Bioinformatics Advance Access published July 21, 2016

Bioinformatics, 2016, 1–8
doi: 10.1093/bioinformatics/btw380
Advance Access Publication Date: 22 June 2016
Original Paper

Sequence analysis

iPTM-mLys: identifying multiple lysine PTM sites and their different types

Wang-Ren Qiu[1,2,3,*], Bi-Qian Sun[1], Xuan Xiao[1,3,*], Zhao-Chun Xu[1] and Kuo-Chen Chou[3,4,5,*]

[1]Computer Department, Jingdezhen Ceramic Institute, Jingdezhen 333403, China, [2]Department of Computer Science and Bond Life Science Center, University of Missouri, Columbia MO, USA; [3]Computational Biology, Gordon Life Science Institute, Boston, MA 02478, USA, [4]Center of Bioinformatics, School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu 610054, China and [5]Center of Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz University, Jeddah 21589, Saudi Arabia

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on May 23, 2016; revised on June 6, 2016; accepted on June 13, 2016

Abstract

Motivation: Post-translational modification, abbreviated as PTM, refers to the change of the amino acid side chains of a protein after its biosynthesis. Owing to its significance for in-depth understanding various biological processes and developing effective drugs, prediction of PTM sites in proteins have currently become a hot topic in bioinformatics. Although many computational methods were established to identify various single-label PTM types and their occurrence sites in proteins, no method has ever been developed for multi-label PTM types. As one of the most frequently observed PTMs, the K-PTM, namely, the modification occurring at lysine (K), can be usually accommodated with many different types, such as 'acetylation', 'crotonylation', 'methylation' and 'succinylation'. Now we are facing an interesting challenge: given an uncharacterized protein sequence containing many K residues, which ones can accommodate two or more types of PTM, which ones only one, and which ones none?

Results: To address this problem, a multi-label predictor called **iPTM-mLys** has been developed. It represents the first multi-label PTM predictor ever established. The novel predictor is featured by incorporating the sequence-coupled effects into the general PseAAC, and by fusing an array of basic random forest classifiers into an ensemble system. Rigorous cross-validations via a set of multi-label metrics indicate that the first multi-label PTM predictor is very promising and encouraging.

Availability and Implementation: For the convenience of most experimental scientists, a user-friendly web-server for **iPTM-mLys** has been established at http://www.jci-bioinfo.cn/iPTM-mLys, by which users can easily obtain their desired results without the need to go through the complicated mathematical equations involved.

Contact: wqiu@gordonlifescience.org, xxiao@gordonlifescience.org, kcchou@gordonlifescience.org

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Post-translational modification (PTM or PTLM) means the change of the amino acid side chains of a protein after its biosynthesis. Since the importance of PTM to basic research and drug development, identification of PTM sites in proteins has become a very hot topic in bioinformatics (Chou, 2015a, b; Jia *et al.*, 2016a, c, d; Qiu, *et al.*, 2014, 2015, 2016a, b; Xu and Chou, 2016; Xu *et al.*, 2013a, b, 2014a, b).

Among the 20 native amino acid residues, the modification at lysine (K), the so-called K-PTM, is one of the most frequently observed PTMs. Furthermore, some Lys residues in proteins can undergo sequential or cascades of covalent modifications; i.e. they can be targeted by various different K-PTM types, such as acetylation, biotinylation, butyrylation, crotonylation, methylation, propionylation, succinylation, ubiquitination and ubiquitin-like modifications. Meanwhile, various computational methods have been developed to predict the modification sites in proteins for different K-PTM types (Chen *et al.*, 2006; Jia *et al.*, 2016c; Qiu *et al.*, 2014, 2015; Shao *et al.*, 2009; Xu *et al.*, 2014a).

But to our best knowledge, so far no computational tool whatsoever can be used to deal with the system that simultaneously contains several different K-PTM types or multiplex Lys residues. Actually, this kind of multiplex Lys residues in proteins may have some exceptional functions worthy of our special notice for both basic research and drug development.

In view of this, the present study was initiated in an attempt to fill such an empty field by establishing a novel method that can be used to predict the multiple K-type modifications in proteins.

According to the Chou's five-step guidelines (Chou, 2011) and followed by many investigators in a series of recent publications (Chen *et al.*, 2016a, b; Jia *et al.*, 2016a, b, c, d, e; Liu *et al.*, 2016a, b, c, d; Qiu *et al.*, 2016a, b, c; Xiao *et al.*, 2016), for developing a new prediction method that can be widely used by broad users, we should consider the following five points: (i) how to construct or select a valid benchmark dataset to train and test the predictor; (ii) how to formulate the biological sequence samples with an effective mathematical expression that can truly reflect their essential correlation with the target concerned; (iii) how to introduce or develop a powerful algorithm (or engine) to run the prediction; (iv) how to properly conduct cross-validation tests to objectively evaluate the anticipated accuracy; (v) how to provide a web-server and user guide to make people very easily to get their desired results. In the rest of this article, we are to address these point by point.

2 Materials and methods

2.1 Benchmark dataset

Note that so far, the statistically significant and experiment-confirmed data are available only for the four K-PTM types: acetylation, crotonylation, methylation and succinylation. Thus, the proteins used in this study were collected according to the following procedures: (i) Open the web site at http://www.uniprot.org/, and click the button 'Advanced'. (ii) Select 'PTM/Processing' and 'Modified residue [FT]' for 'Fields'. (iii) Select 'Any experimental assertion' for 'Evidence'. (iv) Type 'human' for 'Term' to do search. (v) Collected were only those proteins that with keywords of 'acetyllysine', 'crotonyllysine', 'methyllysine' or 'succinyllysine'. (vi) Collected were only those proteins that contain 50 and more amino acid residues. Finally, we obtained 1769 working proteins.

To make the description logically more rigorous and clear, the Chou's scheme (Chou, 2001b) was adopted to formulate the peptide samples, as done recently by many authors in studying the nitrotyrosine sites (Xu *et al.*, 2014b), methylation sites (Qiu *et al.*, 2014) and carbonylation sites (Jia *et al.*, 2016a). According to Chou's scheme, a potential K-PTM site-containing sample can be generally expressed by

$$\mathbf{P}_{\xi}(\mathbb{K}) = \mathrm{R}_{-\xi}\mathrm{R}_{-(\xi-1)}\cdots\mathrm{R}_{-2}\mathrm{R}_{-1}\mathbb{K}\mathrm{R}_{+1}\mathrm{R}_{+2}\cdots\mathrm{R}_{+(\xi-1)}\mathrm{R}_{+\xi}, \tag{1}$$

where the symbol $\mathbb{K}$ denotes the single amino acid code K, the subscript ξ is an integer, $\mathrm{R}_{-\xi}$ represents the ξ-th upstream amino acid residue from the center, the $\mathrm{R}_{+\xi}$ the ξ-th downstream amino acid residue, and so forth.

The detailed procedures to construct the benchmark datasets are as follows. (i) As done in (Chou and Shen, 2007b), slide the $(2\xi+1)$-tuple peptide window along each of the aforementioned 1769 protein sequences, and collected were only those peptide segments that have K (Lys or lysine) at the center (**Equation 1**). (ii) If the upstream or downstream in a protein sequence was less than ξ or greater than $L-\xi$ (L is the length of the protein sequence concerned), the lacking amino acid was filled with the same residue as its nearest one. (iii) The peptide segment samples thus obtained were put into the positive subset if their centers have been experimentally annotated as the acetylation site; otherwise, into the negative subset. (iv) If there were two or more samples sharing a same sequence, kept was only one of them. (v) Repeat the cycle of (i–iv) by successively changing 'acetylation' in (iii) to 'crotonylation', 'methylation' and 'succinylation', respectively. By doing so, we obtained the following four benchmark datasets

$$\begin{cases} \mathbb{S}_{\xi}(\text{acetylation}) = \mathbb{S}_{\xi}^{+}(\text{acetylation}) \cup \mathbb{S}_{\xi}^{-}(\text{acetylation}) \\ \mathbb{S}_{\xi}(\text{crotonylation}) = \mathbb{S}_{\xi}^{+}(\text{crotonylation}) \cup \mathbb{S}_{\xi}^{-}(\text{crotonylation}) \\ \mathbb{S}_{\xi}(\text{methylation}) = \mathbb{S}_{\xi}^{+}(\text{methylation}) \cup \mathbb{S}_{\xi}^{-}(\text{methylation}) \\ \mathbb{S}_{\xi}(\text{succinylation}) = \mathbb{S}_{\xi}^{+}(\text{succinylation}) \cup \mathbb{S}_{\xi}^{-}(\text{succinylation}) \end{cases}, \tag{2}$$

where the positive subset $\mathbb{S}_{\xi}^{+}(\text{acetylation})$ contains only the peptide samples with their center residues K (**Equation 1**) confirmed by experiments being able to be of acetylation, while the negative subset $\mathbb{S}_{\xi}^{-}(\text{acetylation})$ only contains those samples unable to be of acetylation, and the symbol $\cup$ denotes the union in the set theory. Likewise, the remaining three sub-equations in **Equation 2** have exactly the same definition but refer to 'crotonylation', 'methylation' and 'succinylation', respectively. As we can see from **Equations 1** and **2**, with different ξ values, the benchmark datasets would contain length-different samples.

However, many preliminary tests had indicated that the best outcomes were obtained when $\xi = 13$; i.e. the sample's length was $2\xi+1=27$. Accordingly, hereafter we are focused on the 27-tuple peptide samples only. Thus, **Equations 1** and **2** can be reduced to

$$\mathbf{P}(\mathbb{K}) = \mathbf{R}_{-13}\mathbf{R}_{-12}\cdots\mathbf{R}_{-2}\mathbf{R}_{-1}\mathbb{K}\mathbf{R}_{+1}\mathbf{R}_{+2}\cdots\mathbf{R}_{+12}\mathbf{R}_{+13} \tag{3}$$

and

$$\begin{cases} \mathbb{S}(1) = \mathbb{S}^{+}(1) \cup \mathbb{S}^{-}(1) \\ \mathbb{S}(2) = \mathbb{S}^{+}(2) \cup \mathbb{S}^{-}(2) \\ \mathbb{S}(3) = \mathbb{S}^{+}(3) \cup \mathbb{S}^{-}(3) \\ \mathbb{S}(4) = \mathbb{S}^{+}(4) \cup \mathbb{S}^{-}(4) \end{cases}, \tag{4}$$

where the numerical argument 1, 2, 3 or 4 denotes 'acetylation', 'crotonylation', 'methylation' or 'succinylation', respectively. The numbers of samples in the benchmark datasets are outlined in **Table 1**, and their detailed sequences and positions in the proteins are given in Supplementary Material.

2.2 Incorporate sequence-coupled effects into general pseudo amino acid composition

With the avalanche of biological sequence generated in the post-genomic age, one of the most important problems in computational biology is how to formulate a biological sequence with a discrete model or a vector, yet still considerably keep its sequence pattern or essential feature. This is because all the existing machine-learning

Table 1. Distribution of sample numbers in the benchmark datasets[a] for studying the sites of multiple K-PTM types in proteins

Attribute	Ace	Cro	Met	Suc
	$\mathbb{S}(1)$	$\mathbb{S}(2)$	$\mathbb{S}(3)$	$\mathbb{S}(4)$
Positive	3991	115	127	1169
Negative	2403	6279	6267	5225

[a] See Supplementary Material.

Ace, acetylation; Cro, crotonylation; Met, methylation; Suc, succinylation.

algorithms can only handle vector but not sequence samples, as elaborated in Chou (2015a).

To address this problem, the pseudo-amino acid composition (Chou, 2001a, 2005) or PseAAC was proposed. Ever since the concept of pseudo-amino acid composition or Chou's PseAAC (Cao, *et al.*, 2013; Du, *et al.*, 2012; Lin and Lapointe, 2013) was proposed, it has rapidly penetrated into nearly all the areas of computational proteomics (Ahmad, *et al.*, 2016; Dehzangi, *et al.*, 2015; Kabir and Hayat, 2016; Khan *et al.*, 2015; Kumar *et al.*, 2015; Mondal and Pai, 2014; Tang *et al.*, 2016; Wang *et al.*, 2015) and a long list of references cited in Chen *et al.* (2015a), Du *et al.* (2014) and many biomedicine and drug development areas (Chen *et al.*, 2016a; Zhong and Zhou, 2014; Zhou, 2015; Zhou and Zhong, 2016). Because it has been widely and increasingly used, recently three powerful open access soft-wares, called 'PseAAC-Builder' (Du *et al.*, 2012), 'propy' (Cao *et al.*, 2013) and 'PseAAC-General' (Du *et al.*, 2014), were established: the former two are for generating various modes of Chou's special PseAAC; while the third one for those of Chou's general PseAAC (Chou, 2011), including not only all the special modes of feature vectors for proteins but also the higher level feature vectors such as 'Functional Domain' mode (Equations 9 and 10 of Chou, 2011), 'Gene Ontology' mode (Equations 11 and 12 of Chou, 2011) and 'Sequential Evolution' or 'PSSM' mode (Equations 13 and 14 of Chou, 2011). Inspired by the successes of using PseAAC to deal with protein/peptide sequences, three web-servers (Chen *et al.*, 2014, 2015b; Liu *et al.*, 2015a) were developed for generating various feature vectors for DNA/RNA sequences as well. Particularly, recently, a powerful web-server called Pse-in-One (Liu *et al.*, 2015b) has been developed that can be used to generate any desired feature vectors for protein/peptide and DNA/RNA sequences according to the need of users' studies.

According to the general PseAAC (Chou, 2011), the peptide sequence of **Equation 3** can be formulated as

$$\mathbf{P}(\mathbb{K}) = \mathbf{P}^{+}(\mathbb{K}) - \mathbf{P}^{-}(\mathbb{K}), \tag{5}$$

where

$$\mathbf{P}^{+}(\mathbb{K}) = \begin{bmatrix} p^{+}_{-13}(R_{-13}|R_{-12}) \\ p^{+}_{-12}(R_{-12}|R_{-11}) \\ \vdots \\ p^{+}_{-2}(R_{-2}|R_{-1}) \\ p^{+}_{-1}(R_{-1}) \\ p^{+}_{+1}(R_{+1}) \\ p^{+}_{+2}(R_{+2}|R_{+1}) \\ \vdots \\ p^{+}_{+12}(R_{+12}|R_{+11}) \\ p^{+}_{+13}(R_{+13}|R_{+12}) \end{bmatrix} \tag{6}$$

and

$$\mathbf{P}^{-}(\mathbb{K}) = \begin{bmatrix} p^{-}_{-13}(R_{-13}|R_{-12}) \\ p^{-}_{-12}(R_{-12}|R_{-11}) \\ \vdots \\ p^{-}_{-2}(R_{-2}|R_{-1}) \\ p^{-}_{-1}(R_{-1}) \\ p^{-}_{+1}(R_{+1}) \\ p^{-}_{+2}(R_{+2}|R_{+1}) \\ \vdots \\ p^{-}_{+12}(R_{+12}|R_{+11}) \\ p^{-}_{+13}(R_{+13}|R_{+12}) \end{bmatrix} \tag{7}$$

In **Equation 6**, $p^{+}_{-13}(\mathrm{R}_{-13}|\mathrm{R}_{-12})$ is the conditional probability of amino acid R_{-13} occurring at the left first position (**Equation 3**) given that its closest right neighbor is R_{-12}, $p^{+}_{-12}(\mathrm{R}_{-12}|\mathrm{R}_{-11})$ is the conditional probability of amino acid R_{-12} occurring at the left second position given that its closest right neighbor is R_{-11}, and so forth. Note that in **Equation 6**, only $p^{+}_{-1}(\mathrm{R}_{-1})$ and $p^{+}_{+1}(\mathrm{R}_{+1})$ are of non-conditional probability because the right neighbor of R_{-1} and the left neighbor of R_{+1} are always K (Lys). All these probability values can be easily derived from the positive training subsets taken from Supplementary Material, as done in Chou (1996). Likewise, the components in **Equation 7** are the same as those in **Equation 6** except for that they are derived from the negative training subsets taken from the same supporting information.

2.3 Ensemble random forest algorithm

The random forest (RF) algorithm is a powerful algorithm and has been widely used in many areas of computational biology (Jia *et al.*, 2015a, b, 2016b, c, d; Kandaswamy *et al.*, 2011; Lin *et al.*, 2011; Pugalenthi *et al.*, 2012). The algorithm of RF is based on the ensemble of a large number of decision trees, where each tree gives a classification and the forest chooses the final classification via the most votes (over all the trees in the forest). In the most commonly used type of RFs, split selection is performed based on the so-called decrease of Gini impurity. In this study, the RF is used to rank the features using Gini importance that is implemented with the machine learning platform scikit-learn. The detailed procedures of RF and its formulation have been very clearly described in (Breiman, 2001), and hence there is no need to repeat here.

In this study, however, the benchmark datasets are extremely unbalanced. As we can see from **Table 1**, for the case of acetylation, the number of positive samples is much larger than that of the negative ones. But for the case of crotonylation, methylation or succinylation, the situation is just opposite: the number of positive samples is much less than that of the negative ones. A predictor trained by a very skewing dataset would inevitably yield many bias errors. To deal with this problem, we resort to the asymmetric bootstrap approach, as described below.

From the four highly unbalanced benchmark datasets (**Equation 2**), we can construct a set of $4 \times m$ balanced datasets by doing m bagging cycles of randomly picking 2403 positive acetylation samples from $\mathbb{S}^{+}(1)$, 115 negative crotonylation samples from $\mathbb{S}^{-}(2)$, 127 negative methylation samples from $\mathbb{S}^{-}(3)$, and 1169 negative succinylation samples from $\mathbb{S}^-(4)$, respectively. The $4 \times m$ balanced datasets thus obtained can be formulated by

$$\begin{cases} \mathbb{S}^{\text{Balance}}_{\text{Boot}(m)}(1) = \mathbb{S}^-(1) \cup \mathbb{S}^+_{\text{Boot}(m)}(1) \\ \mathbb{S}^{\text{Balance}}_{\text{Boot}(m)}(2) = \mathbb{S}^+(2) \cup \mathbb{S}^-_{\text{Boot}(m)}(2) \\ \mathbb{S}^{\text{Balance}}_{\text{Boot}(m)}(3) = \mathbb{S}^+(3) \cup \mathbb{S}^-_{\text{Boot}(m)}(3) \\ \mathbb{S}^{\text{Balance}}_{\text{Boot}(m)}(4) = \mathbb{S}^+(4) \cup \mathbb{S}^-_{\text{Boot}(m)}(4) \end{cases}, \tag{8}$$

where $m = 1, 2, \cdots, 5$; $\mathbb{S}^+_{\text{Boot}(m)}(1)$ contains 2403 positive samples for acetylation, exactly the same as the number of samples in its negative subset $\mathbb{S}^-(1)$ (see column 2 and row 4 of Table 1); $\mathbb{S}^-_{\text{Boot}(m)}(2)$ contains 115 positive samples for crotonylation, exactly the same as the number of samples in $\mathbb{S}^+(2)$; and so forth.

Now, based on the $4 \times 5 = 20$ balanced benchmark datasets in **Equation 8**, we can establish four ensemble predictors by the fusion approach (Chou and Shen, 2007b; Shen and Chou, 2007a), as formulated by

$$\mathbb{RF}^{\text{E}}(i) = \mathbb{RF}_1(i)\forall\mathbb{RF}_2(i)\forall\cdots\forall\mathbb{RF}_5(i) = \forall_{m=1}^{5}\mathbb{RF}_m(i), \tag{9}$$

where $i = 1, 2, 3, 4$; $\mathbb{RF}^{\text{E}}(1), \mathbb{RF}^{\text{E}}(2), \mathbb{RF}^{\text{E}}(3)$ and $\mathbb{RF}^{\text{E}}(4)$ are the ensemble predictors for identifying the acetylation, crotonylation, methylation and succinylation sites, respectively. The symbol $\forall$ denotes the fusing operator (Chou and Shen, 2007a), and $\mathbb{RF}_m(i)$ is an individual RF predictor based on the benchmark dataset $\mathbb{S}^{\text{Balance}}_{\text{Boot}(m)}(i)$ in **Equation 8** with 150 trees for each of the individual predictors. For more detailed about using the fusion approach to form an ensemble predictor, see a comprehensive review (Chou and Shen, 2007a), where a crystal clear description with a set of elegant equations are given and hence there is no need to repeat here.

Finally, the results obtained by the four ensemble classifiers, $\mathbb{RF}^{\text{E}}(1), \mathbb{RF}^{\text{E}}(2), \mathbb{RF}^{\text{E}}(3)$ and $\mathbb{RF}^{\text{E}}(4)$, will be subjected to a combination to display (1) whether the query sample can be of K-PTM, and (2) what K-PTM type/types it can be of.

The predictor obtained thru the above procedures is called **iPTM-mLys**, where 'i' stands for 'identify', 'PTM' for 'post-translational modification' and 'mLys' for 'multiple lysine sites'.

To provide an intuitive picture, a flowchart is given in Figure 1 to illustrate how the $4 \times 5 = 20$ individual RF predictors are fused into four ensemble classifiers, and how their outputs are combined to yield the final results.

3 Results and discussion

As mentioned in Section 1, among the five guidelines in developing a useful predictor, one of them is how to objectively evaluate its anticipated success rates (Chou, 2011). To fulfil this, the following two things need to consider: one is what metrics should be adopted to measure the predictor's quality; the other is what kind of test method should be used to derive the metrics rates. Below, let us address such two problems

3.1 Metrics for measuring the prediction quality of multi-label systems

As shown in Supplementary Material and **Table 1**, we have a total of 6394 samples, of which 3991 are labeled with 'acetylation', 115 with 'crotonylation', 127 with 'methylation', 1169 with 'succinylation' and 1,750 with 'non-K-PTM'. Note that in the above samples, some have two or more labels. Therefore, in the current study, we are dealing with a multi-label system (Chou, 2013) and hence the conventional metrics (Chen *et al.*, 2013; Jia *et al.*, 2016a; Lin *et al.*, 2014; Xiao *et al.*, 2016) defined for single-label systems are no longer valid.

According to Chou's formulation (Chou, 2013), the metrics for a multi-label system can be formulated as

$$\begin{cases} \text{Aiming} = \dfrac{1}{N}\sum_{k=1}^{N}\left(\dfrac{\|\mathbb{L}_k \cap \mathbb{L}_k^*\|}{\|\mathbb{L}_k^*\|}\right) \\ \text{Coverage} = \dfrac{1}{N}\sum_{k=1}^{N}\left(\dfrac{\|\mathbb{L}_k \cap \mathbb{L}_k^*\|}{\|\mathbb{L}_k\|}\right) \\ \text{Accuracy} = \dfrac{1}{N}\sum_{k=1}^{N}\left(\dfrac{\|\mathbb{L}_k \cap \mathbb{L}_k^*\|}{\|\mathbb{L}_k \cup \mathbb{L}_k^*\|}\right) \\ \text{Absolute-True} = \dfrac{1}{N}\sum_{k=1}^{N}\Delta(\mathbb{L}_k, \mathbb{L}_k^*) \\ \text{Absolute-False} = \dfrac{1}{N}\sum_{k=1}^{N}\left(\dfrac{\|\mathbb{L}_k \cup \mathbb{L}_k^*\| - \|\mathbb{L}_k \cap \mathbb{L}_k^*\|}{M}\right) \end{cases} \tag{10}$$

where N is the total number of the samples concerned, M the total number of labels in the system, $\cup$ and $\cap$ the symbols are for the

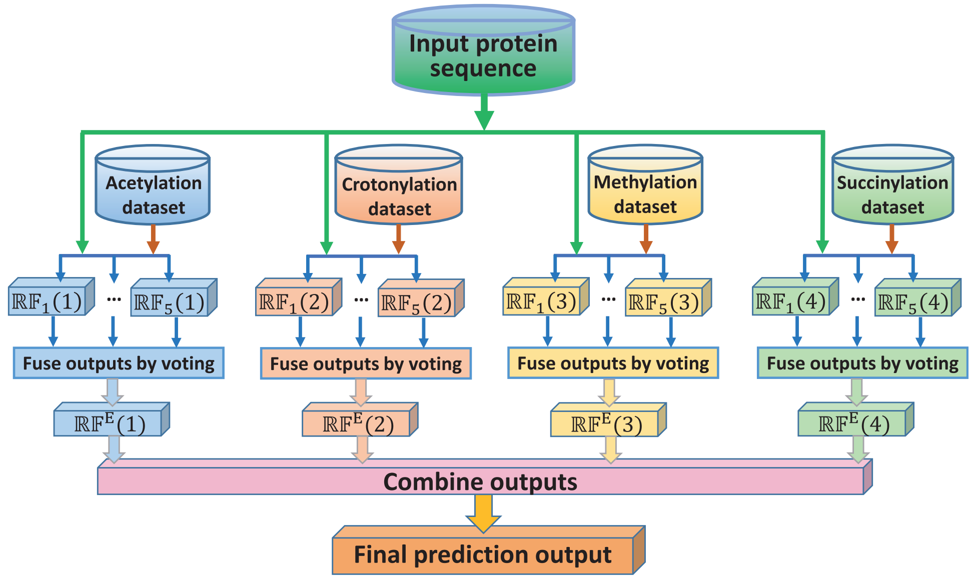

Input protein sequence
Acetylation dataset
Crotonylation dataset
Methylation dataset
Succinylation dataset
$\mathbb{RF}_1(1)$ ··· $\mathbb{RF}_5(1)$
$\mathbb{RF}_1(2)$ ··· $\mathbb{RF}_5(2)$
$\mathbb{RF}_1(3)$ ··· $\mathbb{RF}_5(3)$
$\mathbb{RF}_1(4)$ ··· $\mathbb{RF}_5(4)$
Fuse outputs by voting
Fuse outputs by voting
Fuse outputs by voting
Fuse outputs by voting
$\mathbb{RF}^{\text{E}}(1)$
$\mathbb{RF}^{\text{E}}(2)$
$\mathbb{RF}^{\text{E}}(3)$
$\mathbb{RF}^{\text{E}}(4)$
Combine outputs
Final prediction output

Fig. 1. A flowchart to show how the **iPTM-mLys** predictor works. See the text for further explanation

'union' and 'intersection' in the set theory, || || means the operator acting on the set therein to count the number of its elements, $\mathbb{L}_k$ denotes the subset that contains all the labels experiment-observed for the k-th sample, $\mathbb{L}_k^*$ represents the subset that contains all the labels predicted for the kth sample, and

$$\sum_{k=1}^{N} \Delta\left(\mathbb{L}_k,\ \mathbb{L}_k^*\right) = \begin{cases} 1, & \text{if all the labels in } \mathbb{L}_k^* \text{ are identical with those in } \mathbb{L}_k \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

The above approach had been effectively used to study various multi-label systems, such as those in which a protein may exist in two or more different subcellular locations (Chou *et al.*, 2011, 2012; Lin, *et al.*, 2013; Wu *et al.*, 2011, 2012; Xiao, *et al.*, 2011), or a membrane protein may have two or more different types (Huang and Yuan, 2013), or an antimicrobial peptide may have two or more different types (Xiao *et al.*, 2013).

3.2 Cross-validation

With a set of multi-label metrics clearly defined, the next step is what kind of validation method should be used to derive the metrics values.

The following three cross-validation methods are often used in literature: (i) independent dataset test, (ii) subsampling (or K-fold cross-validation) test and (iii) jackknife test (Chou and Zhang, 1995). 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 (Chou, 2011). Accordingly, the jackknife test has been widely recognized and increasingly used by investigators to examine the quality of various predictors (Ahmad *et al.*, 2015; Chou and Cai, 2005; Dehzangi *et al.*, 2015; Khan *et al.*, 2015; Kumar *et al.*, 2015; Liu *et al.*, 2015c; Nanni *et al.*, 2014; Shen and Chou 2007b; Zhou, 1998; Zhou and Doctor, 2003).

In this study, however, to reduce the computational time, we adopted the 5-fold cross-validation method, as done by many investigators with SVM as the prediction engine.

The 5-fold cross-validation results obtained by the **iPTM-mLys** on the benchmark dataset of Supplementary Material are given by

$$\begin{cases} \text{Aiming} = 69.78\% \\ \text{Coverage} = 74.54\% \\ \text{Accuracy} = 68.37\% \\ \text{Absolute-True} = 60.92\% \\ \text{Absolute-False} = 13.40\% \end{cases} \tag{12}$$

indicating: (i) the rate of "Aiming" or "Precision" (Chou, 2013) is 69.78%, the average ratio of the predicted labels that hit the target of the real labels; (ii) the rate of "Coverage" or "Recall" (Chou, 2013) is 74.54%, the average ratio of the real labels that are covered by the hits of prediction; (iii) the rate of "Accuracy" is 68.37%, the average ratio of the correctly predicted labels over the total labels including correctly and incorrectly predicted ones as well as those real labels but are missed out during the prediction; (iv) the rate of "Absolute-True" is 60.92%, the average ratio of the perfectly correct hits over the total prediction events and (v) the rate of "Absolute-False" or 'Hamming-Loss' (Chou, 2013) is 13.40%, the average ratio of the completely wrong hits over the total prediction events.

Since **iPTM-mLys** is the first multi-label predictor ever developed for identifying multiple PTM sites in proteins, it is hard to demonstrate its power by comparison with its counterparts for exactly the same purpose. Nevertheless, we can show its power by a comparison with some multi-label predictors in other areas via the following analysis and discussion.

In **Equations 10** and **12**, the first four metrics are completely opposite to the last one. For the former, the higher the rate is, the better the multi-label predictor's performance will be; for the latter, the lower the rate is, the better its performance will be.

Among the five metrics in **Equation 10**, the most strict and harsh one is the 'Absolute-True'. To our best knowledge, very few multi-label predictors in biology could reach over 50% for the absolute true rate. For example, **iLoc-Animal,** a very powerful multi-label classifier for predicting subcellular localization of animal proteins (Lin *et al.*, 2013), its reported absolute-true rate was 45.62%. Also, for **iAMP-2L**, a powerful two-level multi-label classifier for identifying antimicrobial peptides and their functional types (Xiao *et al.*, 2013), its reported absolute-true rate was 43.05%. None of them could reach even 50%; in contrast, the absolute-true rate achieved by iPTM-mLys can reach over 60%, as shown in **Equation 12**.

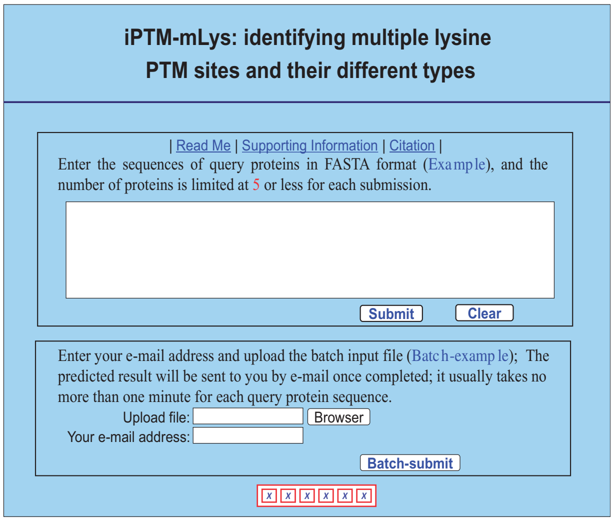
Also, among the same five metrics, the most important is the 'Accuracy'. According to the reports for **iLoc-Animal** (Lin *et al.*, 2013) and **iAMP-2L** (Xiao *et al.*, 2013), their accuracy rates were 62.88% and 66.87%, respectively. In contrast, the accuracy rate achieved by **iPTM-mLys** is 68.37%.

4 Web-server and user guide

To enhance the value of its practical applications, the web-server for **iPTM-mLys** has been established. Furthermore, to maximize the convenience of most experimental scientists, a step-by-step guide is provided below.

Step 1. Opening the web-server at http://www.jci-bioinfo.cn/iPTM-mLys, you will see the top page of **iPTM-mLys** on your computer screen, as shown in Figure 2. Click on the Read Me button to see a brief introduction about the predictor.

Step 2. Either type or copy/paste the query protein sequences into the input box at the center of **Figure 2.** 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.



iPTM-mLys: identifying multiple lysine PTM sites and their different types

| Read Me | Supporting Information | Citation |

Enter the sequences of query proteins in FASTA format (Example), and the number of proteins is limited at 5 or less for each submission.

Enter your e-mail address and upload the batch input file (Batch-example); The predicted result will be sent to you by e-mail once completed; it usually takes no more than one minute for each query protein sequence.

Fig. 2. A semi-screenshot to show the top-page of the **iPTM-mLys** web-server at http://www.jci-bioinfo.cn/iPTM-mLys

Table 2. Comparison between the predicted and experimental results on protein Q16778

	Predicted result				Experimental result			
Sites	Ace	Cro	Met	Suc	Ace	Cro	Met	Suc
6	Yes	Yes	No	No	Yes	Yes	No	No
12	Yes	Yes	No	No	Yes	Yes	No	No
13	Yes	No	No	No	Yes	Yes	No	No
16	Yes	Yes	No	No	Yes	Yes	No	No
17	Yes	Yes	No	No	Yes	Yes	No	No
21	Yes	Yes	No	No	Yes	Yes	No	No
24	Yes	Yes	No	No	Yes	Yes	No	No
25	Yes	No	No	No	No	No	No	No
28	No	No	No	No	No	No	No	No
29	No	No	No	No	No	No	No	No
31	No	No	No	No	No	No	No	No
35	No	Yes	No	No	No	Yes	No	No
44	Yes	No	No	No	No	No	No	No
47	No	No	Yes	No	No	No	Yes	No
58	Yes	No	Yes	No	No	No	Yes	No
86	Yes	No	No	No	Yes	No	Yes	No
109	Yes	No	No	No	No	No	Yes	No
117	Yes	No	No	No	No	No	No	No
121	Yes	No	No	No	Yes	Yes	No	No
126	Yes	No	No	No	No	No	No	No

Ace, acetylation; Cro, crotonylation; Met, methylation; Suc, succinylation.

Step 3. Click on the Submit button to see the predicted result. For instance, if you use the two query protein sequences as an input, after 30 seconds or so since clicking the submit button, you'll see the following results popped on the screen. (1) Sequence-1 (Q16778) contains 20 K residues, of which the residues at the sequence position sites '6', '12', '16', '17', '21' and '24' can be of both 'Acetylation' and 'Crotonylation'; residue '58' can be of both 'Acetylation' and 'Methylation'; residues '35' and '47' can be of both 'Crotonylation' and 'Methylation'; residues '13', '25', '44', '86', '109', '117', '121' and '126' can be of only 'Acetylation'; while residues '28', '29' and '31' can be none of the four PTM. (2) Sequence-2 contains invalid character(s): 'Check your input!' For facilitating comparison, both the predicted and experimental results for protein Q16778 are listed in Table 2. Substituting the consistent and inconsistent scores between the prediction and observation into **Equation 10**, we have the rates of aiming $=67.50\%$, coverage $=65.00\%$, accuracy $=62.50\%$, absolute-true $=55.00\%$ and absolute-false $=15.00\%$, quite similar to the rates obtained by the cross-validation tests as given in Equation 12.

Step 4. As shown on the lower panel of Figure 2, 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.

Step 5. Click on the Citation button to find the relevant papers that document the detailed development and algorithm of **iPTM-mLys.**

Step 6. Click the Supporting Information button to download the benchmark dataset used to train and test the current predictor.

Note: To obtain the predicted result with the anticipated success rate, the entire sequence of the query protein rather than its fragment should be used as an input.

5 Conclusion

There are many existing computational predictors for identifying the PTM sites in proteins; all of them are for the single-label PTM systems but not for the multi-label ones. The **iPTM-mLys** presented in this paper represents the first web-server ever established that can be use to deal with both single- and multi-label PTM systems. It has not escaped our notice that, similar to the current multi-label K-PTM, the other PTM systems such as C-PTM, R-PTM and S-PTM do also have their corresponding multi-label PTM sites at Cys, Arg and Ser residues, respectively. Likewise, the approach and formulations proposed in this article can be used to analyze them as well.

To maximize the users' convenience, a step-by-step guide has been provided, by which users can easily get their desired results without the need to go through the complicated mathematical equations. Although the current iPTM-mLys predictor is limited to analyze the multi-label systems of four different K-PTM types, with more experimental data available in future, we will update it with a new version to cover more types as well, such as biotinylation, butyrylation, propionylation and ubiquitination. By that time, an announcement will be given in the website http://www.jci-bioinfo.cn/iPTM-mLys.

Acknowledgments

The authors wish to thank the three anonymous reviewers, whose constructive comments are very helpful for strengthening the presentation of this paper.

Funding

This work was partially supported by the National Nature Science Foundation of China (No. 61261027, 31260273, 61300139, 31560316), the Natural Science Foundation of Jiangxi Province, China (No. 20142BAB207013), the Scientific Research plan of the Department of Education of JiangXi Province(GJJ14640), the scholarship award of the State Scholarship Fund (No. 201508360047). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Conflict of Interest: none declared.

References

Ahmad,S. *et al.* (2015) Identification of heat shock protein families and J-protein types by incorporating dipeptide composition into Chou's general PseAAC. *Comput. Methods Program. Biomed.*, **122**, 165–174.

Ahmad,K. *et al.* (2016) Prediction of protein submitochondrial locations by incorporating dipeptide composition into chou's general pseudo amino acid composition. *J. Membr. Biol*, 10.1007/s00232-00015-09868-00238.

Breiman,L. (2001) Random forests. *Machine Learn.*, **45**, 5–32.

Cao,D.S. *et al.* (2013) propy: a tool to generate various modes of Chou's PseAAC. *Bioinformatics*, **29**, 960–962.

Chen,H. *et al.* (2006) MeMo: a web tool for prediction of protein methylation modifications. *Nucleic Acids Res.*, **34**, W249–W253.

Chen,W. *et al.* (2013) iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. *Nucleic Acids Res.*, **41**, e68.

Chen,W. *et al.* (2014) PseKNC: a flexible web-server for generating pseudo K-tuple nucleotide composition. *Anal. Biochem.*, **456**, 53–60.

Chen,W. *et al.* (2015a) Pseudo nucleotide composition or PseKNC: an effective formulation for analyzing genomic sequences. *Mol BioSyst*, **11**, 2620–2634.

Chen,W. *et al.* (2015b) PseKNC-General: a cross-platform package for generating various modes of pseudo nucleotide compositions. *Bioinformatics*, **31**, 119–120.

Chen,W. *et al.* (2016a) iACP: a sequence-based tool for identifying anticancer peptides. *Oncotarget*, 7, 16895–16909.

Chen,W. *et al.* (2016b) iRNA-PseU: Identifying RNA pseudouridine sites, *Molecular Therapy - Nucleic Acids*, **5**, e332.

Chou,K.C. (1996) Review: Prediction of human immunodeficiency virus protease cleavage sites in proteins. *Anal. Biochem.*, **233**, 1–14.

Chou,K.C. (2001a) Prediction of protein cellular attributes using pseudo amino acid composition. *Proteins*, **43**, 246–255.

Chou,K.C. (2001b) Prediction of signal peptides using scaled window. *Peptides*, **22**, 1973–1979.

Chou,K.C. (2005) Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinformatics*, **21**, 10–19.

Chou,K.C. (2011) Some remarks on protein attribute prediction and pseudo amino acid composition (50th Anniversary Year Review). *J. Theor. Biol.*, **273**, 236–247.

Chou,K.C. (2013) Some remarks on predicting multi-label attributes in molecular biosystems. *Mol. Biosyst.*, **9**, 1092–1100.

Chou,K.C. (2015a) Impacts of bioinformatics to medicinal chemistry. *Med. Chem.*, **11**, 218–234.

Chou,K.C. (2015b) An unprecedented revolution in medicinal science. *Proc. MOL2NET (International Conference on Multidisciplinary Sciences)*, **1**, 1–10. doi:10.3390/MOL2NET-1-b040.

Chou,K.C. and Cai,Y.D. (2005) Prediction of membrane protein types by incorporating amphipathic effects. *J. Chem. Inf. Model.*, **45**, 407–413.

Chou,K.C. and Shen,H.B. (2007a) Review: Recent progresses in protein subcellular location prediction. *Anal. Biochem.*, **370**, 1–16.

Chou,K.C. and Shen,H.B. (2007b) Signal-CF: a subsite-coupled and window-fusing approach for predicting signal peptides. *Biochem. Biophys. Res. Comm.*, **357**, 633–640.

Chou,K.C. and Zhang,C.T. (1995) Review: prediction of protein structural classes. *Crit. Rev. Biochem. Mol. Biol.*, **30**, 275–349.

Chou,K.C. *et al.* (2011) iLoc-Euk: a multi-label classifier for predicting the subcellular localization of singleplex and multiplex eukaryotic proteins. *PLoS One*, **6**, e18258.

Chou,K.C. *et al.* (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.

Dehzangi,A. *et al.* (2015) Gram-positive and Gram-negative protein subcellular localization by incorporating evolutionary-based descriptors into Chou's general PseAAC. *J. Theor. Biol.*, **364**, 284–294.

Du,P. *et al.* (2012) PseAAC-Builder: A cross-platform stand-alone program for generating various special Chou's pseudo-amino acid compositions. *Anal. Biochem.*, **425**, 117–119.

Du,P. *et al.* (2014) PseAAC-General: Fast building various modes of general form of Chou's pseudo-amino acid composition for large-scale protein datasets. *Int. J. Mol. Sci.*, **15**, 3495–3506.

Huang,C. and Yuan,J.Q. (2013) 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**, 327–334.

Jia,J. *et al.* (2015a) 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**, 47–56.

Jia,J. *et al.* (2015b) 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.* doi:10.1080/07391102.2015.1095116.

Jia,J. *et al.* (2016a) iCar-PseCp: identify carbonylation sites in proteins by Monto Carlo sampling and incorporating sequence coupled effects into general PseAAC. *Oncotarget*, **7**, 34558–34570

Jia,J. *et al.* (2016b) iPPBS-Opt: a sequence-based Ensemble classifier for identifying protein-protein binding sites by optimizing imbalanced training datasets. *Molecules*, **21**, 95.

Jia,J. *et al.* (2016c) iSuc-PseOpt: identifying lysine succinylation sites in proteins by incorporating sequence-coupling effects into pseudo components and optimizing imbalanced training dataset. *Anal. Biochem*, **497**, 48–56.

Jia,J. *et al.* (2016d) pSuc-Lys: predict lysine succinylation sites in proteins with PseAAC and ensemble random forest approach. *J. Theor. Biol.*, **394**, 223–230.

Jia,J. *et al.* (2016e) pSumo-CD: Predicting sumoylation sites in proteins with covariance discriminant algorithm by incorporating sequence-coupled effects into general PseAAC. *Bioinformatics*, doi: 10.1093/bioinformatics/btw387.

Kabir,M. and Hayat,M. (2016) iRSpot-GAEnsC: identifing recombination spots via ensemble classifier and extending the concept of Chou's PseAAC to formulate DNA samples. *Mol. Genet. Genomics*, **291**, 285–296.

Kandaswamy,K.K. *et al.* (2011) AFP-Pred: a random forest approach for predicting antifreeze proteins from sequence-derived properties. *J. Theor. Biol.*, **270**, 56–62.

Khan,Z.U. *et al.* (2015) Discrimination of acidic and alkaline enzyme using Chou's pseudo amino acid composition in conjunction with probabilistic neural network model. *J. Theor. Biol.*, **365**, 197–203.

Kumar,R. *et al.* (2015) Prediction of beta-lactamase and its class by Chou's pseudo-amino acid composition and support vector machine. *J. Theor. Biol.*, **365**, 96–103.

Lin,H. *et al.* (2014) iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition. *Nucleic Acids Res.*, **42**, 12961–12972.

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.

Lin,W.Z. *et al.* (2011) iDNA-Prot: identification of DNA binding proteins using random forest with grey model. *PLoS ONE*, **6**, e24756.

Lin,W.Z. *et al.* (2013) iLoc-Animal: a multi-label learning classifier for predicting subcellular localization of animal proteins. *Mol. BioSyst.*, **9**, 634–644.

Liu,B. *et al.* (2015a) repDNA: a Python package to generate various modes of feature vectors for DNA sequences by incorporating user-defined physicochemical properties and sequence-order effects. *Bioinformatics*, **31**, 1307–1309.

Liu,B. *et al.* (2015b) Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. *Nucleic Acids Res.*, **43**, W65–W71.

Liu,B. *et al.* (2015c) PseDNA-Pro: DNA-binding protein identification by combining Chou's PseAAC and physicochemical distance transformation. *Mol. Informatics*, **34**, 8–17.

Liu,B. *et al.* (2016a) iMiRNA-PseDPC: microRNA precursor identification with a pseudo distance-pair composition approach. *J. Biomol. Struct. Dyn.*, **34**, 223–235.

Liu,B. *et al.* (2016b) iEnhancer-2L: a two-layer predictor for identifying enhancers and their strength by pseudo k-tuple nucleotide composition. *Bioinformatics*, **32**, 362–389.

Liu,B. *et al.* (2016c) iDHS-EL: Identifying DNase I hypersensi-tivesites by fusing three different modes of pseudo nucleotide composition into an en-semble learning framework. *Bioinformatics*. doi:10.1093/bioinformatics/btw186.

Liu,Z. *et al.* (2016d) pRNAm-PC: Predicting N-methyladenosine sites in RNA sequences via physical-chemical properties. *Anal. Biochem.*, **497**, 60–67.

Mondal,S. and Pai,P.P. (2014) Chou's pseudo amino acid composition improves sequence-based antifreeze protein prediction. *J. Theor. Biol.*, **356**, 30–35.

Nanni,L. *et al.* (2014) Prediction of protein structure classes by incorporating different protein descriptors into general Chou's pseudo amino acid composition. *J. Theor. Biol.*, **360**, 109–116.

Pugalenthi,G. *et al.* (2012) RSARF: prediction of residue solvent accessibility from protein sequence using random forest method. *Protein Peptide Letters*, **19**, 50–56.

Qiu,W.R. *et al.* (2014) iMethyl-PseAAC: Identification of Protein Methylation Sites via a Pseudo Amino Acid Composition Approach. *Biomed Res Int (BMRI)*, **2014**, 947416.

Qiu,W.R. *et al.* (2015) iUbiq-Lys: Prediction of lysine ubiquitination sites in proteins by extracting sequence evolution information via a grey system model. *Journal of Biomolecular Structure and Dynamics (JBSD)*, **33**, 1731–1742.

Qiu,W.R. *et al.* (2016a) iPhos-PseEvo: identifying human phosphorylated proteins by incorporating evolutionary information into general PseAAC via grey system theory. *Mol. Informatics*. doi:10.1002/minf.201600010.

Qiu,W.R. *et al*. (2016b) iPhos-PseEn: identifying phosphorylation sites in proteins by fusing different pseudo components into an ensemble classifier. *Oncotarget*. doi:10.18632/oncotarget.9987.

Qiu,W.R. *et al*. (2016c) iHyd-PseCp: Identify hydroxyproline and hydroxylysine in proteins by incorporating sequence-coupled effects into general PseAAC. *Oncotarget*, **7**, 44310–44321.

Shao,J. *et al*. (2009) Computational identification of protein methylation sites through bi-profile Bayes feature extraction. *PLoS One*, **4**, e4920.

Shen,H.B. and Chou,K.C. (2007a) Signal-3L: a 3-layer approach for predicting signal peptide. *Biochem. Biophys. Res. Comm.*, **363**, 297–303.

Shen,H.B. and Chou,K.C. (2007b) Virus-PLoc: A fusion classifier for predicting the subcellular localization of viral proteins within host and virus-infected cells. *Biopolymers*, **85**, 233–240.

Tang,H. *et al*. (2016) Identification of immunoglobulins using Chou's pseudo amino acid composition with feature selection technique. *Mol. Biosyst.*, **12**, 1269–1275.

Wang,X. *et al*. (2015) MultiP-SChlo: multi-label protein subchloroplast localization prediction with Chou's pseudo amino acid composition and a novel multi-label classifier. *Bioinformatics*, **31**, 2639–2645.

Wu,Z.C. *et al*. (2011) iLoc-Plant: a multi-label classifier for predicting the subcellular localization of plant proteins with both single and multiple sites. *Mol. BioSyst.*, **7**, 3287–3297.

Wu,Z.C. *et al*. (2012) iLoc-Gpos: A Multi-Layer Classifier for Predicting the Subcellular Localization of Singleplex and Multiplex Gram-Positive Bacterial Proteins. *Protein Peptide Lett.*, **19**, 4–14.

Xiao,X. *et al*. (2011) 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**, 42–51.

Xiao,X. *et al*. (2013) iAMP-2L: A two-level multi-label classifier for identifying antimicrobial peptides and their functional types. *Anal. Biochem.*, **436**, 168–177.

Xiao,X. *et al*. (2016) iROS-gPseKNC: predicting replication origin sites in DNA by incorporating dinucleotide position-specific propensity into general pseudo nucleotide composition. *Oncotarget*, **7**, 34180–34189.

Xu,Y. and Chou,K.C. (2016) Recent progress in predicting posttranslational modification sites in proteins. *Curr Top Med Chem*, **16**, 591–603.

Xu,Y. *et al*. (2013a) 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.

Xu,Y. *et al*. (2013b) iSNO-AAPair: incorporating amino acid pairwise coupling into PseAAC for predicting cysteine S-nitrosylation sites in proteins. *PeerJ*, **1**, e171.

Xu,Y. *et al*. (2014a) iHyd-PseAAC: Predicting hydroxyproline and hydroxylysine in proteins by incorporating dipeptide position-specific propensity into pseudo amino acid composition, *Int. J. Mol. Sci.*, **15**, 7594–7610.

Xu,Y. *et al*. (2014b) iNitro-Tyr: Prediction of nitrotyrosine sites in proteins with general pseudo amino acid composition. *PLoS One*, **9**, e105018.

Zhong,W.Z. and Zhou,S.F. (2014) Molecular science for drug development and biomedicine. *Int. J. Mol. Sci.*, **15**, 20072–20078.

Zhou,G.P. (1998) An intriguing controversy over protein structural class prediction. *J. Protein Chem*, **17**, 729–738.

Zhou,G.P. (2015) Current progress in structural bioinformatics of protein-biomolecule interactions. *Med. Chem.*, **11**, 216-216.

Zhou,G.P. and Doctor,K. (2003) Subcellular location prediction of apoptosis proteins. *Proteins*, **50**, 44–48.,

Zhou,G.P. and Zhong,W.Z. (2016) Perspectives in medicinal chemistry. *Curr. Topics Med. Chem.*, **16**, 381–382.