Bioinformatics, 32(20), 2016, 3133–3141
doi: 10.1093/bioinformatics/btw387
Advance Access Publication Date: 26 June 2016
Original Paper

pSumo-CD: predicting sumoylation sites in proteins with covariance discriminant algorithm by incorporating sequence-coupled effects into general PseAAC

Jianhua Jia[1,2,*], Liuxia Zhang[1], Zi Liu[3], Xuan Xiao[1,2,*] and Kuo-Chen Chou[2,4,5,*]

[1]Computer Department, Jing-De-Zhen Ceramic Institute, Jing-De-Zhen 333403, China, [2]Gordon Life Science Institute, Boston, MA 02478, USA, [3]School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China, [4]Center of Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz University, Jeddah 21589, Saudi Arabia and [5]Center of Bioinformatics, School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu 610054, China

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on May 12, 2016; revised on June 12, 2016; accepted on June 15, 2016

Abstract

Motivation: Sumoylation is a post-translational modification (PTM) process, in which small ubiquitin-related modifier (SUMO) is attaching by covalent bonds to substrate protein. It is critical to many different biological processes such as replicating genome, expressing gene, localizing and stabilizing proteins; unfortunately, it is also involved with many major disorders including Alzheimer's and Parkinson's diseases. Therefore, for both basic research and drug development, it is important to identify the sumoylation sites in proteins.

Results: To address such a problem, we developed a predictor called pSumo-CD by incorporating the sequence-coupled information into the general pseudo-amino acid composition (PseAAC) and introducing the covariance discriminant (CD) algorithm, in which a bias-adjustment term, which has the function to automatically adjust the errors caused by the bias due to the imbalance of training data, had been incorporated. Rigorous cross-validations indicated that the new predictor remarkably outperformed the existing state-of-the-art prediction method for the same purpose.

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

Contact: jjia@gordonlifescience.org, xxiao@gordonlifescience.org or kcchou@gordonlifescience.org

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

1 Introduction

In vivo, protein post-translational modification (PTM or PTLM) is an important biological mechanism for expanding the genetic code as well as regulating cellular physiology. Small ubiquitin-like modifier (SUMO) proteins are a type of PTM that play an important role in subcellular transport, transcription, DNA repair and signal transduction. Recent researches have indicated that sumoylation can promote the binding ability of proteins, and that some proteins, such as claspin, whose binding function is sumoylation-dependent. Many diseases and disorders, such as Alzheimer's and Parkinson's diseases, have been found to be closely related with sumoylation. Therefore, identification of sumoylation sites in proteins is important not only for in-depth understanding many important biological processes but also for developing effective drugs.

Identification of sumoylation sites with biological and chemical approaches is more laborious and slow; particularly the sumoylation is reversible and instable. With the avalanche of protein sequences generated in the post-genomic era, it is highly demanded to develop computational methods as a complimentary tool to the pure experimental methods.

Actually, considerable efforts have been made in this regard, including the method called SUMOsp developed by Xue *et al.* (2006), the method SUMOsp2.0 by Ren *et al.* (2009) and the method GPS-SUMO by Zhao *et al.* (2014); all of them were developed using the group-based phosphorylation scoring algorithm. Meanwhile, based on support vector machine (SVM), the methods SUMOPre and SUMOhydro were developed by Xu *et al.* (2008) and Chen *et al.* (2012b), respectively. Very recently, based on the linear discriminant analysis, Xu *et al.* (2016) proposed the method SUMO_LDA. Each of these methods did make contribution in stimulating the development of such an important area. Particularly, the most recent method SUMO_LDA (Xu *et al.*, 2016), which was established by combining three different types of sequence features into the general pseudo-amino acid composition (PseAAC) (Chou, 2011), achieved quite decent success rates. In considering, however, the topic's importance and also the urgency of demanding more powerful computational tools in this area, further efforts aiming at prediction of protein sumoylation sites are definitely needed.

The present study was initiated in an attempt to develop a new and more powerful predictor by (i) incorporating the vectorized sequence-coupling model into the general form of pseudo amino acid composition and (ii) installing the covariance discriminant operation engine into the prediction system. Rigorous cross-validations indicated that the new predictor significantly outperformed the existing state-of-the-art predictor (Xu *et al.*, 2016) in both overall accuracy ($>10\%$) and stability ($>16\%$).

According to the Chou's 5-step rule (Chou, 2011) and concurred by many investigators in a series of recent publications (Chen *et al.*, 2016a, c; Jia *et al.*, 2016a, b, c, d; Liu *et al.*, 2016a, b, d, e; Qiu *et al.*, 2016a; 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) a good benchmark dataset used to train or test the new model; (ii) an effective mathematical formulation to represent the statistical samples concerned; (iii) a powerful algorithm to operate the calculation; (iv) a compelling demonstration to show its prediction quality being improved over the existing counterparts and (v) the prediction method should be with a web-server accessible to public. Below, we are to address these points one by one, making them very natural in logic and crystal clear in description.

2 Materials and methods

2.1 Benchmark datasets

The benchmark dataset used in this study was derived from the same 510 proteins as used by Xu *et al.* (2016). The complete amino acid sequences of these proteins can be obtained from UniProt (Apweiler *et al.*, 2004). In the last decade or so, various consensus motifs for SUMO have been suggested. Regardless of their many differences in details, there is one thing in common, i.e. they all contain the amino acid residue Lys or K. To make the description mathematically more rigorous and clear, the Chou's scheme (Chou, 2001c) was adopted to formulate peptide samples, as done recently by many authors in studying the nitrotyrosine sites (Xu *et al.*, 2014), methylation sites (Qiu *et al.*, 2014) and protein–protein-binding sites (Jia *et al.*, 2015b). According to Chou's scheme, a potential hydroxylation site-containing peptide sample can be generally expressed by

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

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 $(2\xi+1)$-tuple peptide sample $\mathbf{P}_{\xi}(\mathbb{K})$ can be further classified into the following two categories:

$$\mathbf{P}_{\xi}(\mathbb{K}) \in \begin{cases} \mathbf{P}_{\xi}^{+}(\mathbb{K}), & \text{if its center belongs to sumoylation site} \\ \mathbf{P}_{\xi}^{-}(\mathbb{K}), & \text{otherwise} \end{cases} \quad (2)$$

where $\mathbf{P}_{\xi}^{+}(\mathbb{K})$ denotes a true sumoylation segment, $\mathbf{P}_{\xi}^{-}(\mathbb{K})$ a false sumoylation segment and the symbol $\in$ means 'a member of' in the set theory.

In literature the benchmark dataset usually consists of a training dataset and a testing dataset: the former is used for training a model, while the latter for testing the model. But as pointed out in a comprehensive review (Chou and Shen, 2007a), there is no need to artificially separate a benchmark dataset into the two parts if the prediction model is examined by the jackknife test or sub-sampling (K-fold) cross-validation since the outcome thus obtained is actually from a combination of many different independent dataset tests. Thus, the benchmark dataset for the current study can be formulated as

$$\mathbb{S}_{\xi} = \mathbb{S}_{\xi}^{+} \cup \mathbb{S}_{\xi}^{-} \quad (3)$$

where the positive subset $\mathbb{S}_{\xi}^{+}$ contains only the true sumoylation samples $\mathbf{P}_{\xi}^{+}(\mathbb{K})$, and the negative subset $\mathbb{S}_{\xi}^{-}$ contains the false sumoylation segments $\mathbf{P}_{\xi}^{-}(\mathbb{K})$ only [see Equation (2)]; while $\cup$ denotes the symbol of 'union' in the set theory.

Since the length of peptide sample $\mathbf{P}_{\xi}(\mathbb{K})$ is $2\xi+1$ [see Equation (1)], the benchmark dataset with different ξ value will contain peptide segments with different number of amino acid residues. In the study carried out recently by Xu *et al.* (2016), however, they selected

$$\xi = 10; \text{i.e. the length of peptide sample is 21} \quad (4)$$

In order to facilitate comparison with their method, in this study, let us also assign the value 10 for the window parameter ξ. Thus, Equations (1) and (3) can be reduced to

$$\begin{cases} \mathrm{P}(\mathbb{K}) = \mathrm{R}_{-10}\mathrm{R}_{-9}\cdots\mathrm{R}_{-2}\mathrm{R}_{-1}\mathbb{K}\mathrm{R}_{+1}\mathrm{R}_{+2}\cdots\mathrm{R}_{+9}\mathrm{R}_{+10} \\ \mathbb{S} = \mathbb{S}^{+} \cup \mathbb{S}^{-} \end{cases} \quad (5)$$

The detailed procedures in constructing the benchmark dataset $\mathbb{S}$ are as follows. (i) As done in (Chou, 2001c), slide the 21-tuple peptide window along each of the aforementioned 510 protein sequences, and collected were only those peptide segments that have K (Lys or lysine) at the center [see Equation (1)]. (ii) If the upstream or downstream in a protein sequence was less than 10 or greater than $(L-10)$ where L is the length of the protein sequence concerned, the lacking amino acid was filled with a dummy residue X. (iii) The peptide segment samples thus obtained were put into the positive subset if their centers have been experimentally annotated as the sumoylation sites; otherwise, into the negative subset. (iv) The peptide samples thus obtained were subject to a screening procedure to window those that had $\geq 40\%$ pairwise sequence identity to any other in a same subset. By following the above procedures, we finally obtained 755 positive samples, 9944 negative samples. For readers' convenience, their detailed sequences are given in Supplementary Materials S1 and S2, respectively.

2.2 Incorporating sequence-coupled information 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 order information. This is because all the existing machine-learning algorithms can only handle vector but not sequence samples, as elaborated in Chou (2015).

To address this problem, the pseudo-amino acid composition (Chou, 2001a, b) 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 (see, e.g. 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 as well as a long list of references cited in Chen *et al.* 2015b; Du *et al.* 2014 and many biomedicine and drug development areas (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 [see Equations (9)–(10) of Chou, 2011], 'Gene Ontology' mode [see Equations (11)–(12) of Chou, 2011] and 'Sequential Evolution' or 'PSSM' mode [see Equations (13)–(14) of Chou, 2011]. Inspired by the successes of using PseAAC to deal with protein/peptide sequences, three web-servers (Chen *et al.*, 2014c, 2015c; Liu *et al.*, 2015c) 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.*, 2015d) 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 in Equation (5) can be formulated as

$$\mathbf{P}(\mathbb{K}) = \mathbf{P}^{+}(\mathbb{K}) - \mathbf{P}^{-}(\mathbb{K}) \tag{6}$$

where

$$\mathbf{P}^{+}(\mathbb{K}) = \begin{bmatrix} p^{+}_{-10}(\mathrm{R}_{-10}|\mathrm{R}_{-9}) \\ p^{+}_{-9}(\mathrm{R}_{-9}|\mathrm{R}_{-8}) \\ \vdots \\ p^{+}_{-2}(\mathrm{R}_{-2}|\mathrm{R}_{-1}) \\ p^{+}_{-1}(\mathrm{R}_{-1}) \\ p^{+}_{+1}(\mathrm{R}_{+1}) \\ p^{+}_{+2}(\mathrm{R}_{+2}|\mathrm{R}_{+1}) \\ \vdots \\ p^{+}_{+9}(\mathrm{R}_{+9}|\mathrm{R}_{+8}) \\ p^{+}_{+10}(\mathrm{R}_{+10}|\mathrm{R}_{+9}) \end{bmatrix} \tag{7}$$

and

$$\mathbf{P}^{-}(\mathbb{K}) = \begin{bmatrix} p^{-}_{-10}(\mathrm{R}_{-10}|\mathrm{R}_{-9}) \\ p^{-}_{-9}(\mathrm{R}_{-9}|\mathrm{R}_{-8}) \\ \vdots \\ p^{-}_{-2}(\mathrm{R}_{-2}|\mathrm{R}_{-1}) \\ p^{-}_{-1}(\mathrm{R}_{-1}) \\ p^{-}_{+1}(\mathrm{R}_{+1}) \\ p^{-}_{+2}(\mathrm{R}_{+2}|\mathrm{R}_{+1}) \\ \vdots \\ p^{-}_{+9}(\mathrm{R}_{+9}|\mathrm{R}_{+8}) \\ p^{-}_{+10}(\mathrm{R}_{+10}|\mathrm{R}_{+9}) \end{bmatrix} \tag{8}$$

In the above Equation (7), $p^{+}_{-10}(\mathrm{R}_{-10}|\mathrm{R}_{-9})$ is the conditional probability of amino acid R_{-10} occurring at the left 1st position [see Equation (5)] given that its closest right neighbor is R_{-9}; $p^{+}_{-9}(\mathrm{R}_{-9}|\mathrm{R}_{-8})$ is the conditional probability of amino acid R_{-9} occurring at the left 2nd position given that its closest right neighbor is R_{-8}; and so forth. Note that in Equation (7), only $p^{+}_{-1}(\mathrm{R}_{-1})$ and $p^{+}_{+1}(\mathrm{R}_{+1})$ are of non-conditional probability since the right neighbor of R_{-1} and the left neighbor of R_{+1} are always K or Lys. All these probability values can be easily derived from the positive samples in Supplementary Material S1, as done in Chou (1996). Likewise, the components in Equation (8) are the same as those in Equation (7) except for that they are derived from the negative samples in Supplementary Material S2.

2.3 Generalized covariance discriminant algorithm

The covariant discriminant (CD) algorithm has been widely used in bioinformatics, such as predicting protein structural classes (Chou, 1999; Chou *et al.*, 1998; Chou and Maggiora, 1998; Chou and Zhang, 1994; Liu and Chou, 1998; Zhou, 1998; Zhou and Assa-Munt, 2001), protein subcellular localization (Chou, 2000a; Chou and Elrod, 1998; Chou and Elrod, 1999b; Zhou and Doctor, 2003), membrane protein type (Chou and Elrod, 1999a), GPCR types (Chou, 2005a; Chou and Elrod, 2002; Elrod and Chou, 2002), enzyme family classes (Chou and Elrod, 2003) and nucleosome positioning (Chen *et al.*, 2012a).

To make the feature vector as defined in Equation (6) easier to be expressed in the CD algorithm, let us define

$$\mathbb{P} = \mathbf{P}(\mathbb{K}) = [\Psi_1 \quad \Psi_2 \quad \cdots \quad \Psi_{10} \quad \Psi_{11} \quad \cdots \quad \Psi_{20}\ ]^{\mathrm{T}} \tag{9}$$

where T is the transposing operator, and

$$\begin{cases} \Psi_1 = \ p^{+}_{-10}(\mathrm{R}_{-10}\mathrm{R}_{-9}) - p^{-}_{-10}(\mathrm{R}_{-10}\mathrm{R}_{-9}) \\ \Psi_2 = p^{+}_{-9}(\mathrm{R}_{-9}\mathrm{R}_{-8}) - p^{-}_{-9}(\mathrm{R}_{-9}\mathrm{R}_{-8}) \\ \quad\vdots \\ \Psi_{10} = p^{+}_{-1}(\mathrm{R}_{-1}) - p^{-}_{-1}(\mathrm{R}_{-1}) \\ \Psi_{11} = p^{+}_{+1}(\mathrm{R}_{+1}) - p^{-}_{+1}(\mathrm{R}_{+1}) \\ \quad\vdots \\ \Psi_{20} = p^{+}_{+10}(\mathrm{R}_{+10}\mathrm{R}_{+9}) - p^{-}_{+10}(\mathrm{R}_{+10}\mathrm{R}_{+9}) \end{cases} \tag{10}$$

Suppose the standard feature vectors for the peptide samples in $\mathbb{S}^+$ and $\mathbb{S}^-$ are, respectively, expressed by

$$\begin{cases} \overline{\mathbb{P}^+} = \begin{bmatrix} \overline{\Psi_1^+} & \overline{\Psi_2^+} & \cdots & \overline{\Psi_{10}^+} & \overline{\Psi_{11}^+} & \cdots & \overline{\Psi_{20}^+} \end{bmatrix}^{\mathbf{T}} \\ \overline{\mathbb{P}^-} = \begin{bmatrix} \overline{\Psi_1^-} & \overline{\Psi_2^-} & \cdots & \overline{\Psi_{10}^-} & \overline{\Psi_{11}^-} & \cdots & \overline{\Psi_{20}^-} \end{bmatrix}^{\mathbf{T}} \end{cases} \tag{11}$$

where

$$\begin{cases} \overline{\Psi_u^+} = \dfrac{1}{N^+} \sum_{k=1}^{N^+} \Psi_{u,k}^+ \\ \overline{\Psi_u^-} = \dfrac{1}{N^-} \sum_{k=1}^{N^-} \Psi_{u,k}^- \end{cases} \quad (u = 1, 2, \ \cdots, \ 20) \tag{12}$$

where $\Psi_{u,k}^+$ is the uth component of the feature vector for the kth peptide sample in the positive dataset $\mathbb{S}^+$, $\Psi_{u,k}^-$ for the kth peptide sample in the negative dataset $\mathbb{S}^-$, N^+ the total number of peptide samples in $\mathbb{S}^+$ and N^- the total number of peptide samples in $\mathbb{S}^-$.

Thus, whether a query peptide sequence sample $\mathbb{P}$ belongs to the sumoylation site subset $\mathbb{S}^+$ or non-sumoylation site subset $\mathbb{S}^-$ will be judged by

$$\mathrm{Sgn}(\delta) = \mathbf{arg\ min}_\delta \left\{ \mathbb{F}(\mathbb{P}, \overline{\mathbb{P}^\delta} \right\}, \ (\delta = + \text{ or } -) \tag{13}$$

where $\mathrm{Sgn}(\delta)$ is the argument of δ that minimize $\mathbb{F}(\mathbb{P}, \overline{\mathbb{P}^\delta})$, which according to the CD algorithm is defined as (Chou and Maggiora, 1998)

$$\mathbb{F}(\mathbb{P}, \overline{\mathbb{P}^\delta}) = D^2(\mathbb{P}, \overline{\mathbb{P}^\delta}) + \ln|\mathbb{C}_\delta| - 2\ln\left[\mathfrak{P}(\mathbb{S}^\delta)\right] + \Lambda \ln(2\pi) \tag{14}$$

In Equation (14), Λ is the dimension of the feature vector that is a constant and hence the term $\Lambda\ln(2\pi)$ can be ignored in this study. $\mathfrak{P}(\mathbb{S}^\delta)$ is the prior probability of subset $\mathbb{S}^\delta$. For the current study, we have

$$\mathfrak{P}(\mathbb{S}^\delta) = \begin{cases} N^+/(N^+ + N^-), & \text{if } \delta = + \\ N^-/(N^+ + N^-), & \text{if } \delta = - \end{cases} \tag{15}$$

where N^+=755 and N^-=9,944 (see Supplementary Materials S1 and S2). $\mathbb{C}_\delta$ in Equation (14) is the covariance matrix of the subset $\mathbb{S}^\delta$, as given by

$$\mathbb{C}_\delta = \begin{bmatrix} c_{1,1}^\delta & c_{1,2}^\delta & \cdots & c_{1,20}^\delta \\ c_{2,1}^\delta & c_{2,2}^\delta & \cdots & c_{2,20}^\delta \\ \vdots & \vdots & \ddots & \vdots \\ c_{20,1}^\delta & c_{20,2}^\delta & \cdots & c_{20,20}^\delta \end{bmatrix} \tag{16}$$

and the elements therein are given by

$$c_{i,j}^\delta = \frac{1}{N^\delta - 1} \sum_{u=1}^{N^\delta} \left(\Psi_{u,i}^\delta - \overline{\Psi_{u,i}^\delta} \right) \left(\Psi_{u,j}^\delta - \overline{\Psi_{u,j}^\delta} \right) \ (i, \ j = 1, 2, \ \cdots, \ 20) \tag{17}$$

Its determinant is denoted by $|\mathbb{C}_\delta|$ and its inverse matrix by $\mathbb{C}_\delta^{-1}$. Thus, the Mahalanobis distance (Mahalanobis, 1936) is given by (Chou, 1995b; Chou and Zhang, 1994)

$$D^2(\mathbb{P}, \overline{\mathbb{P}^\delta}) = \left(\mathbb{P} - \overline{\mathbb{P}^\delta} \right)^{\mathbf{T}} \mathbb{C}_\delta^{-1} \left(\mathbb{P} - \overline{\mathbb{P}^\delta} \right) \tag{18}$$

It is important to point out, however, that the 20 components in Equation (9) are defined by the probability distribution [see Equation (10)] and hence they are constrained by some sort of condition (Chou, 1995b). Therefore, of the 20 components, only 19 are completely independent. As a consequence, the covariance $\mathbb{C}_\delta$ of Equation (16) must be a singular one (Chou and Zhang, 1994), making the $\mathbb{F}$ function [Equation (14)] divergent and meaningless. To avoid such a situation, the dimension-reducing procedure (Chou and Zhang, 1994) was adopted. The concrete procedure is as follows. Instead of 20-D space, a peptide sample is defined in a 19-D space by leaving out one of its 20 components. The remaining 19 components will be completely independent; therefore, the corresponding covariance matrix $\mathbb{C}_\delta$ will no longer be singular. However, a question might be raised: which one of the 20 components should be left out? The answer is any one. Will it yield different outcome by removing a different component? The answer is no. The reason is that, according to Chou's invariance theorem, the outcome of the Mahalanobis distance will remain the same regardless of which one of the components is left out. For the rigorous proof of Chou's invariance theorem, see the Appendix A of Chou (1995b). For a brief introduction about Chou's invariance theorem, see the Wikipedia article at https://en.wikipedia.org/wiki/Chou's_invariance_theorem. Accordingly, in practical calculation, instead of the 20 × 20 matrix version as shown in Equation (16), we should use its 19 × 19 matrix version for $\mathbb{C}_\delta$. Furthermore, the 2nd term in the $\mathbb{F}$ function of Equation (14) can be formulated as (Chou and Elrod, 1999a; Chou and Elrod, 1999b)

$$\ln|\mathbb{C}_\delta| = \begin{cases} \ln(\lambda_1\lambda_2 \cdots \lambda_{19}\lambda_{20}) \text{ for the } 20{\times}20 \text{ version} \\ \ln(\lambda_2\lambda_3\lambda_4 \cdots \lambda_{19}) \text{ for the } 19{\times}19 \text{ version} \end{cases} \tag{19}$$

where $0 = \lambda_1 < \lambda_2 < \lambda_3 < \cdots$ are the eigenvalue values of the determinant $|\mathbb{C}_\delta|$.

Note that in the early studies of using CD algorithm to predict protein structural classes (Chou, 1995b; Chou and Zhang, 1994; Chou, 1995a), only the term of Mahalanobis distance [Equation (18)] was used to calculate the $\mathbb{F}$ function of Equation (14). In a series of subsequent studies on predicting protein structural classes (Chou, 1999; Chou *et al.*, 1998; Chou and Maggiora, 1998; Liu and Chou, 1998; Zhou, 1998; Zhou and Assa-Munt, 2001), protein subcellular localization (Chou, 2000a; Chou and Elrod, 1998; Chou and Elrod, 1999b; Zhou and Doctor, 2003), membrane protein types (Chou and Elrod, 1999a), GPCR types (Chou, 2005a; Chou and Elrod, 2002; Elrod and Chou, 2002), enzyme family classes (Chou and Elrod, 2003) and nucleosome positioning (Chen *et al.*, 2012a), the 2nd term $\ln|\mathbb{C}_\delta|$ in Equation (14) was included as well, and the prediction quality was remarkably improved. In this study, we are to further include the 3rd term $2\ln\left[\mathfrak{P}(\mathbb{S}^\delta)\right]$ in Equation (14). The reason to do so is that in the current case the prior probability of subset $\mathbb{S}^+$ is much smaller than that of subset $\mathbb{S}^-$ [see Equation (15)]. In other words, when the CD algorithm was used to study a system in which the subsets of benchmark dataset was not very skewing, it would be OK by just using the 1st and 2nd terms of Equation (14). But when the benchmark dataset is very unbalanced or highly skewing, the 3rd term must be taken into account. Otherwise, if $\mathbb{S}^+ \ll \mathbb{S}^-$, many positive sample might be incorrectly predicted as negative; and vice versa. To fix such a bias problem when using the operation engines rather than CD algorithm, special treatments, such as dataset optimization (Xiao *et al.*, 2015), Monte Carlo sampling (Jia *et al.*, 2016a) and fusion approach (Jia *et al.*, 2016b; Jia *et al.*, 2016d; Qiu *et al.*, 2016b), were needed. The advantage of using CD algorithm is that it has automatically included the function to deal with the skewing dataset problem, but it was just ignored by the aforementioned investigators and hence missing the contribution from the 3rd term of Equation (14).

The predictor thus established is called ‘pSumo-CD’, where ‘p’ stands for ‘prediction’, ‘Sumo’ for ‘sumoylation site’ and ‘CD’ for ‘covariance discriminant algorithm’.

3 Results and discussion

As pointed out in the Introduction section, one of the keys in establishing a useful predictor is how to properly evaluate its anticipated success rates. To realize this, we need to consider two things: one is what metrics or scales should be used to quantitatively measure its prediction quality; the other is what validation method should be adopted to calculate or derive the metrics values. Below, let us address the two problems

3.1 A set of four metrics

The following four metrics are usually used in literature to measure the quality of binary classification: (i) overall accuracy or Acc; (ii) Mathew’s correlation coefficient or MCC; (iii) sensitivity or Sn and (iv) specificity or Sp (see, e.g. Chen *et al*., 2007). Unfortunately, the conventional formulations for the four are not intuitive and that most experimental scientists feel difficult to understand them, particularly for the one of MCC. Interestingly, using the Chou’s symbols and derivation in studying signal peptides (Chou, 2001b), the aforementioned four metrics can be easily converted into a set of following equations (Chen *et al*., 2013; Xu *et al*., 2013):

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

where N^{+} represents the total number of sumoylation sites investigated, whereas N_{-}^{+} the number of true sumoylation sites incorrectly predicted to be of non-sumoylation site; N^{-} the total number of the non-sumoylation sites investigated, whereas N_{+}^{-} the number of non-sumoylation sites incorrectly predicted to be of sumoylation site.

According to Equation (20), it is crystal clear to see the following. When $N_{-}^{+} = 0$ meaning none of the true sumoylation sites are incorrectly predicted to be of non-sumoylation site, we have the sensitivity $\mathrm{Sn} = 1$. When $N_{-}^{+} = N^{+}$ meaning that all the sumoylation sites are incorrectly predicted to be of non-sumoylation site, we have the sensitivity $\mathrm{Sn} = 0$. Likewise, when $N_{+}^{-} = 0$ meaning none of the non-sumoylation sites are incorrectly predicted to be of sumoylation site, we have the specificity $\mathrm{Sp} = 1$; whereas $N_{+}^{-} = N^{-}$ meaning that all the non-sumoylation sites are incorrectly predicted to be of sumoylation sites, we have the specificity $\mathrm{Sp} = 0$. When $N_{-}^{+} = N_{+}^{-} = 0$ meaning that none of sumoylation sites in the positive dataset and none of the non-sumoylation sites in the negative dataset are incorrectly predicted, we have the overall accuracy $\mathrm{Acc} = 1$ and $\mathrm{MCC} = 1$; when $N_{-}^{+} = N^{+}$ and $N_{+}^{-} = N^{-}$ meaning that all the sumoylation sites in the positive dataset and all the non-sumoylation sites in the negative dataset are incorrectly predicted, we have the overall accuracy $\mathrm{Acc} = 0$ and $\mathrm{MCC} = -1$; whereas when $N_{-}^{+} = N^{+}/2$ and $N_{+}^{-} = N^{-}/2$, we have $\mathrm{Acc} = 0.5$ and $\mathrm{MCC} = 0$ meaning no better than random guess. Therefore, using Equation (20) has made the meanings of sensitivity, specificity, overall accuracy and Mathew’s correlation coefficient much more intuitive and easier-to-understand, particularly for the meaning of MCC, as concurred recently by many investigators (see, e.g. Chen *et al*., 2014a, b, 2015a, 2016a, b, c; Ding *et al*., 2014; Jia *et al*., 2015a, b, 2016a; Liu *et al*., 2015a, b, 2016b, c, d, e; Qiu *et al*., 2016a, b, c; Xiao *et al*., 2015; Xiao *et al*., 2016).

Note that, however, the set of equations defined in Equation (20) is valid only for the single-label systems. For the multi-label systems whose emergence has become more frequent in system biology (Chou *et al*., 2012; Lin *et al*., 2013; Xiao *et al*., 2011) and system medicine (Xiao *et al*., 2013; Qiu *et al*., 2016d), a completely different set of metrics are needed as elaborated in Chou (2013).

3.2 Cross-validation

With a set of well-defined metrics to measuring the quality of a predictor, the next thing is what kind of validation method should be used to score these metrics. In predictive analytics, the following three cross-validation methods are often used: (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 (see, e.g. Ahmad *et al*., 2016; Cai *et al*., 2003; Chou and Cai, 2003, 2005; Dehzangi *et al*., 2015; Fan *et al*., 2015; Ju *et al*., 2016; Kabir and Hayat, 2016; Khan *et al*., 2015; Kumar *et al*., 2015; Mondal and Pai, 2014; Shen *et al*., 2007; Tang *et al*., 2016; Zhou, 1998; Zhou and Assa-Munt, 2001; Zhou and Doctor, 2003). However, to reduce the computational time, in this study, we adopted the 10-fold cross-validation, as done by most investigators with SVM and random forests algorithms as the prediction engine. Besides, doing so would also facilitate the comparison since the reported results by Xu *et al*. (2016) was also derived from 10-fold cross-validation.

3.3 Result analysis and comparison

The success rates achieved by the new predictor iSumo-CD via the 10-fold cross-validation test on the same 510 proteins used by Xu *et al*. (2016) are given in Table 1, where for facilitating comparison, the corresponding rates obtained by the predictor SUMO_LDA (Xu *et al*., 2016) are also listed. As we can see from the table, the rate of Acc by new predictor pSumo-CD is 97.88%, which is about 11% higher than that by SUMO_LDA (Xu *et al*., 2016). The rate of MCC by pSumo-CD is 0.846, about 16% higher than that of SUMO_LDA. It is instructive to point out that, of the four metrics defined in Equation (20), the most important are the Acc and MCC (Chen *et al*., 2016a, c): the former reflects the overall accuracy of a predictor; while the latter, its stability in practical applications. The metrics Sn and Sp are used to measure a predictor from two opposite angles. When, and only when, both Sn and Sp of the predictor A are higher than those of the predictor B, can we say A is better than B (Liu *et al*., 2016d). In other words, Sn and Sp are actually constrained with each other (Chou, 1993). Therefore, it is meaningless to use only one of the two for comparing the quality of two predictors. A meaningful comparison in this regard should consider the rates of both Sn and Sp, or even better consider the rate of their combination that is none but just the score of MCC as shown in Table 1.

Table 1. A comparison of the proposed predictor with the state-of-the-art method in identifying the sumoylation sites in proteins[a]

Predictor	Acc (%)[d]	MCC[d]	Sn (%)[d]	Sp (%)
SUMO_LDA[b]	86.92	0.6845	98.71	84.51
pSumo-CD[c]	97.88	0.8460	82.01	99.21

[a] The scores here were generated by the 10-fold cross-validation on the 510 proteins as used by Xu et al. (2016).

[b] The predictor developed by Xu et al. (2016).

[c] The predictor proposed in this paper.

[d] See Equation (20) for the metrics definition.

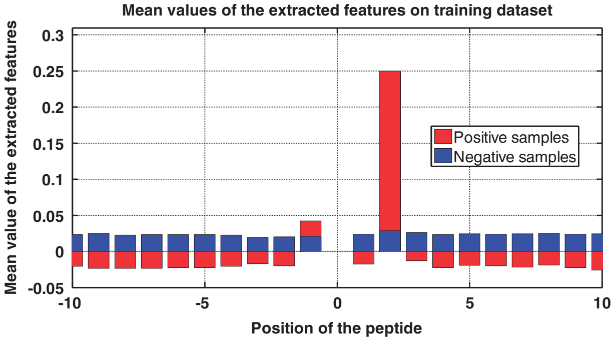


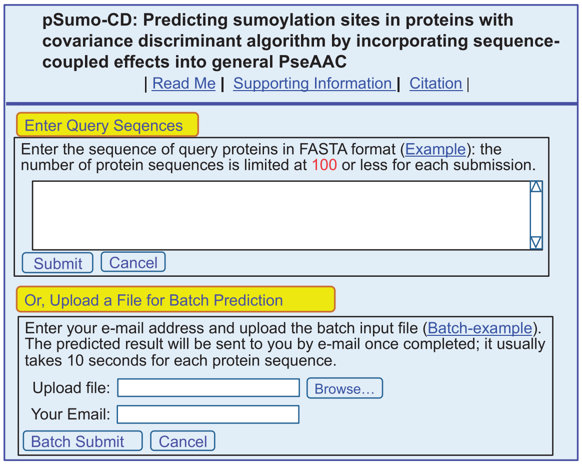
Fig. 1. Histograms to show the clusters of the true- and false-sumoylation peptide samples that are expressed by the general PseAAC of Equation (10). Each of the components is marked on the horizontal axis, and its average value on the vertical axis. The red histogram is for the mean value derived from the positive subset, and the blue for that from the negative subset. See the text for further explanation

Why could the proposed method be able to increase the prediction quality so substantially? First, a special term in the CD algorithm, namely $2\ln[\mathfrak{P}(\mathbb{S}^{\delta})]$ of Equation (14), has been taken into account in the current study. The term has the function to adjust the errors caused by the bias in a highly unbalanced benchmark dataset. Second, the amino acid-coupled effects around the sumoylation sites have been taken into account via the conditional probability approach as formulated in Equations (6)–(10). As a result, the cluster of the true-sumoylation samples (Fig. 1) can be more distinctly separated with that of the false-sumoylation samples, leading to a better success rates in discriminating them from each other. Similar remarkable successes have also been observed in predicting beta-turns (Zhang and Chou, 1997), alpha-turns (Chou, 1997), tight turns and their types in proteins (Chou, 2000b), specificity of GalNAc-transferase (Chou, 1995c), HIV protease cleavage sites (Chou, 1993; Chou *et al.*, 1996; Zhang and Chou, 1993), as well as signal peptide cleavage sites (Chou, 2001d; Chou and Shen, 2007b; Shen and Chou, 2007).

3.4 Web server and user guide

To enhance the value of its practical applications, the web-server for pSumo-CD has been established at http://www.jci-bioinfo.cn/pSumo-CD. Furthermore, to maximize the convenience for the majority of experimental scientists, a step-by-step guide is provided below.

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



pSumo-CD: Predicting sumoylation sites in proteins with covariance discriminant algorithm by incorporating sequence-coupled effects into general PseAAC

| Read Me | Supporting Information | Citation |

Enter Query Seqences

Enter the sequence of query proteins in FASTA format (Example): the number of protein sequences is limited at 100 or less for each submission.

Submit Cancel

Or, Upload a File for Batch Prediction

Enter your e-mail address and upload the batch input file (Batch-example). The predicted result will be sent to you by e-mail once completed; it usually takes 10 seconds for each protein sequence.

Upload file: Browse...

Your Email:

Batch Submit Cancel

Fig. 2. A semi-screenshot of the top-page for the web-server pSumo-CD at http://www.jci-bioinfo.cn/pSumo-CD

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.

Step 3. Click on the Submit button to see the predicted result. For example, if you use the three query protein sequences in the Example window as the input, in about 20 s after your submitting, you will see the following on the screen of your computer: (i) The 1st query protein (O95644) contains 51 K residues, of which residue 277, 293 and 703 are highlighted with red, meaning able to be of sumoylation. (ii) The 2nd query protein (B7ZR65) contains 25 K residues, of which residues 55, 253 and 365 are able to be of sumoylation. (iii) The 3rd query protein (P03496) contains 12 K residues, of which only the one at position 70 is highlighted with red meaning able to be of sumoylation. All the $(51+25+12)=88$ predicted outcomes are fully consistent with experimental observations except for the following two cases: residue 55 in the 2nd query protein was over-predicted (false positive) and residue 219 in the 3rd query protein was missed (false negative).

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 the Supporting Information button to download the benchmark dataset used in this study.

Step 6. Click on the Citation button to find the relevant papers that play the key roles in developing the new prediction method.

4 Conclusion

The pSumo-CD predictor is a new bioinformatics tool for identifying the sumoylation sites in proteins. Compared with the existing state-of-the-art predictor in this area, its prediction quality is much better, with remarkably higher overall accuracy and stability. For the convenience of most experimental scientists, we have provided its web-server and a step-by-step guide, by which users can easily obtain their desired results without the need to go through the detailed mathematics. The reason of including them in this paper is for the integrity of the new prediction method, and also for that some interesting techniques, such as incorporating the sequence-coupled approach into the general PseAAC, introducing the prior probability term in the CD algorithm to adjust the bias errors caused by unbalanced training dataset, and applying Chou's invariance theorem to overcome the divergence problem, may be of use as well in developing other tools in computational biology.

We anticipate that pSumo-CD will become a very useful high throughput tool for both basic research and drug development in the areas relevant to the protein sumoylation.

Acknowledgements

The authors wish to thank the three anonymous reviewers for their constructive comments, which were 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, 31560316), the Natural Science Foundation of Jiangxi Province and China (No. 20122BAB211033, 20122BAB201044, 20132BAB201053). 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,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.

Apweiler,R. *et al.* (2004) UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res.*, **32**, D115–D119.

Cai,Y.D. *et al.* (2003) Support vector machines for predicting membrane protein types by using functional domain composition. *Biophys. J.*, **84**, 3257–3263.

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

Chen,J. *et al.* (2007) Prediction of linear B-cell epitopes using amino acid pair antigenicity scale. *Amino Acids*, **33**, 423–428.

Chen,W. *et al.* (2012a) iNuc-PhysChem: a sequence-based predictor for identifying nucleosomes via physicochemical properties. *PLoS One*, **7**, e47843.

Chen,Y.Z. *et al.* (2012b) SUMOhydro: a novel method for the prediction of sumoylation sites based on hydrophobic properties. *PLoS One*, **7**, e39195.

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

Chen,W. *et al.* (2014a) iTIS-PseTNC: a sequence-based predictor for identifying translation initiation site in human genes using pseudo trinucleotide composition. *Anal. Biochem.*, **462**, 76–83.

Chen,W. *et al.* (2014b) iSS-PseDNC: identifying splicing sites using pseudo dinucleotide composition. *Biomed. Res. Int. (BMRI)*, **2014**, 623149.

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

Chen,W. *et al.* (2015a) iRNA-Methyl: Identifying N6-methyladenosine sites using pseudo nucleotide composition. *Anal. Biochem.*, **490**, 26–33. (Also, Data in Brief, 2015, 5: 376–378)

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

Chen,W. *et al.* (2015c) 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) Using deformation energy to analyze nucleosome positioning in genomes. *Genomics*, **107**, 69–75.

Chen,W. *et al.* (2016c) iRNA-PseU: identifying RNA pseudouridine sites. *Mol. Ther. Nucleic Acids*, **5**, e332.

Chou,K.C. (1993) A vectorized sequence-coupling model for predicting HIV protease cleavage sites in proteins. *J. Biol. Chem.*, **268**, 16938–16948.

Chou,K.C. (1995a) Does the folding type of a protein depend on its amino acid composition? *FEBS Lett.*, **363**, 127–131.

Chou,K.C. (1995b) A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space. *Proteins Struct. Funct. Genet.*, **21**, 319–344.

Chou,K.C. (1995c) A sequence-coupled vector-projection model for predicting the specificity of GalNAc-transferase. *Protein Sci.*, **4**, 1365–1383.

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

Chou,K.C. (1997) Prediction and classification of alpha-turn types. *Biopolymers*, **42**, 837–853.

Chou,K.C. (1999) A key driving force in determination of protein structural classes. *Biochem. Biophys. Res. Commun. (BBRC)*, **264**, 216–224.

Chou,K.C. (2000a) Prediction of protein subcellular locations by incorporating quasi-sequence-order effect. *Biochem. Biophys. Res. Commun. (BBRC)*, **278**, 477–483.

Chou,K.C. (2000b) Review: prediction of tight turns and their types in proteins. *Anal. Biochem.*, **286**, 1–16.

Chou,K.C. (2001a) Prediction of protein cellular attributes using pseudo amino acid composition. *Proteins Struct. Funct. Genet.*, **44**, 246–255. (Erratum: ibid., (2001) Vol. 43).

Chou,K.C. (2001b) Prediction of protein signal sequences and their cleavage sites. *Proteins Struct. Funct. Genet.*, **42**, 136–139.

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

Chou,K.C. (2001d) Using subsite coupling to predict signal peptides. *Protein Eng.*, **14**, 75–79.

Chou,K.C. (2005a) Prediction of G-protein-coupled receptor classes. *J. Proteome Res.*, **4**, 1413–1418.

Chou,K.C. (2005b) 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. (2015) Impacts of bioinformatics to medicinal chemistry. *Med. Chem.*, **11**, 218–234.

Chou,K.C. and Cai,Y.D. (2003) Prediction and classification of protein subcellular location: sequence-order effect and pseudo amino acid composition. *J. Cell. Biochem.*, **91**, 1250–1260. (Addendum, ibid. (2004) 90)

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

Chou,K.C. and Elrod,D.W. (1998) Using discriminant function for prediction of subcellular location of prokaryotic proteins. *Biochem. Biophys. Res. Commun. (BBRC)*, **252**, 63–68.

Chou,K.C. and Elrod,D.W. (1999a) Prediction of membrane protein types and subcellular locations, *Proteins: Struct. Funct., Genet.*, **34**, 137–153.

Chou,K.C. and Elrod,D.W. (1999b) Protein subcellular location prediction. *Protein Eng.*, **12**, 107–118.

Chou,K.C. and Elrod,D.W. (2002) Bioinformatical analysis of G-protein-coupled receptors. *J. Proteome Res.*, **1**, 429–433.

Chou,K.C. and Elrod,D.W. (2003) Prediction of enzyme family classes. *J. Proteome Res.*, **2**, 183–190.

Chou,K.C. *et al.* (1998) Prediction and classification of domain structural classes. *Proteins Struct. Funct. Genet.*, **31**, 97–103.

Chou,K.C. and Maggiora,G.M. (1998) Domain structural class prediction. *Protein Eng.*, **11**, 523–538.

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. Commun. (BBRC)*, **357**, 633–640.

Chou,K.C. *et al.* (1996) Predicting HIV protease cleavage sites in proteins by a discriminant function method. *Proteins: Struct. Funct. Genet.*, **24**, 51–72.

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. *Mol. Biosyst.*, **8**, 629–641.

Chou,K.C. and Zhang,C.T. (1994) Predicting protein folding types by distance functions that make allowances for amino acid interactions. *J. Biol. Chem.*, **269**, 22014–22020.

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

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.

Ding,H. *et al.* (2014) iCTX-Type: A sequence-based predictor for identifying the types of conotoxins in targeting ion channels. *BioMed Res. Int. (BMRI)*, **2014**, 286419.

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.

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.

Elrod,D.W. and Chou,K.C. (2002) A study on the correlation of G-protein-coupled receptor types with amino acid composition. *Protein Eng.*, **15**, 713–715.

Fan,G.L. *et al.* (2015) DSPMP: discriminating secretory proteins of malaria parasite by hybridizing different descriptors of Chou's pseudo amino acid patterns. *J. Comput. Chem.*, **36**, 2317–2327.

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. (JBSD)*. 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.

Ju,Z. *et al.* (2016) Predicting lysine phosphoglycerylation with fuzzy SVM by incorporating k-spaced amino acid pairs into Chou's general PseAAC. *J. Theor. Biol.*, **397**, 145–150.

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

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,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. Biomed. Sci. Eng. (JBiSE)*, **6**, 435–442.

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) Identification of real microRNA precursors with a pseudo structure status composition approach. *PLoS One*, **10**, e0121501.

Liu,B. *et al.* (2015b) Identification of microRNA precursor with the degenerate K-tuple or Kmer strategy. *J. Theor. Biol.*, **385**, 153–159.

Liu,B. *et al.* (2015c) 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.* (2015d) 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,Z. *et al.* (2015e) iDNA-Methyl: Identifying DNA methylation sites via pseudo trinucleotide composition. *Anal. Biochem.*, **474**, 69–77. (also, Data in Brief, 2015, 4: 87–89),

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) repRNA: a web server for generating various feature vectors of RNA sequences. *Mol. Genet. Genomics*, **291**, 473–481.

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

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

Liu,W. and Chou,K.C. (1998) Prediction of protein structural classes by modified Mahalanobis discriminant algorithm. *J. Protein Chem.*, **17**, 209–217.

Mahalanobis,P.C. (1936) On the generalized distance in statistics. *Proc. Natl. Inst. Sci. India*, **2**, 49–55.

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.

Qiu,W.R. *et al.* (2016a) iPhos-PseEvo: identifying human phosphorylated proteins by incorporating evolutionary information into general PseAAC via grey system theory. *Mol. Inf.*, 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. (2016c) iHyd-PseCp: Identify hydroxyproline and hydroxylysine in proteins by incorporating sequence-coupled effects into general PseAAC. *Oncotarget*, **7**, 44310–44321.

Qiu,W.R. *et al.* (2016d) iPTM-mLys: identifying multiple lysine PTM sites and their different types. *Bioinformatics*, doi:10.1093/bioinformatics/btw380.

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.

Ren,J. *et al.* (2009) Systematic study of protein sumoylation: development of a site-specific predictor of SUMOsp 2.0. *Proteomics*, **9**, 3409–3412.

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

Shen,H.B. *et al.* (2007) Euk-PLoc: an ensemble classifier for large-scale eukaryotic protein subcellular location prediction. *Amino Acids*, **33**, 57–67.

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.

Xiao,X. *et al.* (2015) 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**, 2221–2233.

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.* (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*. (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,J. *et al*. (2008) A novel method for high accuracy sumoylation site prediction from protein sequences. *BMC Bioinformatics*, **9**, 1.

Xu,Y. *et al*. (2013) iSNO-PseAAC: predict cysteine S-nitrosylation sites in proteins by incorporating position specific amino acid propensity into pseudo amino acid composition. *PLoS One*, **8**, e55844.

Xu,Y. *et al*. (2016) Prediction of sumoylation sites in proteins using linear discriminant analysis. *Gene*, **576**, 99–104.

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

Xue,Y. *et al*. (2006) SUMOsp: a web server for sumoylation site prediction. *Nucleic Acids Res*., **34**, W254–W257.

Zhang,C.T. and Chou,K.C. (1993) An alternate-subsite-coupled model for predicting HIV protease cleavage sites in proteins. *Protein Eng*., **7**, 65–73.

Zhang,C.T. and Chou,K.C. (1997) Prediction of beta-turns in proteins by 1-4 and 2-3 correlation model. *Biopolymers*, **41**, 673–702.

Zhao,Q. *et al*. (2014) GPS-SUMO: a tool for the prediction of sumoylation sites and SUMO-interaction motifs. *Nucleic Acids Res*., gku383.

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**, 216–217.

Zhou,G.P. and Assa-Munt,N. (2001) Some insights into protein structural class prediction. *Proteins Struct. Funct. Genet*., **44**, 57–59.

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

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