GPS-Palm Manual

Prediction of *S*-Palmitoylation sites

18/02/2020

Author: Wanshan Ning, Peiran Jiang and Yu Xue[*]

Index

Statement

1. **Implementation**. The softwares of the CUCKOO Workgroup are implemented in Python (3.6) and package PyQt5. Usually, online service or local stand-alone packages will be provided.

2. **Availability**. Our softwares are freely available for academic researches. For non-profit users, you can copy, distribute and use the softwares for your scientific studies. Our softwares are not free for commercial usage.

3. **GPS**. Previously, we used the GPS to denote our Group-based Scoring algorithm. Currently, we are developing an integrated computational platform for post-translational modifications (PTMs) of proteins. We re-denote the GPS as Group-based Prediction Systems. Here, we incorporated data quality discrimination (DQD) into our previous GPS, and used number-to-image transformation (NIT) to transform GPS features into images. We rename GPS as Graphic Presentation System. This software is an indispensable part of GPS.

4. **Usage**. Our softwares are designed in an easy-to-use manner. Also, we invite you to read the manual before using the softwares.

5. **Updation**. Our softwares will be updated routinely based on users' suggestions and advices. Thus, your feedback is greatly important for our future updation. Please do not hesitate to contact with us if you have any concerns.

6. **Citation**. Usually, the latest published articles will be shown on the software websites. We wish you could cite the article if the software has been helpful for your work.

Introduction

As an important and special class of post-translational modifications (PTMs), lipid modifications mainly comprise *S*-palmitoylation (C16), *N*-myristoylation (C14), *S*-farnesylation (C15), *S*-geranylgeranylation (C20), cholesterylation and glycosylphosphatidylinositol (GPI)-anchor, depending on the type of lipids covalently attached to modified substrate proteins [1, 2]. Unlike other tethering lipid modifications, *S*-palmitoylation reversibly adds one or multiple palmitoyl moieties to internal cysteine residues in proteins through the thioesterification reaction [3-7]. *S*-palmitoylation effectively increases the hydrophobicity of protein surfaces to dynamically regulate membrane-protein interactions [1, 8], and participates in regulating a broad spectrum of biological processes, such signal transduction [2, 7], neuronal transmission [3], metabolism [9], autophagy [10], and immunological response [11]. In addition, the dysregulation of *S*-palmitoylation is associated with numerous human diseases such as cancer [11, 12], neurodegenerative disorders [13] and type 1 diabetes [14]. Although the biological importance of protein *S*-palmitoylation has been gradually recognized in recent years, its underlying mechanisms are still unclear.

Here, we collected 3098 known *S*-palmitoylation sites experimentally identified from small- or large-scale studies, and developed a new method named **data quality discrimination** (DQD) to measure **data quality weights** (DQWs) of the sites derived from the two types of experiments. We incorporated DQD into our previous methods, transformed sequence similarity values into images, and designed a novel algorithm of **graphic presentation system** (GPS) 6.0. Using the same **number-to-image transformation** (NIT), we further integrated 9 additional types of sequence and structure features, implemented a framework of **parallel convolutional neural networks** (pCNNs), and constructed a new predictor called GPS-Palm. Compared with other existing tools, GPS-Palm showed a > 31.3% improvement of area under the curve (AUC) value (0.855 vs. 0.651) for predicting *S*-palmitoylation sites. Finally, we constructed the **GPS-Palm software package for the convenient use**.
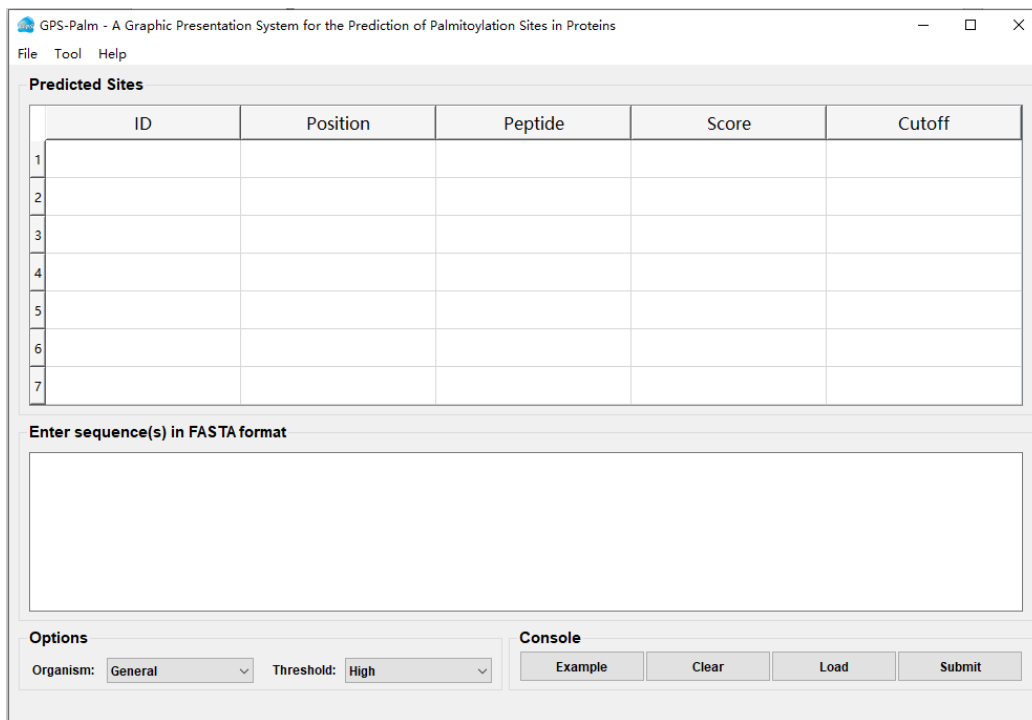
GPS-Palm software User Interface

Download & Installation

The software of GPS-Palm was implemented in Python, and could be installed on Windows systems. GPS-Palm distributions for Windows can be found at http://gpspalm.biocuckoo.cn/download.php. We recommend that users could download the latest release.

After downloading, please double-click on the file *GPS-Palm_windows.exe* to begin installation. Follow the user prompts through the installation. And snapshots of the setup program are shown below:
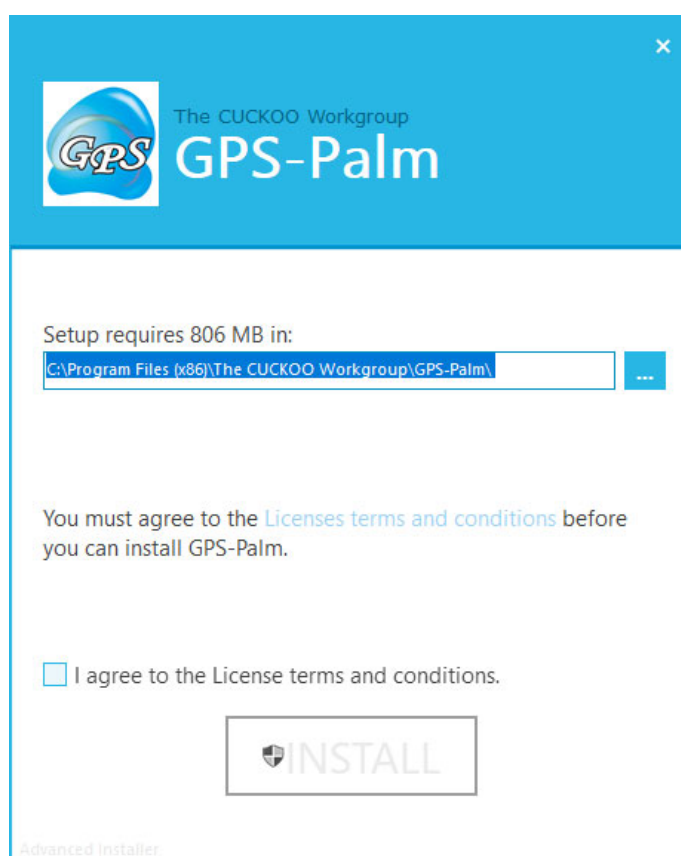
Click on the **I agree to the Licenses terms and conditions** button. Users can change the directory. Attention, administrator account is required for saving your predicted results if users install GPS-Palm in authority directory. This could be avoided by converting to the non-administrative installation directory.
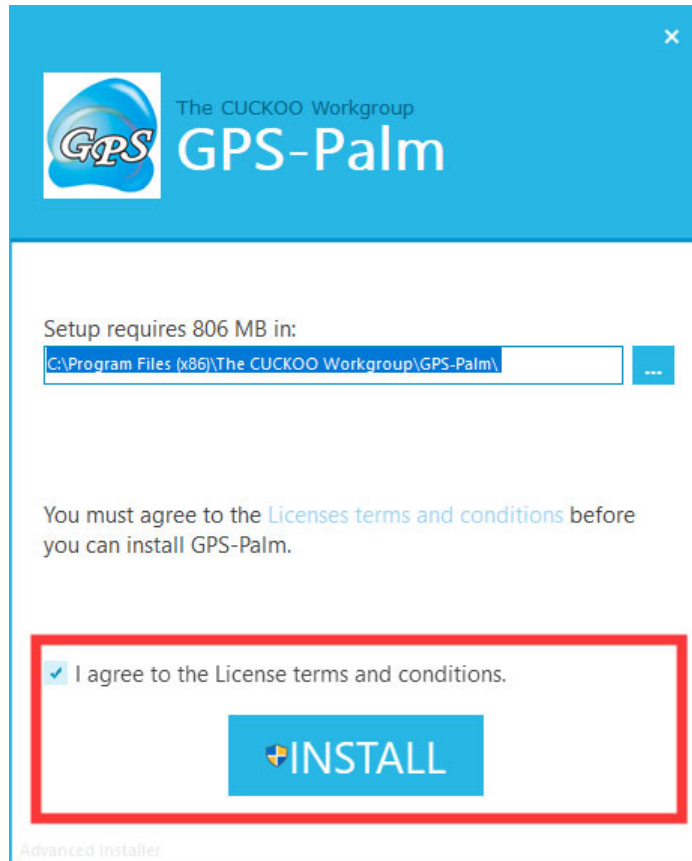
Click on the **Install** button to start the setup program.
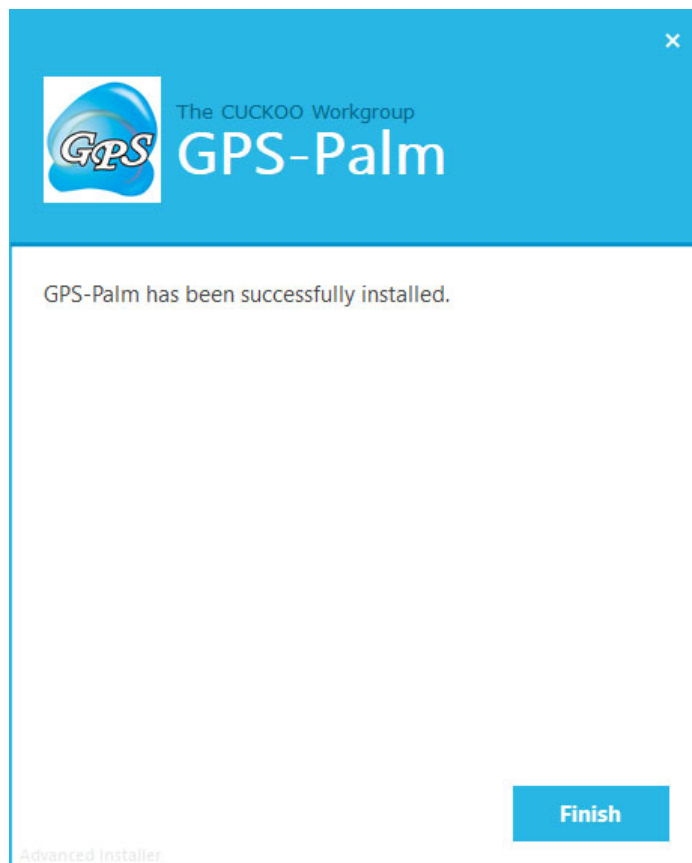
Click on the **Finish** button to complete the setup program.

Prediction of Protein *S*-palmitoylation Sites

Direct Prediction

For convenience, the GPS-Palm allows users to input their protein sequences into the "TEXT form" for prediction. One or more protein sequences should be prepared in **FATSA** format as below:

>protein1
XXXXXXXXXXXXX
XXXXXXXX
>protein2
XXXXXXXXXXXXXXXX…
>protein3
XXXXXXXXXXXX

Please note: All irregular words, including non-amino acid word (eg, number) and blank, will be removed automatically. As an instance, we put **human Programmed cell death 1 ligand 1 (PD-L1)** protein sequence as an example for GPS-Palm. Users could click on the "Example" button to access the example.

Choose an **Organism to pinpoint** what you need, default is **General (no speices-specific)**. More precise prediction is provided for human and mouse

Choose a **Threshold** what you need, default is **High**.

Click on the **Submit** button, then the predicted *S*-Palmitoylation sites will be shown.

Batch Prediction

We also provide an alternative approach for processing multiple protein sequences. If the file is large, the **Batch Predictor** will be convenient for users.

The following steps show you how to use it:
Put protein sequences into a file with **FATSA** format as below:

>protein1
XXXXXXXXXXXXX
XXXXXXXX
>protein2
XXXXXXXXXXXXXXXX…
>protein3
XXXXXXXXXXXX

The names of proteins are necessary (the line with ">" and a protein name/accession number).

To run the Batch Predictor just select the **Load** option in the **File** menu.
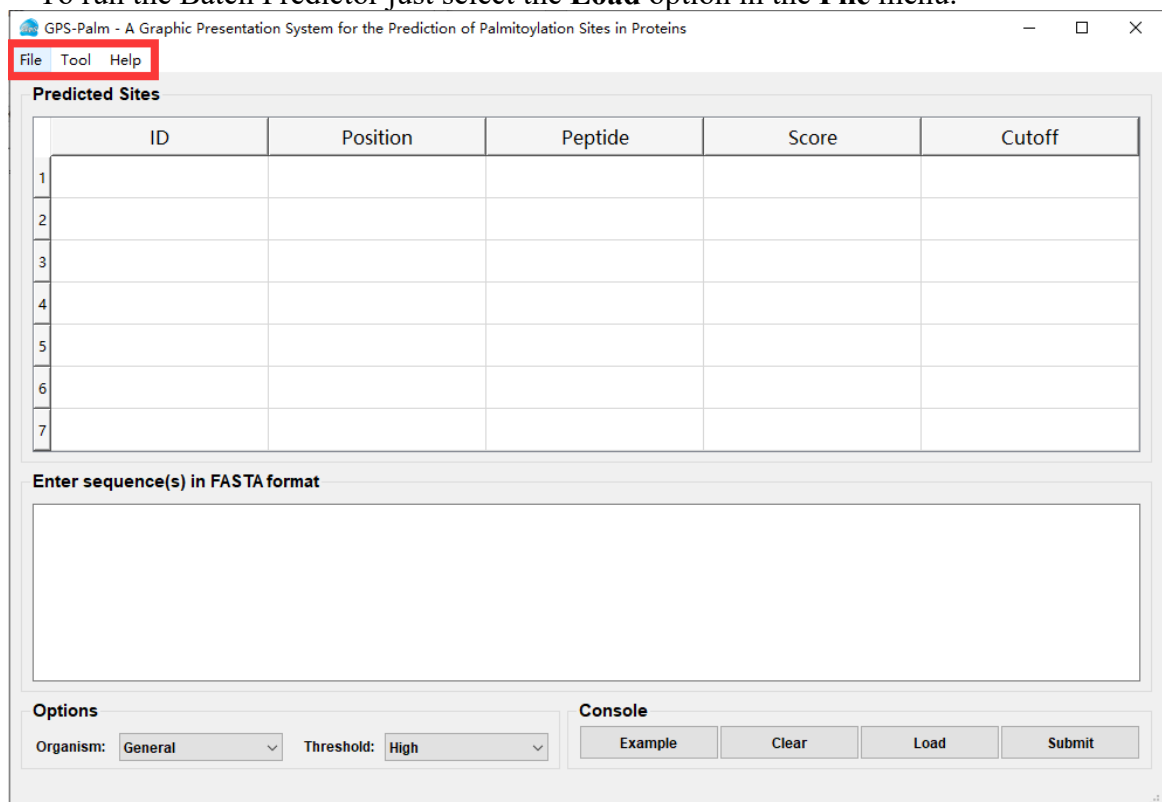
Click on the **Add File** button and add one or more protein sequence files in your hard disk.
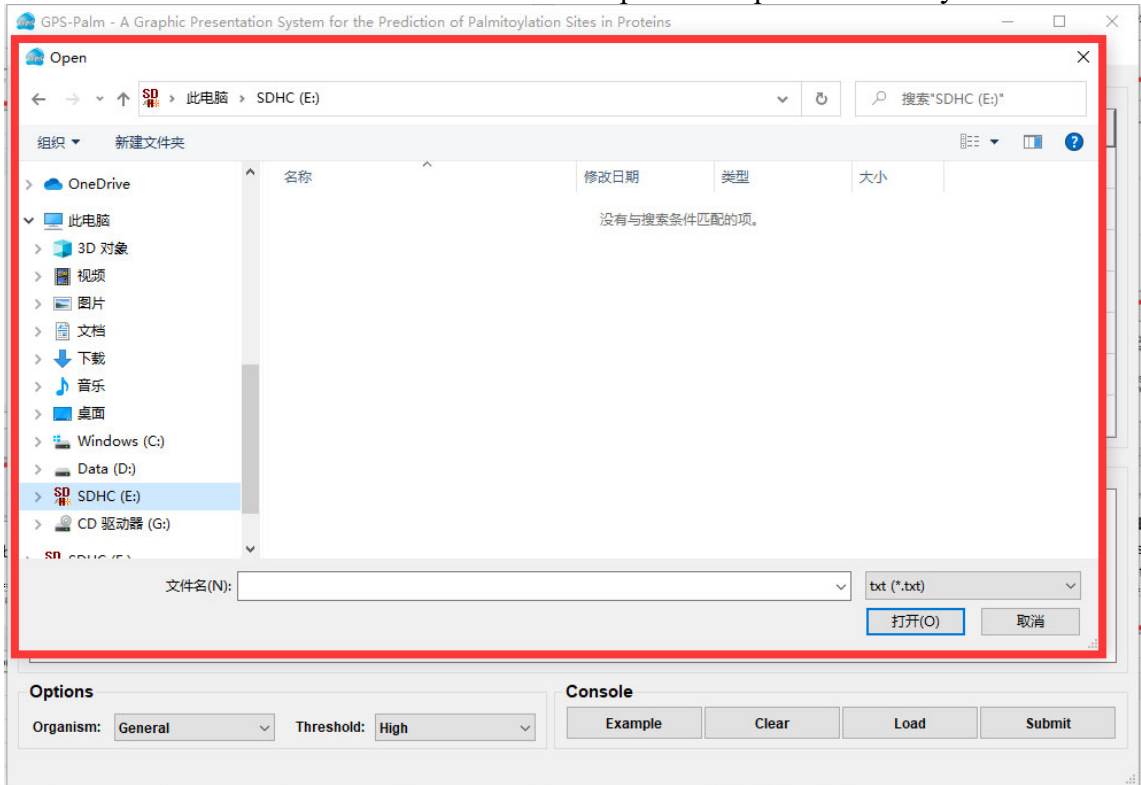
The name of added proteins will be shown in the text window.

Click on the Submit button, then the predicted *S*-Palmitoylation sites will be shown.

The output directory of prediction results should also be defined. Please click on the **File** button and **Export** bottom to specify the export file fold. Users can also export the predicted results by right-click and select the Export Result option.

Algorithms and Prediction Performance

The GPS 6.0 Algorithm Design

In 2004, we developed the GPS 1.0 algorithm with the full name of group-based phosphorylation site predicting and scoring platform to measure the local sequence similarity between a given phosphorylation site (p-site) peptide and all known p-sites in positive data, based on a hypothesis of similar peptides potentially sharing similar properties [15]. Later, we renamed it into group-based prediction system [16], whereas the basic scoring strategy was never altered in all versions of GPS algorithms, and the latest GPS 5.0 algorithm implemented two additional approaches including position weight determination (PWD) and scoring matrix optimization (SMO) for performance improvement (http://gps.biocuckoo.cn/).

In the scoring strategy, the similarity score between two PSP(10, 10) peptide A and B was formulated as below:

$$S(A, B) = \sum_{1 \leq i \leq 20} W_i \, M(A[i], B[i])$$

Here, W_i was the weight of position i, and the $M(A[i], B[i])$ was the substitution score of the amino acid pair $(A[i], B[i])$ occurring at position i. The substitution score is symmetrical with $M(a, b)=M(b, a)$. Initially, all position values in W were taken as 1, while the BLOSUM62 matrix was used as the starting matrix. Then in the step of performance improvement, PWD and SMO were iteratively adopted to optimize the trainable parameters in W and M, respectively, until the AUC value of the 10-fold cross-validations was not increased any longer. In GPS 5.0, the original PLR algorithm with the ridge (L2) penalty was used for training models.

In this study, we further developed two new methods including DQD and NIT, and incorporated them into our previous approaches to develop the GPS 6.0 algorithm. For DQD, we defined the average similarity score $PS(A)$ between a given PSP(10, 10) A and the whole positive data set P with T_+ peptides as:

$$PS(A) = \frac{1}{T_+} \sum_{j=1}^{T_+} S(A, P_j) \, pDQW_j$$

Where $pDQW_j$ was the DQW value of P_j in P, and $pDQW$ was a weight vector of trainable DQWs for the positive data set. Also, we defined the average similarity score $NS(A)$ between a given PSP(10, 10) A and the whole negative data set N with T_- peptides as:

$$NS(A) = \frac{1}{T_-} \sum_{j=1}^{T_-} S(A, N_j) \, nDQW_j$$

Where $nDQW_j$ was the DQW value of N_j in N, and $nDQW$ was a weight vector of trainable DQWs for the negative data set. All values in $pDQW$ and $nDQW$ were initialized as 1, and the improved PLR algorithm was used to optimize DQWs. DQD, PWD and SMO were iteratively performed until the 10-fold cross-validation AUC value was not enhanced any longer.

In order to further improve the accuracy, we used NIT to transform GPS features into images, and rename GPS as graphic presentation system. For the positive data set, we first transformed individual $PS(A)$ values into a similarity matrix $Mat_+(A)$, in which the 21 rows represented 21 types of pseudo amino acids (A, C, D, …, Y, $*$) shown in alphabetical order and the 20 columns denoted 20 positions in PSP(10, 10) peptide A (From -10 to 10). The central S-palmitoylated cysteine residues were not taken into the consideration to avoid overfitting. The matrix was shown as following:

$$Mat_+(A) = \begin{pmatrix} Mat_+(A)[A,-10] & \cdots & Mat_+(A)[A,10] \\ \vdots & \ddots & \vdots \\ \vdots & \ddots & \vdots \\ Mat_+(A)[*,-10] & \cdots & Mat_+(A)[*,10] \end{pmatrix}_{21 \times 20}$$

In $Mat_+(A)$, any value $Mat_+(A)[a, i]$ for an amino acid a in the position i of A could be calculated as below:

$$Mat_+(A)[a, i] = \frac{1}{T_+(a,i)} \sum_{j=1}^{T_+(a,i)} W_i M(a, A[i]) pDQW_j$$

Where $T_+(a, i)$ was the number of PSP(10, 10) items in the positive data set with the residue a at position i. Analogously, the similar matrix $Mat_-(A)$ was also determined between A and the whole negative data set.

For each given PSP(10, 10) peptide A, two similarity matrices $Mat_+(A)$ and $Mat_-(A)$ were generated and transformed into an image in RGB formation with 2 layers. The red and green channels were used for representing $Mat_+(A)$ and $Mat_-(A)$, respectively. For the red channel, any element E in the $Mat_+(A)$ was normalized to 0~255 as below:

$$E_{Noramlized} = \frac{E - E_{min}}{E_{max} - E_{min}} \times 255$$

where E_{max} and E_{min} were the maximum and minimum elements in the $Mat_+(A)$, respectively. The same procedure was also conducted to $Mat_-(A)$ by using the green channel. Two channels merged together to output an intact image, and the blue channel was not utilized in this work. Thus, the images with two layers contained both the similarity values of A against the whole positive and negative data sets, respectively. The ultimate graphic presentation of GPS features for a given PSP(10, 10) peptide was an informative 21×20-pixel doubly colored image.

Evaluation of Prediction Performances

Performance measurements

To evaluate the prediction performances, four standard measurements were used, including accuracy (*Ac*), sensitivity (*Sn*), specificity (*Sp*) and Mathew correlation coefficient (*MCC*). Accuracy (*Ac*) represents the correct ratio between both positive (+) and negative (-) data sets, while sensitivity (*Sn*) and specificity (*Sp*) illustrate the correct prediction ratios of positive (+) and negative data (-) sets respectively. Since the number of positive data and negative data differed too much from each other, the Mathew correlation coefficient (*MCC*) was also included. The value of *MCC* ranges from -1 to 1, and a larger *MCC* value stands for better prediction performance.

Among the data with positive hits by GPS-Palm, the real positives were defined as *true positives* (*TP*), while the others were defined as *false positives* (*FP*). Among the data with negative predictions, the real positives were defined as *false negatives* (*FN*), while the others were defined as *true negatives* (*TN*). The four measurements of sensitivity (*Sn*), specificity (*Sp*), accuracy (*Ac*), and Mathew correlation coefficient (*MCC*) were defined as below:

$$Sn = \frac{TP}{TP + FN},$$

$$Sp = \frac{TN}{TN + FP},$$

$$Ac = \frac{TP + TN}{TP + FP + TN + FN},$$

$$Pr = \frac{TP}{TP + FP}$$

$$MCC = \frac{(TP \times TN) - (FN \times FP)}{\sqrt{(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)}}$$

References

1.   Ray A, Jatana N, Thukral L. Lipidated proteins: Spotlight on protein-membrane binding interfaces, Prog Biophys Mol Biol 2017;128:74-84.
2.   Casey PJ. Protein lipidation in cell signaling, Science 1995;268:221-225.
3.   Roth AF, Wan J, Bailey AO *et al*. Global analysis of protein palmitoylation in yeast, Cell 2006;125:1003-1013.
4.   Dietrich LE, Ungermann C. On the mechanism of protein palmitoylation, EMBO Rep 2004;5:1053-1057.
5.   Greaves J, Chamberlain LH. Palmitoylation-dependent protein sorting, J Cell Biol 2007;176:249-254.
6.   Linder ME, Deschenes RJ. Palmitoylation: policing protein stability and traffic, Nat Rev Mol Cell Biol 2007;8:74-84.
7.   Smotrys JE, Linder ME. Palmitoylation of intracellular signaling proteins: regulation and function, Annu Rev Biochem 2004;73:559-587.
8.   Kleuss C, Krause E. Galpha(s) is palmitoylated at the N-terminal glycine, Embo J 2003;22:826-832.
9.   Shen LF, Chen YJ, Liu KM *et al*. Role of S-Palmitoylation by ZDHHC13 in Mitochondrial function and Metabolism in Liver, Sci Rep 2017;7:2182.
10.   Kim SW, Kim DH, Park KS *et al*. Palmitoylation controls trafficking of the intracellular Ca(2+) channel MCOLN3/TRPML3 to regulate autophagy, Autophagy 2018:1-14.
11.   Yao H, Lan J, Li C *et al*. Inhibiting PD-L1 palmitoylation enhances T-cell immune responses against tumours, Nat Biomed Eng 2019;3:306-317.
12.   Chen S, Zhu B, Yin C *et al*. Palmitoylation-dependent activation of MC1R prevents melanomagenesis, Nature 2017;549:399-403.
13.   Andrew RJ, Fernandez CG, Stanley M *et al*. Lack of BACE1 S-palmitoylation reduces amyloid burden and mitigates memory deficits in transgenic mouse models of Alzheimer's disease, Proc Natl Acad Sci U S A 2017;114:E9665-e9674.
14.   Berchtold LA, Storling ZM, Ortis F *et al*. Huntingtin-interacting protein 14 is a type 1 diabetes candidate protein regulating insulin secretion and beta-cell apoptosis, Proc Natl Acad Sci U S A 2011;108:E681-688.
15.   Zhou FF, Xue Y, Chen GL *et al*. GPS: a novel group-based phosphorylation predicting and scoring method, Biochem Biophys Res Commun 2004;325:1443-1448.
16.   Xue Y, Ren J, Gao X *et al*. GPS 2.0, a tool to predict kinase-specific phosphorylation sites in hierarchy, Mol Cell Proteomics 2008;7:1598-1608.

Release Note

1. Jan. 15th, 2008, the online service and the local stand-alone packages of CSS-Palm 2.0 were released.
2. Jan. 29th, 2008, a bug was found that the version 2.0 couldn't be properly used under non-English Operating Systems.
3. Feb. 16th, 2008, the CSS-Palm 2.0.1 manual was written and included in the packages.
4. Aug. 28th, 2008, DOG (Domain Graph) 1.0 was integrated into CSS-Palm 2.0.2 and a new function of visualizing the predicted sites was added. Feb. 2nd, 2009，CSS-Palm version 2.0.3 was released. We moved the CSS-Palm web server to a new website (http://csspalm.biocuckoo.org) and a new GPS logo was put into use.
5. Jul. 23rd, 2009, CSS-Palm version 2.0.4 was released. Check for update function was added. DOG (Domain Graph) was updated to version 1.0.5.
6. Aug. 31th, 2013, CSS-Palm version 4.0 beta was released.
7. Nov. 20th, 2019, The GPS-Palm was released, with local packages and support three major Operating Systems including Windows, Linux/Unix and Mac. In this version, we incorporated data quality discrimination (DQD) into our previous methods, transformed sequence similarity values into images, and designed a novel algorithm of graphic presentation system (GPS) 6.0. Using the same number-to-image transformation (NIT), we further integrated 9 additional types of sequence and structure features, implemented a framework of parallel convolutional neural networks (pCNNs), and constructed the GPS-Palm. Meanwhile, the GPS-Palm manual was updated and included in the packages.