshev filter, Amino Acids 28 (2005) 373–376.
[89] X. Xiao, S. Shao, Y. Ding, Z. Huang, X. Chen, Using cellular automata to generate image representation for biological sequences, Amino Acids 28 (2005) 29–35.
[90] H.B. Shen, K.C. Chou, Predicting protein subnuclear location with optimized evidence-theoretic K-nearest classifier and pseudo amino acid composition, Biochem Biophys Res Comm. (BBRC) 337 (2005) 752–756.
[91] Y.D. Cai, Predicting membrane protein type by functional domain composition and pseudo amino acid composition, J. Theor. Biol. 238 (2006) 395–400.
[92] X. Xiao, S.H. Shao, Y.S. Ding, Z.D. Huang, Using cellular automata images and

pseudo amino acid composition to predict protein subcellular location, Amino Acids 30 (2006) 49–54.
[93] S.Q. Wang, J. Yang, Using stacked generalization to predict membrane protein types based on pseudo amino acid composition, J. Theor. Biol. 242 (2006) 941–946.
[94] G.P. Zhou, Y.D. Cai, Predicting protease types by hybridizing gene ontology and pseudo amino acid composition, PROTEINS: Structure, Function, and Bioinformatics 63 (2006) 681–684.
[95] X.B. Zhou, C. Chen, Z.C. Li, X.Y. Zou, Using Chou's amphiphilic pseudo amino acid composition and support vector machine for prediction of enzyme subfamily classes, J. Theor. Biol. 248 (2007) 546–551.
[96] L. Nanni, A. Lumini, Genetic programming for creating Chou's pseudo amino acid based features for submitochondria localization, Amino Acids 34 (2008) 653–660.
[97] D.N. Georgiou, T.E. Karakasidis, J.J. Nieto, A. Torres, Use of fuzzy clustering technique and matrices to classify amino acids and its impact to Chou's pseudo amino acid composition, J. Theor. Biol. 257 (2009) 17–26.
[98] M. Esmaeili, H. Mohabatkar, S. Mohsenzadeh, Using the concept of Chou's pseudo amino acid composition for risk type prediction of human papillomaviruses, J. Theor. Biol. 263 (2010) 203–209.
[99] H. Mohabatkar, Prediction of cyclin proteins using Chou's pseudo amino acid composition, Protein Pept. Lett. 17 (2010) 1207–1214.
[100] H. Mohabatkar, M. Mohammad Beigi, A. Esmaeili, Prediction of GABA(a) receptor proteins using the concept of Chou's pseudo amino acid composition and support vector machine, J. Theor. Biol. 281 (2011) 18–23.
[101] B.M. Mohammad, M. Behjati, H. Mohabatkar, Prediction of metalloproteinase family based on the concept of Chou's pseudo amino acid composition using a machine learning approach, J. Struct. Funct. Genom. 12 (2011) 191–197.
[102] M. Hayat, A. Khan, Discriminating Outer Membrane Proteins with Fuzzy K-Nearest Neighbor Algorithms based on the General form of Chou's PseAAC, Protein Pept. Lett. 19 (2012) 411–421.
[103] L. Nanni, S. Brahnam, A. Lumini, Wavelet images and Chou's pseudo amino acid composition for protein classification, Amino Acids 43 (2012) 657–665.
[104] M. Khosravian, F.K. Faramarzi, M.M. Beigi, M. Behbahani, H. Mohabatkar, Predicting Antibacterial Peptides by the Concept of Chou's Pseudo amino Acid Composition and Machine Learning Methods, Protein Pept. Lett. 20 (2013) 180–186.
[105] Z. Hajisharifi, M. Piryaiee, M. Mohammad Beigi, M. Behbahani, H. Mohabatkar, Predicting anticancer peptides with Chou's pseudo amino acid composition and investigating their mutagenicity via Ames test, J. Theor. Biol. 341 (2014) 34–40.
[106] M. Hayat, N. Iqbal, Discriminating protein structure classes by incorporating Pseudo Average Chemical Shift to Chou's general PseAAC and support Vector Machine, Comput. Methods Prog. Biomed. 116 (2014) 184–192.
[107] R. Kumar, A. Srivastava, B. Kumari, M. Kumar, Prediction of beta-lactamase and its class by Chou's pseudo amino acid composition and support vector machine, J. Theor. Biol. 365 (2015) 96–103.
[108] K. Ahmad, M. Waris, M. Hayat, Prediction of Protein Submitochondrial Locations by Incorporating Dipeptide Composition into Chou's General Pseudo Amino Acid Composition, J. Membr. Biol. 249 (2016) 293–304.
[109] M. Behbahani, H. Mohabatkar, M. Nosrati, 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. 411 (2016) 1–5.
[110] P.K. Meher, T.K. Sahu, V. Saini, A.R. Rao, Predicting antimicrobial peptides with improved accuracy by incorporating the compositional, physico-chemical and structural features into Chou's general PseAAC, Sci. Rep. 7 (2017) 42362.
[111] M. Rahimi, M.R. Bakhtiarizadeh, A. Mohammadi-Sangcheshmeh, OOgenesis_Pred: a sequence-based method for predicting oogenesis proteins by six different modes of Chou's pseudo amino acid composition, J. Theor. Biol. 414 (2017) 128–136.
[112] M. Arif, M. Hayat, Z. Jan, 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. 442 (2018) 11–21.
[113] M.A. Al Maruf, S. Shatabda, iRSpot-SF: Prediction of recombination hotspots by incorporating sequence based features into Chou's Pseudo components, Genomics (2018), https://doi.org/10.1016/j.ygeno.2018.06.003.
[114] S. Akbar, M. Hayat, 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), https://doi.org/10.1016/j.jtbi.2018.07.018.
[115] E. Contreras-Torres, Predicting structural classes of proteins by incorporating their global and local physicochemical and conformational properties into general Chou's PseAAC, J. Theor. Biol. 454 (2018) 139–145.
[116] Z. Ju, S.Y. Wang, Prediction of citrullination sites by incorporating k-spaced amino acid pairs into Chou's general pseudo amino acid composition, Gene 664 (2018) 78–83.
[117] M.S. Krishnan, 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. 445 (2018) 62–74.
[118] Y. Liang, S. Zhang, 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. 454 (2018) 22–29.
[119] J. Mei, Y. Fu, J. Zhao, Analysis and prediction of ion channel inhibitors by using feature selection and Chou's general pseudo amino acid composition, J Theor Biol (2018), https://doi.org/10.1016/j.jtbi.2018.07.040.
[120] J. Mei, J. Zhao, Prediction of HIV-1 and HIV-2 proteins by using Chou's pseudo amino acid compositions and different classifiers, Sci. Rep. 8 (2018) 2359.
[121] W. Qiu, S. Li, X. Cui, Z. Yu, M. Wang, J. Du, Y. Peng, B. Yu, Predicting protein submitochondrial locations by incorporating the pseudo-position specific scoring matrix into the general Chou's pseudo-amino acid composition, J. Theor. Biol. 450 (2018) 86–103.
[122] S.M. Rahman, S. Shatabda, S. Saha, M. Kaykobad, M. Sohel Rahman, DPP-PseAAC: a DNA-binding Protein Prediction model using Chou's general PseAAC, J. Theor. Biol. 452 (2018) 22–34.
[123] E.S. Sankari, D.D. Manimegalai, Predicting membrane protein types by incorporating a novel feature set into Chou's general PseAAC, J Theor Biol (2018), https://doi.org/10.1016/j.jtbi.2018.07.032.
[124] S. Zhang, X. Duan, Prediction of protein subcellular localization with oversampling approach and Chou's general PseAAC, J. Theor. Biol. 437 (2018) 239–250.
[125] K.C. Chou, An unprecedented revolution in medicinal chemistry driven by the progress of biological science, Curr. Top. Med. Chem. 17 (2017) 2337–2358.
[126] W. Chen, T.Y. Lei, D.C. Jin, H. Lin, PseKNC: a flexible web-server for generating pseudo K-tuple nucleotide composition, Anal. Biochem. 456 (2014) 53–60.
[127] W. Chen, P.M. Feng, H. Lin, iSS-PseDNC: identifying splicing sites using pseudo dinucleotide composition, Biomed Research International (BMRI) 2014 (2014) 623149.
[128] W. Chen, P.M. Feng, E.Z. Deng, H. Lin, iTIS-PseTNC: a sequence-based predictor for identifying translation initiation site in human genes using pseudo trinucleotide composition, Anal. Biochem. 462 (2014) 76–83.
[129] W. Chen, H. Lin, Pseudo nucleotide composition or PseKNC: an effective formulation for analyzing genomic sequences, Mol. BioSyst. 11 (2015) 2620–2634.
[130] B. Liu, L. Fang, R. Long, X. Lan, iEnhancer-2L: a two-layer predictor for identifying enhancers and their strength by pseudo k-tuple nucleotide composition, Bioinformatics 32 (2016) 362–369.
[131] B. Liu, R. Long, iDHS-EL: Identifying DNase I hypersensi-tivesites by fusing three different modes of pseudo nucleotide composition into an ensemble learning framework, Bioinformatics 32 (2016) 2411–2418.
[132] B. Liu, K. Li, D.S. Huang, iEnhancer-EL: Identifying enhancers and their strength with ensemble learning approach, Bioinformatics (2018), https://doi.org/10.1093/bioinformatics/bty458.
[133] M.F. Sabooh, N. Iqbal, M. Khan, M. Khan, H.F. Maqbool, Identifying 5-methylcytosine sites in RNA sequence using composite encoding feature into Chou's PseKNC, J. Theor. Biol. 452 (2018) 1–9.
[134] L. Zhang, L. Kong, iRSpot-ADPM: Identify recombination spots by incorporating the associated dinucleotide product model into Chou's pseudo components, J. Theor. Biol. 441 (2018) 1–8.
[135] L. Zhang, L. Kong, iRSpot-PDI: Identification of recombination spots by incorporating dinucleotide property diversity information into Chou's pseudo components, Genomics (2018), https://doi.org/10.1016/j.ygeno.2018.03.003.
[136] B. Liu, F. Liu, X. Wang, J. Chen, L. Fang, Pse-in-one: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences, Nucleic Acids Res. 43 (2015) W65–W71.
[137] B. Liu, H. Wu, 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 (2017) 67–91.
[138] C.T. Zhang, Monte Carlo simulation studies on the prediction of protein folding types from amino acid composition, Biophys. J. 63 (1992) 1523–1529.
[139] K.C. Chou, A vectorized sequence-coupling model for predicting HIV protease cleavage sites in proteins, J. Biol. Chem. 268 (1993) 16938–16948.
[140] C.T. Zhang, An analysis of protein folding type prediction by seed-propagated sampling and jackknife test, J. Protein Chem. 14 (1995) 583–593.
[141] Y.D. Cai, K.Y. Feng, W.C. Lu, Using LogitBoost classifier to predict protein structural classes, J. Theor. Biol. 238 (2006) 172–176.
[142] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, J. Artif. Intell. Res. 16 (2011) 321–357.
[143] J. Jia, Z. Liu, X. Xiao, B. Liu, Identification of protein-protein binding sites by incorporating the physicochemical properties and stationary wavelet transforms into pseudo amino acid composition (iPPBS-PseAAC), J Biomol Struct Dyn (JBSD) 34 (2016) 1946–1961.
[144] W.R. Qiu, B.Q. Sun, X. Xiao, Z.C. Xu, iPTM-mLys: identifying multiple lysine PTM sites and their different types, Bioinformatics 32 (2016) 3116–3123.
[145] X. Cheng, S.G. Zhao, X. Xiao, iATC-mHyb: a hybrid multi-label classifier for predicting the classification of anatomical therapeutic chemicals, Oncotarget 8 (2017) 58494–58503.
[146] K.C. Chou, C.T. Zhang, Review: Prediction of protein structural classes, Crit. Rev. Biochem. Mol. Biol. 30 (1995) 275–349.
[147] G.P. Zhou, N. Assa-Munt, Some insights into protein structural class prediction, Proteins Struct. Funct. Genet. 44 (2001) 57–59.
[148] D.W. Elrod, Prediction of enzyme family classes, J. Proteome Res. 2 (2003) 183–190.
[149] K.C. Chou, H.B. Shen, MemType-2L: a Web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM, Biochem Biophys Res Comm (BBRC) 360 (2007) 339–345.
[150] F. Ali, M. Hayat, Classification of membrane protein types using Voting Feature Interval in combination with Chou's Pseudo Amino Acid Composition, J. Theor. Biol. 384 (2015) 78–83.
[151] M. Tahir, M. Hayat, iNuc-STNC: a sequence-based predictor for identification of nucleosome positioning in genomes by extending the concept of SAAC and Chou's PseAAC, Mol. BioSyst. 12 (2016) 2587–2593.
[152] M. Khan, M. Hayat, S.A. Khan, N. Iqbal, Unb-DPC: Identify mycobacterial membrane protein types by incorporating un-biased dipeptide composition into Chou's general PseAAC, J. Theor. Biol. 415 (2017) 13–19.
[153] J.Z. Cao, W.Q. Liu, H. Gu, Predicting Viral Protein Subcellular Localization with Chou's Pseudo Amino Acid Composition and Imbalance-Weighted Multi-Label K-Nearest Neighbor Algorithm, Protein and Peptide Letters 19 (2012) 1163–1169.

[154] J. He, H. Gu, W. Liu, Imbalanced multi-modal multi-label learning for subcellular localization prediction of human proteins with both single and multiple sites, PLoS One 7 (2012) e37155.
[155] L.Q. Li, Y. Zhang, L.Y. Zou, Y. Zhou, X.Q. Zheng, Prediction of Protein Subcellular Multi-Localization based on the General form of Chou's Pseudo Amino Acid Composition, Protein Pept. Lett. 19 (2012) 375–387.
[156] S. Mei, Predicting plant protein subcellular multi-localization by Chou's PseAAC formulation based multi-label homolog knowledge transfer learning, J. Theor. Biol. 310 (2012) 80–87.
[157] X. Wang, G.Z. Li, A multi-label predictor for identifying the subcellular locations of singleplex and multiplex eukaryotic proteins, PLoS One 7 (2012) e36317.
[158] W. Chen, P.M. Feng, H. Lin, iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition, Nucleic Acids Research 41 (2013) e68.
[159] Y. Xu, X.J. Shao, L.Y. Wu, N.Y. Deng, iSNO-AAPair: incorporating amino acid pairwise coupling into PseAAC for predicting cysteine S-nitrosylation sites in proteins, PeerJ 1 (2013) e171.
[160] K.C. Chou, Using subsite coupling to predict signal peptides, Protein Eng. 14 (2001) 75–79.
[161] K.C. Chou, Prediction of signal peptides using scaled window, Peptides 22 (2001) 1973–1979.
[162] H. Lin, E.Z. Deng, H. Ding, W. Chen, iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition, Nucleic Acids Res. 42 (2014) 12961–12972.
[163] H. Ding, E.Z. Deng, L.F. Yuan, L. Liu, H. Lin, W. Chen, iCTX-type: a sequence-based predictor for identifying the types of conotoxins in targeting ion channels, BioMed Research International (BMRI) 2014 (2014) 286419.
[164] W.R. Qiu, X. Xiao, iRSpot-TNCPseAAC: Identify recombination spots with trinucleotide composition and pseudo amino acid components, Int J Mol Sci (IJMS) 15 (2014) 1746–1766.
[165] R. Xu, J. Zhou, B. Liu, Y.A. He, Q. Zou, X. Wang, Identification of DNA-binding proteins by incorporating evolutionary information into pseudo amino acid composition via the top-n-gram approach, Journal of Biomolecular Structure & Dynamics (JBSD) 33 (2015) 1720–1730.
[166] B. Liu, L. Fang, S. Wang, X. Wang, H. Li, Identification of microRNA precursor with the degenerate K-tuple or Kmer strategy, J. Theor. Biol. 385 (2015) 153–159.
[167] J. Jia, Z. Liu, X. Xiao, iPPI-Esml: an ensemble classifier for identifying the interactions of proteins by incorporating their physicochemical properties and wavelet transforms into PseAAC, J. Theor. Biol. 377 (2015) 47–56.
[168] W. Chen, P. Feng, H. Ding, H. Lin, iRNA-methyl: Identifying N6-methyladenosine sites using pseudo nucleotide composition, Anal. Biochem. 490 (2015) 26–33.
[169] M. Kabir, M. Hayat, iRSpot-GAEnsC: identifing recombination spots via ensemble classifier and extending the concept of Chou's PseAAC to formulate DNA samples, Mol. Gen. Genomics. 291 (2016) 285–296.
[170] W. Chen, H. Ding, P. Feng, H. Lin, iACP: a sequence-based tool for identifying anticancer peptides, Oncotarget 7 (2016) 16895–16909.
[171] W. Chen, P. Feng, H. Ding, H. Lin, Using deformation energy to analyze nucleosome positioning in genomes, Genomics 107 (2016) 69–75.
[172] J. Jia, L. Zhang, Z. Liu, X. Xiao, pSumo-CD: predicting sumoylation sites in proteins with covariance discriminant algorithm by incorporating sequence-coupled effects into general PseAAC, Bioinformatics 32 (2016) 3133–3141.
[173] W.R. Qiu, B.Q. Sun, X. Xiao, D. Xu, iPhos-PseEvo: Identifying human phosphorylated proteins by incorporating evolutionary information into general PseAAC via grey system theory, Molecular Informatics 36 (2017) UNSP 1600010.
[174] P.M. Feng, W. Chen, H. Lin, iHSP-PseRAAAC: Identifying the heat shock protein families using pseudo reduced amino acid alphabet composition, Anal. Biochem. 442 (2013) 118–125.
[175] H.B. Shen, Recent advances in developing web-servers for predicting protein attributes, Nat. Sci. 1 (2009) 63–92.
[176] J. Jia, Z. Liu, X. Xiao, B. Liu, iCar-PseCp: identify carbonylation sites in proteins by Monto Carlo sampling and incorporating sequence coupled effects into general PseAAC, Oncotarget 7 (2016) 34558–34570.
[177] W.R. Qiu, B.Q. Sun, X. Xiao, Z.C. Xu, iHyd-PseCp: Identify hydroxyproline and hydroxylysine in proteins by incorporating sequence-coupled effects into general PseAAC, Oncotarget 7 (2016) 44310–44321.
[178] W.R. Qiu, X. Xiao, Z.C. Xu, iPhos-PseEn: identifying phosphorylation sites in proteins by fusing different pseudo components into an ensemble classifier, Oncotarget 7 (2016) 51270–51283.
[179] W. Chen, P. Feng, H. Yang, H. Ding, H. Lin, iRNA-AI: identifying the adenosine to inosine editing sites in RNA sequences, Oncotarget 8 (2017) 4208–4217.
[180] B. Liu, H. Wu, D. Zhang, X. Wang, Pse-Analysis: a python package for DNA/RNA and protein/peptide sequence analysis based on pseudo components and kernel methods, Oncotarget 8 (2017) 13338–13343.
[181] J. Jia, Z. Liu, X. Xiao, B. Liu, pSuc-Lys: Predict lysine succinylation sites in proteins with PseAAC and ensemble random forest approach, J. Theor. Biol. 394 (2016) 223–230.
[182] Y. Xu, X. Wen, L.S. Wen, L.Y. Wu, N.Y. Deng, iNitro-Tyr: Prediction of nitrotyrosine sites in proteins with general pseudo amino acid composition, PLoS One 9 (2014) e105018.
[183] Z. Liu, X. Xiao, D.J. Yu, J. Jia, W.R. Qiu, pRNAm-PC: predicting N-methyladenosine sites in RNA sequences via physical-chemical properties, Anal. Biochem. 497 (2016) 60–67.
[184] Z. Chen, P.Y. Zhao, F. Li, A. Leier, T.T. Marquez-Lago, Y. Wang, G.I. Webb, A.I. Smith, R.J. Daly, J. Song, iFeature: a python package and web server for features extraction and selection from protein and peptide sequences, Bioinformatics 34 (2018) 2499–2502.
[185] J. Song, F. Li, A. Leier, T.T. Marquez-Lago, T. Akutsu, G. Haffari, G.I. Webb, R.N. Pike, PROSPERous: high-throughput prediction of substrate cleavage sites for 90 proteases with improved accuracy, Bioinformatics 34 (2018) 684–687.
[186] J. Wang, B. Yang, A. Leier, T.T. Marquez-Lago, M. Hayashida, A. Rocker, Z. Yanju, T. Akutsu, R.A. Strugnell, J. Song, T. Lithgow, Bastion6: a bioinformatics approach for accurate prediction of type VI secreted effectors, Bioinformatics (2018), https://doi.org/10.1093/bioinformatics/bty155.
[187] J. Wang, B. Yang, J. Revote, A. Leier, T.T. Marquez-Lago, G. Webb, J. Song, T. Lithgow, POSSUM: a bioinformatics toolkit for generating numerical sequence feature descriptors based on PSSM profiles, Bioinformatics 33 (2017) 2756–2758.
[188] X. Xuao, X. Cheng, G. Chen, Q. Mao, pLoc_bal-mGpos: predict subcellular localization of Gram-positive bacterial proteins by quasi-balancing training dataset and PseAAC, Genomics (2018), https://doi.org/10.1016/j.ygeno.2018.05.017.
[189] F. Li, C. Li, T.T. Marquez-Lago, A. Leier, T. Akutsu, A.W. Purcell, A.I. Smith, T. Lightow, R.J. Daly, J. Song, Quokka: a comprehensive tool for rapid and accurate prediction of kinase family-specific phosphorylation sites in the human proteome, Bioinformatics (2018), https://doi.org/10.1093/bioinformatics/bty522.
[190] Y. Zuo, Y. Li, Y. Chen, G. Li, Z. Yan, L. Yang, PseKRAAC: a flexible web server for generating pseudo K-tuple reduced amino acids composition, Bioinformatics 33 (2017) 122–124.
[191] P. Feng, H. Lin, W. Chen, Y. Zuo, Predicting the types of J-proteins using clustered amino acids, Biomed. Res. Int. 2014 (2014) 935719.
[192] Y.C. Zuo, Q.Z. Li, Using K-minimum increment of diversity to predict secretory proteins of malaria parasite based on groupings of amino acids, Amino Acids 38 (2010) 859–867.
[193] Y. Zuo, Y. Lv, Z. Wei, L. Yang, G. Li, G. Fan, iDPF-PseRAAAC: a Web-Server for Identifying the Defensin Peptide Family and Subfamily using Pseudo Reduced Amino Acid Alphabet Composition, PLoS One 10 (2015) e0145541.
[194] Y.C. Zuo, Q.Z. Li, Using reduced amino acid composition to predict defensin family and subfamily: Integrating similarity measure and structural alphabet, Peptides 30 (2009) 1788–1793.