6

Computational Prediction of Post-Translational Modification Sites in Proteins

Yu Xue[1], Zexian Liu[2], Jun Cao[2] and Jian Ren[3]
[1]*Huazhong University of Science & Technology, Wuhan, Hubei,*
[2]*University of Science & Technology of China, Hefei, Anhui,*
[3]*Sun Yat-sen University, Guangzhou, Guangdong,*
China

1. Introduction

The last several decades have witnessed rapid progresses in identification and functional analysis of post-translational modifications (PTMs) in proteins. Through temporal and spatial modification of proteins by covalent attachment of additional chemical groups and other small proteins, proteolytic cleavage or intein splicing, PTMs greatly expand the proteome diversity and play important roles in regulating the stability and functions of the proteins (Mann and Jensen, 2003; Walsh, 2005; Walsh and Jefferis, 2006). To date, more than 350 types of distinct PTMs were experimentally discovered *in vivo*, while subsequently functional assays have detected a number of exciting observations. In 1992, the Nobel Prize in Physiology or Medicine was awarded to Edmond H. Fischer and Edwin G. Krebs for their seminal discovery that reversible protein phosphorylation is a biological regulatory mechanism (Kresge et al., 2011), while Leland H. Hartwell, Tim Hunt, and Paul M. Nurse shared the Nobel Prize in Physiology or Medicine 2001 for their profound contributions in identification of key regulators including cyclin-dependent kinases (CDKs) and cyclins that precisely orchestrate the cell cycle process through phosphorylation (Balter and Vogel, 2001). Moreover, Aaron Ciechanover, Avram Hershko and Irwin Rose became laureates of the Nobel Prize in Chemistry 2004 for their discovery of ubiquitin-mediated protein degradation (Vogel, 2004).

Although virtually all PTMs play their major roles as regulating the biological processes, different ones have their aspects with emphasis. For example, phosphorylation is preferentially implicated in signal-transduction cascades, while ubiquitination regulates the lifetime of proteins by targeting specific substrates for degradation. Recently, protein lysine acetylation was observed to play a predominant role in regulation of cellular metabolism (Wang et al., 2010; Zhao et al., 2010). Other types of PTMs such as sumoylation, glycosylation and palmitoylation are also critical for exactly orchestrating distinct cellular processes (Fukata and Fukata, 2010; Linder and Deschenes, 2007). Furthermore, the crosstalk among different PTMs is ubiquitous, especially on histones, which is regarded as the "histone code" (Jenuwein and Allis, 2001). The aberrances of PTMs are highly associated in diseases and cancers, while a variety of regulatory enzymes involved in PTMs have been drug targets (Lahiry et al., 2010; Norvell and McMahon, 2010). In this regard, elucidation of PTMs regulatory roles is fundamental for understanding molecular mechanisms of diseases and cancers, and further biomedical design.

Recently, with the developments of "state-of-the-art" techniques especially the high-throughput mass spectrometry (HTP-MS), large-scale identification of PTMs substrates with their sites has become a popular and near-routine assay (Choudhary and Mann, 2010). For example, combined with efficient isolation and enrichment methods such as antibodies which specifically recognize modified peptides and subsequently HTP-MS profiling, thousands of PTMs sites, eg., phosphorylation (Olsen et al., 2006; Villen et al., 2007), acetylation (Choudhary et al., 2009), or glycosylation (Zielinska et al., 2010) sites can be accurately determined in a single experiment. These high-throughout approaches can provide systematic insights into the biological roles of PTMs, especially a global view. However, due to the technical limitations such as low-sensitive detection of modifications in low expressed proteins (Ackermann and Berna, 2007; Boschetti and Righetti, 2009; Yates et al., 2009), and error-prone determination of multiple modified proteins (Hunter, 2007; Young et al., 2010), it is still a great challenge for fully charactering the whole PTM events *in vivo*.

In contrast with conventional experimental methods, computational analysis of PTMs has also been an alternative and attractive approach for its accuracy, fast-speed and convenience. The computational predictors can narrow down the number of potentially candidates and rapidly generate useful information for further experimental investigations. In one of our recent reviews, we specifically summarized more than 50 computational resources including public databases and prediction tools for phosphorylation (Xue et al., 2010a). Currently, although there have been ~170 databases and computational tools developed for PTM analysis (http://www.biocuckoo.org/link.php), accurate prediction of PTM sites in given proteins is still not a simple job. Again, although protein 3D structure information can be helpful for prediction of PTMs sites (Kumar and Mohanty, 2010), mainstream computational approaches were designed mainly based on protein primary sequence features (Xue et al., 2010a). A variety of algorithms have been introduced into this field, such as position-specific scoring matrix (PSSM) (Obenauer et al., 2003), support vector machines (SVMs) (Kim et al., 2004), artificial neural network (ANN) (Blom et al., 2004), Hidden Markov Model (HMM) (Huang et al., 2005), Bayesian decision theory (Xue et al., 2006a), and Conditional Random Field (CRF) (Dang et al., 2008). These methods were largely introduced from the fields of informatics or statistics and originally designed for general propose.

Previously, we developed a series of GPS algorithms (Initially defined as Group-based Phosphorylation Scoring and later renamed as Group-based Prediction System), which have been exclusively and successfully used for the prediction of kinds of PTM sites, such as phosphorylation (Xue et al., 2005; Xue et al., 2008; Xue et al., 2011; Zhou et al., 2004), sumoylation (Ren et al., 2009; Xue et al., 2006b), palmitoylation (Ren et al., 2008; Zhou et al., 2006a), *S*-Nitrosylation (Xue et al., 2010b) and nitration (Liu et al., 2011). The prediction performance of GPS 1.x (1.0 and 1.1) could be comparative to other analogous approaches, while GPS 2.x versions (2.0 and 2.1) are much better than other strategies (Xue et al., 2010a). Recently, we greatly improved our previous method and released the GPS 3.0 algorithm, while has been successfully adopted for predicting *S*-nitrosylation (Xue et al., 2010b) and nitration sites (Liu et al., 2011), with further enhanced performances. During the past several years, a considerable number of bioinformatists or experimentalists have communicated with us on GPS algorithm details, which were never thoroughly described in our previously research articles due to page limitation. In this regard, the aim of this chapter is to provide a comprehensive description of GPS series algorithms. First, we gave a historical introduction of GPS 1.x, GPS 2.x and the latest GPS 3.0 algorithms. Also, we used palmitoylation as an

application to design a site-specific predictor of CSS-Palm 3.0 with GPS 3.0. The procedures of benchmark data preparation, scoring strategies, performance evaluation and comparison, and software package implementation were clearly described. The online service and local packages of CSS-Palm 3.0 are freely available at: http://csspalm.biocuckoo.org/. For convenience, the computational tools developed in GPS series algorithms were summarized in Table 1.

Name	PTM type	Website Link	Reference
GPS 1.0 & 1.1	Phosphorylation	http://gps.biocuckoo.org/1.1/	Xue *et al*., 2005; Zhou *et al*., 2004
CSS-Palm 1.0	Palmitoylation	http://csspalm.biocuckoo.org/1.0/	Zhou *et al*., 2006a
SUMOsp 1.0	Sumoylation	http://sumosp.biocuckoo.org/1.0/	Xue *et al*., 2006b
GPS 2.0 & 2.1	Phosphorylation	http://gps.biocuckoo.org/	Xue *et al*., 2008; Xue *et al*., 2011
CSS-Palm 2.0	Palmitoylation	http://csspalm.biocuckoo.org/	Ren *et al*., 2008
SUMOsp 2.0	Sumoylation	http://sumosp.biocuckoo.org/	Ren *et al*., 2009
GPS-SNO 1.0	*S*-Nitrosylation	http://sno.biocuckoo.org/	Xue *et al*., 2010b
GPS-YNO2 1.0	nitration	http://yno2.biocuckoo.org/	Liu *et al*., 2011

Table 1. The computational tools constructed with GPS series algorithms.

2. GPS series algorithms

In this section, we described the theoretical basis and developmental history of GPS series algorithms. The chief hypothesis of the algorithm is established on consensus experimental observations that if two short peptides share high sequence homology, they may exhibit similar 3D structures and biochemical properties. This hypothesis is widely adopted by conventional experimentalists, who usually compare a given protein to homologous modified proteins by sequence alignment. If a conserved peptide is observed around aligned modified residue, they may obtain confidence that the peptide in the given protein can also be modified. We borrowed this hypothesis and implemented it into an automatic algorithm. First, we used the amino acid substitution matrix BLOSUM62 to evaluate the similarity between two *phosphorylation site peptides* PSP(*m, n*) with *m* residues upstream and *n* residues downstream flanking the phosphorylated site, while *m* and *n* were arbitrarily determined as 3 for phosphorylation site peptide (Xue et al., 2005; Zhou et al., 2004). In GPS (Group-based Phosphorylation Scoring) 1.0 and 1.1, we clustered the phosphorylated peptides with Markov cluster algorithm (MCL for short) with an additional hypothesis of that one protein kinase (PK) can recognize more than one motif in substrates (Xue et al., 2005; Zhou et al., 2004). Later, based on the observation of different matrix generating different performance, we developed the matrix mutation (MaM) approach for the performance improvement in GPS 2.0, which was refined as Group-based Prediction System (Xue et al., 2008). In GPS 2.0, the MCL clustering was discarded for its low efficiency while the informative peptide was selected as PSP(7, 7) (Xue et al., 2008). For the prediction of sumoylation (Ren et al., 2009) and palmitoylation sites (Ren et al., 2008), we testingly classified modification sites based on known linear motifs together with GPS 2.0 algorithm, and achieved increased performances. Later, we improved the algorithm to version 2.1 for the prediction of phosphorylation sites with a additional motif length selection method (MLS) (Xue et al., 2011). Recently, with two additional approaches of *k*-means clustering and

weight training (WT), we further designed GPS 3.0 algorithm for the prediction of *S*-nitrosylation (Xue et al., 2010b) and nitration sites (Liu et al., 2011). The details of GPS series algorithms were described below.

2.1 GPS 1.x algorithms

The GPS 1.x algorithms include two versions of GPS 1.0 and GPS 1.1. In 2004, we initially designed the GPS 1.0 algorithm for the prediction of kinase-specific phosphorylation sites, while the full name of GPS was "Group-based Phosphorylation Scoring" (Zhou et al., 2004). From public databases and literature curation, we collected 2,001 experimentally identified phosphorylation sites with their cognate PKs (Zhou et al., 2004). Since similar PKs can modify similar sequences in the substrates, we clustered the available PKs into 52 PK groups according to the BLAST results and Swiss-Prot/TrEMBL annotations. Then the known phosphorylation sites were classified into one of multiple of the 52 PK groups based on their regulatory PK information (Zhou et al., 2004). Based on the hypothesis of similar short peptides bearing similar biological properties, we designed a simple scoring strategy. Given a *phosphorylation site peptide* PSP(m, n) as a serine (S), threonine (T) or tyrosine (Y) amino acid flanked by m residues upstream and n residues downstream, we employed the amino acid substitution matrix BLOSUM62 to evaluate the similarity of peptides. In GPS 1.0 (Zhou et al., 2004), the m and n were arbitrarily chosen as 3. For two PSP(m, n), we scored the similarity of two PSP(m, n) as:

$$S(A,B) = \sum_{-m \le i \le n} Score(A[i], B[i]) \tag{1}$$

$Score(A[i], B[i])$ represents the substitution score of the two amino acid of $A[i]$ and $B[i]$ in BLOSUM62. If S(A, B)<0, we simply redefined it as S(A, B)=0. With an additional hypothesis that one PK can recognize multiple motifs, we automatically clustered the phosphorylation site peptides into more than one groups with Markov cluster algorithm (MCL for short). Thus, given any peptide for the prediction of kinase-specific phosphorylation sites, we calculated the average score between the peptide and the experimentally identified phosphorylated site peptides in each cluster, while the maximum score among the clusters was decided as the final score. The prediction performance can be comparative with other tools such as Scansite (Obenauer et al., 2003). Later, we slightly refined the algorithm and released GPS 1.1 version, which can predict phosphorylation sites for 216 unique kinases in 71 kinase groups (Xue et al., 2005). Again, the m and n in GPS 1.1 were still selected as 3. The only improvement in GPS 1.1 is that the classification of PKs is better than GPS 1.0 (Xue et al., 2005).

Furthermore, we applied the GPS 1.x algorithm to predict a variety of PTMs such as sumoylation and palmitoylation, with additional refinement if necessary (Xue et al., 2006b; Zhou et al., 2006a). For the prediction of sumoylation sites in SUMOsp 1.0 (Xue et al., 2006b), both GPS 1.x and Motif-X algorithms (Schwartz and Gygi, 2005) were employed because a large proportion of sumoylation sites follow a consensus motif ψ-K-X-E (ψ is a hydrophobic amino acid) or ψ-K-X-E/D (Johnson, 2004). Thus, all known sumoylation sites were classified into two groups with consensus and non-consensus. A given peptide will be compared to known sumoylation sites of both two groups by calculating the average similarity scores, respectively. The maximum score was decided as the final score. And if the score is higher than a pre-determined threshold, the peptide will be predicted as potential sumoylation site (Xue et al., 2006b). In SUMOsp 1.0, the sumoylation site peptide for the prediction was arbitrarily selected as SSP(7, 7). For the prediction of palmitoylation sites in

CSS-Palm 1.0 (Zhou et al., 2006a), the palmitoylation site peptide was casually chosen as PSP(7, 7). Because no common canonical consensus sequence/motif for palmitoylation was reported, we developed a BLOSUM62-based Clustering method (BBC) based on the graph theory, and classified all known palmitoylation sites into three clusters (Zhou et al., 2006a).

2.2 GPS 2.x algorithms

The GPS 1.x algorithms were too preliminary, while a variety of issues were not addressed. By personal communications, several researchers asked us a number of questions. For example, why we arbitrarily chose BLOSUM62 rather than other amino acid substitution matrices? Why we classified the PKs based on BLAST searching rather than using pre-established classification information? Why we selected PSP(3, 3) or PSP(7, 7)? The aim of GPS 2.x algorithms was to resolve these problems.

To evaluate the prediction performance and robustness of a predictor, we usually preformed the self-consistency validation, the leave-one-out validation (LOO) and *n*-fold cross-validations. The self-consistency validation used the training positive data (+) and negative data (-) directly to evaluate the prediction performance, and represented the computational power of the prediction system. However, the prediction system might be overtrained and only perfect for the training data set, with low prediction ability for new data. In this regard, the LOO validation and *n*-fold cross-validations should be performed to evaluate the robustness and the stability on an independent data set. In the LOO validation, each site in the data set was picked out in turn as an independent test sample, and all the remaining sites were regarded as training data. This process was repeated until each site was used as test data one time. In *n*-fold cross-validations, all the (+) sites and (-) sites were combined and then divided equally into *n* parts, keeping the same distribution of (+) and (-) sites in each part. Then *n*-1 parts were merged into a training data set while the remnant part was taken as a testing data set. This process was repeated 20 times and the average performance of *n*-fold cross-validations was used to estimate the performance. In our previous study, when the training data set is large enough (number of positive sites ≥ 30), the results of *n*-fold cross-validations are similar with the LOO result. In this regard, we merely used the LOO validation to evaluate the robustness and stability.

In GPS 1.x algorithms, the amino acid substitution matrix was arbitrarily chosen as BLOSUM62, while performances of other matrices were not evaluated. For the sake of better performance, we tested other matrices such as BLOSUM30, 45, 62, 90 and PAM10, 90, 250, 500, and found different matrices could generate various performances (Xue et al., 2008). For example, PAM10-based scoring can easily generate a perfect self-consistency result with sensitivity (*Sn*) of 100% and specificity (*Sp*) of 100%, while the LOO result is very poor that denotes the prediction model is highly over-fitting and unstable. In this regard, a key challenge is that whether we can obtain an optimal or near-optimal matrix with the highest LOO values. To address this issue, we developed GPS 2.0 algorithm with an additional approach of matrix mutation (MaM) (Xue et al., 2008). First BLOSUM62 matrix was chosen as the initial matrix. The performance (*Sn* and *Sp*) of LOO validation was calculated. For the prediction of kinase-specific phosphorylation sites, we fixed *Sp* at 90% to improve *Sn* by randomly picking out an element of the matrix for +1 or -1. The procedure was terminated when the *Sn* value was not increased any further. Although matrix mutation in other types was also valid, the MaM strategy can improve the LOO result significantly, whereas the self-consistency was only influenced moderately. Thus, such a procedure made the predictor more robust and stable. By comparison, the GPS 2.0 exhibited superior performance against other analogous tools (Xue et al., 2008).

In GPS 1.x algorithms, numerous PKs were casually classified into several groups simply based on sequence comparison by BLAST, based on the hypothesis that PKs in a same group/subfamily will recognize similar sequence patterns of substrates for modification (Xue et al., 2005; Zhou et al., 2004). Because the kinomes of several eukaryotic organisms have been comprehensively identified, phylogenetically analyzed, and classified into a hierarchical structure, including group, family, subfamily, and single PK (Caenepeel et al., 2004; Manning et al., 2002), and because most of the phosphorylation sites in the public database have been experimentally verified in mammals (~97.6%), we adopted the well established rule for human PK classification (Xue et al., 2008) in GPS 2.0 to cluster various PKs with their verified sites into a hierarchical structure with four levels, including group, family, subfamily, and single PK. The PK groups with less than three sites were singled out (Xue et al., 2008). The training data could be reused several times and included in different PK clusters based on their known PK information. GPS 2.0 can predict kinase-specific phosphorylation sites for 408 human PKs in hierarchy.

In GPS 2.0, the PSP(7, 7) was arbitrarily determined (Xue et al., 2008). Later, we carefully studied how different combinations of PSP(*m*, *n*) influenced prediction performance and robustness (Xue et al., 2011). The self-consistency validation and LOO validation were thoroughly carried out for each PK group. We observed that the self-consistency results will be always increased with longer PSP(*m*, *n*). However, when the phosphorylated peptide was elongated, the LOO results will first reach a peak value then decrease. In this regard, we developed GPS 2.1 algorithm with an additional approach of motif length selection (MLS) (Xue et al., 2011), which could automatically detect the optimal length of PSP(*m*, *n*) with the highest LOO performance. We exhaustively tested all combinations of PSP(*m*, *n*) (*m* = 1,…, 15; *n* = 1, . . ., 15). The *Sn* values were calculated under the *Sp* of 85, 90 and 95%. Then the average *Sn* was calculated as the final *Sn* value. By comparing to GPS 2.0 software (Xue et al., 2008), the average *Sn* of the LOO was significantly increased by 15.62%, whereas the average *Sn* value of the self-consistency was slightly reduced by 2.28%. The *Sp* score was fixed at 90% for comparison. In this regard, the MLS method could efficiently narrow down the difference between the LOO validation and self-consistency validation to improve the robustness of prediction system (Xue et al., 2011).

The GPS 2.1 algorithm was also adopted for the prediction of sumoylation and palmitoylation sites, with additional improvements (Ren et al., 2009; Ren et al., 2008). The experimentally identified sumoylation sites were classified into two types including Type I (consensus) and Type II (non-consensus) sites in SUMOsp 2.0 (Ren et al., 2009). Type I sites followed the ψKXE (ψ is A, I, L, M, P, F, or V and X is any amino acid residue) motif, while Type II sites contained other non-canonical sites. Also, we clustered known palmitoylation sites in CSS-Palm 2.0 into three groups, including Type I (sites follow a –CC– motif, C is a cysteine residue), Type II (sites follow a –CXXC– motif, C is a cysteine residue and X is a random residue) and Type III (other sites) group (Ren et al., 2008). In SUMOsp 1.0 and CSS-Palm 1.0, the threshold is the same for different clusters. However, for SUMOsp 2.0 and CSS-Palm 2.0, we set different threshold for each group, separately. The prediction performance of GPS 2.x is much better than GPS 1.x (Ren et al., 2009; Ren et al., 2008; Xue et al., 2008).

2.3 GPS 3.0 algorithm

Although GPS 1.x and 2.x algorithms were successfully applied in the prediction of phosphorylation, sumoylation and palmitoylation sites, they exhibited poor performance for

other PTMs, such as *S*-nitrosylation (Xue et al., 2010b) and nitration sites (Liu et al., 2011). Thus, the GPS algorithm still need further improvements.

We hypothesized that one type of PTM can recognize multiple sequence patterns/motifs. If this hypothesis is correct, the prediction performance can be enhanced by clustering known PTM sites into multiple groups. In GPS 1.x, the MCL algorithm was adopted to automatically classify the known phosphorylation site peptides into multiple clusters if available. However, only eight PK groups obtained more than one cluster (Xue et al., 2005; Zhou et al., 2004). In CSS-Palm 1.0, we adopted a graph-based BBC method for clustering known palmitoylation sites, but it can not significantly improve performance for other types of PTMs (unpublished). In this regard, the clustering strategy was dropped for its low efficiency in GPS 2.0 for the prediction of phosphorylation sites (Xue et al., 2008). For the prediction of sumoylation (SUMOsp 2.0) and palmitoylation (CSS-Palm 2.0) sites, we clustered known sumoylation and palmitoylation into two and three groups based on reported motifs, although the palmitoylation motifs are much weak (Ren et al., 2009; Ren et al., 2008). However, for *S*-nitrosylation and nitration sites, even very weak motifs are not available (Liu et al., 2011; Xue et al., 2010b). In this regard, an interesting question is how to classify PTM sites without any obvious motifs? To address this problem, we developed GPS 3.0 algorithm with an additional *k*-mean clustering method (Liu et al., 2011; Xue et al., 2010b), which was extensively used in a variety of aspects (Herwig et al., 1999; Yoon et al., 2007). With the algorithm, we successfully classified 504 experimentally identified *S*-nitrosylation into 3 groups in GPS-SNO 1.0 (Xue et al., 2010b), while the 1,066 known nitration sites were clustered into 5 groups in GPS-YNO2 1.0 (Liu et al., 2011).

Again, in GPS 1.x and GPS 2.x, the contribution of each residue for substrate recognition by enzymes was regarded as equal. However, there were various amino acid preferences in the residues around the phosphorylation sites for different PKs (Schwartz and Gygi, 2005). For example, the substrates of CDKs follow a pS-P-X-K motif (pS is the phosphorylated serine), which indicates that the adjacent proline is critical for the CDK-specific phosphorylation (Schwartz and Gygi, 2005). Furthermore, the glutamine residue adjacent to the serine/threonine (S/T-Q) was found to be important for ATM (ataxia telangiectasia mutated)/ATR (ATM and Rad3-related) recognition (Matsuoka et al., 2007). In this regard, the different contributions of distinct positions around the PTM sites should be considered and included in the computational model. In this regard, an additional approach of weight training (WT) was added in GPS 3.0 algorithm. We optimized the weight of each position in the *PTM site peptide* PSP(*m*, *n*) for every cluster according to the leave-one-out performance (Liu et al., 2011; Xue et al., 2010b).

Together with MaM and MLS approaches, we determined the order of training processes to be: *k*-means clustering, MLS, WT and MaM. By exhaustively testing, it was found that this training order cannot be changed (Liu et al., 2011; Xue et al., 2010b). The prediction performance of GPS 3.0 is much better than GPS 1.x and GPS 2.x algorithms. The GPS 3.0 was firstly introduced and described in the construction of GPS-SNO 1.0 and GPS-YNO2 1.0 (Liu et al., 2011; Xue et al., 2010b). Below, we used palmitoylation as an example to depict the implementation process in detail.

3. An application: Prediction of palmitoylation sites with GPS 3.0 algorithm

In order to describe the GPS series algorithms thoroughly, here we employed the GPS 3.0 algorithm to predict palmitoylation sites as an example. Palmitoylation is the only type of

reversible lipid modification, and dynamically regulates protein trafficking and functions through addition of saturated 16-carbon palmitic acids to specific cysteine residues by DHHC palmitoyltransferases (Fukata and Fukata, 2010; Linder and Deschenes, 2007). First, we manually collected the experimentally identified palmitoylation sites from scientific literatures in PubMed. Redundant homologous sites were cleared, while the positive and negative data sets were prepared. The procedures of performance improvement with an order of *k*-means clustering, MLS, WT and MaM were described in detail. Finally, the CSS-Palm 3.0 software packages were implemented in JAVA. The full process of CSS-Palm 3.0 construction is shown in Fig. 1.

3.1 Data preparation

Previously, we manually collected the experimental identified palmitoylation sites from scientific literature which was published before October 8th, 2007 (Ren et al., 2008). Since a large number of experimental studies were reported after CSS-Palm 2.0 was developed, here we further searched the literature in PubMed with the keywords of "palmitoylation" and "palmitoylated" to obtain additional verified palmitoylation sites (before February 14th, 2010). The protein sequences were retrieved from the UniProt database (UniProt, 2010).

In general, if the training data set is highly redundant with too many homologous sites, the prediction accuracy will be overestimated. To avoid such overestimation, we clustered the protein sequences with a threshold of 40% identity by CD-HIT (Li and Godzik, 2006). If two proteins were similar with ≥40% identity, we re-aligned the proteins with BL2SEQ, a program in the BLAST package (Johnson et al., 2008), and checked the results by hand. If two palmitoylation sites from two homologous proteins were at the same position after sequence alignment, only one item was preserved, the other was discarded. Finally, the non-redundant benchmark data set for training contained 439 positive sites from 194 unique substrates.

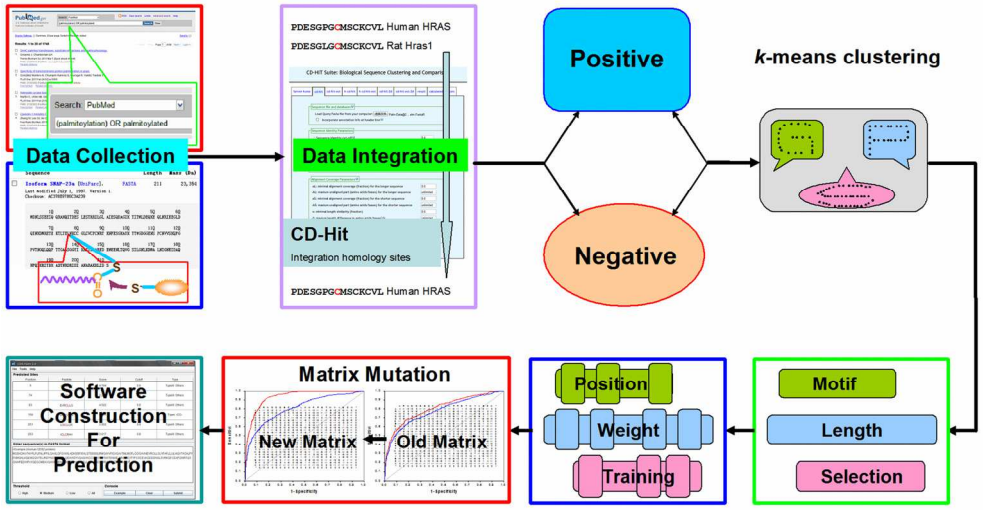


Fig. 1. The full procedures of constructing CSS-Palm 3.0 for the prediction of palmitoylation sites.

We defined a *palmitoylation site peptide* PSP(*m*, *n*) as a palmitoylated cysteine flanked by *m* residues upstream and *n* residues downstream. As previously described (Ren et al., 2008; Zhou et al., 2006a), we regarded all experimentally verified palmitoylation sites as positive data (+), while all other non-palmitoylated cysteine sites in the same substrates were taken as negative data (-). If a palmitoylated cysteine locates at the N- or C-terminus of the protein and the length of the peptide is smaller than *m*+*n*+*1*, we added one or multiple "*" characters as pseudo amino acids to complement the PSP(*m*, *n*). Finally, we got 439 positive sites and 2,171 negative sites.

3.2 The GPS 3.0 algorithm

The GPS 3.0 algorithm contains two major components of scoring strategy and performance improvement.

Given the hypothesis of similar short peptides bearing similar biochemical properties, the similarity between two PSP(*m*, *n*) of *A* and *B* can be calculated with equation (1). Again, if S (*A*, *B*) <0, we simply redefined it as S (*A*, *B*)=0. A putative PSP(*m*, *n*) is compared with each of the experimentally verified palmitoylated peptides in a pairwise manner to calculate the similarity score. The average value of the substitution scores is regarded as the final score. The schematic description of the scoring strategy with examples was shown in Fig. 2. The performance improvement processes with of four sequential steps of k-means clustering, MLS, WT and MaM were presented below.

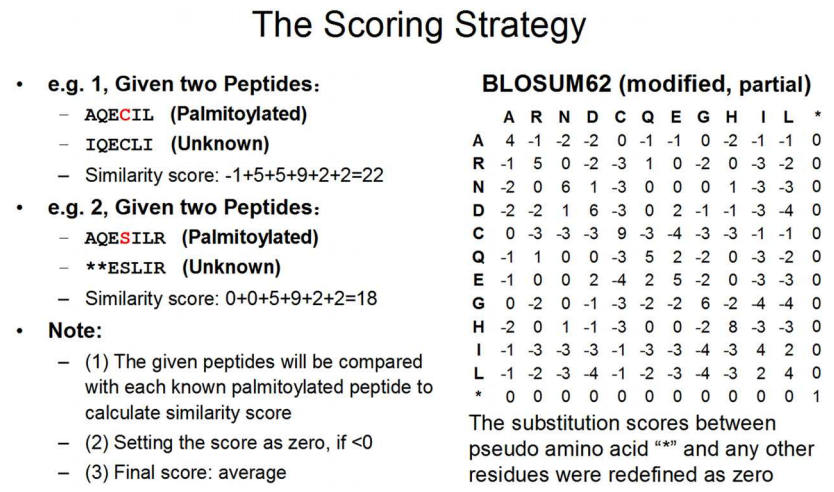


Fig. 2. Schematic description of the scoring strategy in GPS algorithm.

3.2.1 *k*-means clustering

In CSS-Palm 2.0, we clustered experimental identified palmitoylation sites into three groups based on known motifs (Ren et al., 2008). Because the palmitoylation motifs are very weak, the prediction performance was only considerably improved (Ren et al., 2008). For grouping known palmitoylation sites, here we used the *k*-mean clustering approach, which was widely adopted and got successful performance in our previous studies (Liu et al., 2011; Xue et al., 2010b). The clustering process was described as below:

Given two PSP(m, n) peptides A and B, the similarity was defined and measured as: $s(A, B) = N_s/N$. The N is the number of all substitutions, whereas the N_s is the number of conserved substitutions with $Score(a, b) > 0$ in the BLOSUM62 matrix. The $s(A, B)$ ranges from 0 to 1. Thus, the distance between them can be defined as: $D(A, B) = 1/s(A, B)$. If $s(A, B) = 0$, $D(A, B) = \infty$. By exhaustive testing, the k was roughly set to 3, while PSP(7, 7) was adopted. First, three palmitoylation sites from the positive data (+) were randomly chosen as the centroids. Second, the other positive sites were compared in a pairwise manner with the three centroids and clustered into groups with the highest similarity values. Third, the centroid of each cluster was updated with the highest average similarity (HAS). The second and third steps were iteratively repeated until the clusters did not change any longer. After the three clusters for the positive sites had been determined, we put each negative site into the cluster with the HAS.

Given a potential PSP(7, 7) for prediction, we firstly determined which cluster it belongs to, by calculating the average similarity score of the PSP(7, 7) against each cluster (Fig. 3). For example, the PSP(7, 7) P_1 will be regarded as Cluster 1 type site, while the P_2 and P_3 will be determined to be Cluster 2 and 3 type sites, respectively (Fig. 3).

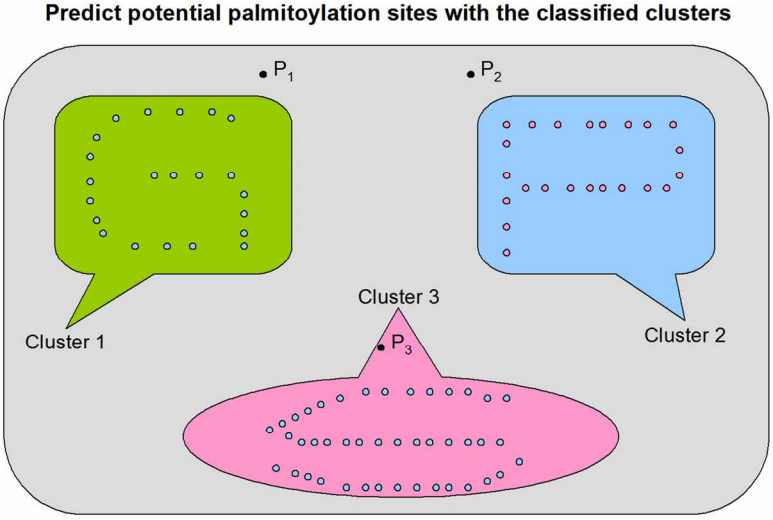


Fig. 3. Prediction of potential palmitoylation sites with the classified clusters.

3.2.2 Motif length selection (MLS)

Previously, the m and n in PSP(m, n) was arbitrarily determined (Xue et al., 2008; Zhou et al., 2004; Zhou et al., 2006a). In this step, we determined the optimized combination of PSP(m, n) for optimal performance. The combinations of PSP(m, n) (m = 1, …, 30; n = 1, …, 30) were extensively tested. The optimal PSP(m, n) for each cluster was separately selected, with the highest LOO performance. From our previous experience, a higher Sp value is more important than a higher Sn to avoid too many false positive hits (Ren et al., 2009; Ren et al., 2008; Xue et al., 2005). Thus, to improve the prediction performance and robustness in the region of high Sp is more important than other regions. In this study, we fixed the Sp at 90% to compare Sn values.

3.2.3 Weight training (WT)

We updated the substitution score between two PSP(m, n) peptides A and B as below:

$$S'(A,B) = \sum_{-m \le i \le n} w_i Score(A[i], B[i]) \quad (2)$$

The w_i is the weight of position i. Again, if $S'(A, B)<0$, we redefined it as $S'(A, B)=0$. Initially, the w was chosen as 1 for each position. We randomly picked out the weight of any position for +1 or -1, and adopted the manipulation if the Sn score of the re-calculated LOO result with the Sp fixed at 90% was increased. The process was repeated until convergence was reached.

3.2.4 Matrix mutation (MaM)

Previously, we chose the BLOSUM62 matrix to evaluate the similarity between PSP(m, n). Later, we observed that different matrices generate various performances (Xue et al., 2008). For palmitoylation, we also tested a variety of matrices such as BLOSUM30, 45, 62, 90, and PAM 10, 90, 250 and 500. The self-consistency (red) and LOO (blue) validations were preformed (Fig. 4). To balance the prediction performance and robustness of the prediction system, the BLOSUM62 matrix was adopted as the initial matrix in CSS-Palm 3.0.

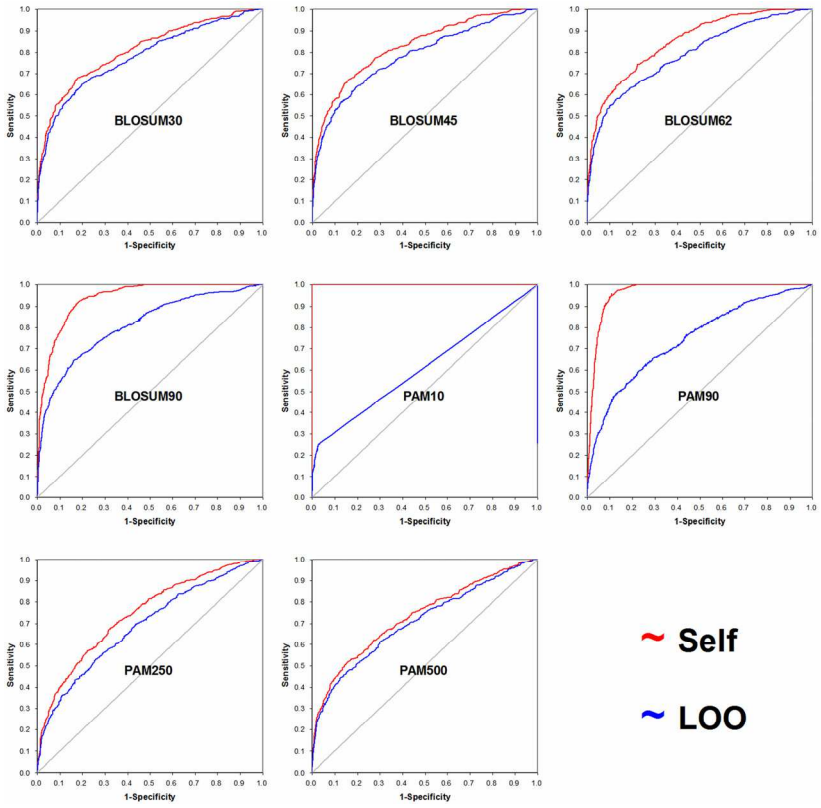


Fig. 4. Different matrices generate various performances. The ROC curves of self-consistency and LOO validations were drawn in red and blue, respectively.

In the MaM process, the LOO result with BLOSUM62 was first calculated. In BLOSUM62, the substitution score between "*" and other residues is -4 but redefined as 0. Then we fixed the *Sp* at 90% to improve *Sn* by randomly picking out one value from the BLOSUM62 matrix for mutation (+1 or -1). If the *Sn* value increased, the mutation was adopted. This process was terminated when the *Sn* value was not increased any further. Interestingly, we observed that when the training time of MaM is long enough, the mutated matrix generated from other matrices, e.g., BLOSUM45, is exactly identical to the one from BLOSUM62 (Data not shown). In this regard, the final mutated matrix is not dependent on the initial matrix.

3.3 Software construction

Previously, we only developed the online services for the predictions with PHP/PERL. We also discussed the general user interface for the predictors of PTMs sites (Zhou et al., 2006b). When the number of users becomes large, the server will bear high burden with very low speed. In this regard, we recently constructed computational tools in JAVA, whereas the online service and local packages were both provided (Liu et al., 2011; Ren et al., 2008; Ren et al., 2009; Xue et al., 2008; Xue et al., 2010b; Xue et al., 2011). The online service and local packages of CSS-Palm 3.0 were implemented in JAVA. For the online service, we tested the CSS-Palm 3.0 on a variety of internet browsers, including Internet Explorer 6.0, Netscape Browser 8.1.3 and Firefox 2 under the Windows XP Operating System (OS), Mozilla Firefox 1.5 of Fedora Core 6 OS (Linux), and Safari 3.0 of Apple Mac OS X 10.4 (Tiger) and 10.5 (Leopard). For the Windows and Linux systems, the latest version of Java Runtime Environment (JRE) package (JAVA 1.4.2 or later versions) of Sun Microsystems should be pre-installed. However, for Mac OS, CSS-Palm 3.0 can be directly used without any additional packages. For convenience, we also developed local packages of CSS-Palm 3.0, which worked with the three major Operating Systems, Windows, Linux and Mac. The software and the online sever are freely available at: http://csspalm.biocuckoo.org/.

3.4 Performance evaluation and comparison

As previously described (Ren et al., 2008; Zhou et al., 2006a), We adopted four standard measurements, including accuracy (*Ac*), sensitivity (*Sn*), specificity (*Sp*) and Mathew correlation coefficient (*MCC*). *Ac* illustrates the correct ratio between both positive (+) and negative (-) data sets, while *Sn* and *Sp* represent the correct prediction ratios of positive (+) and negative data (-) sets respectively. However, when the number of positive data and negative data differ too much from each other, *MCC* should be included to evaluate the prediction performance. 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 CSS-Palm 3.0, the real positives are defined as true positives (*TP*), while the others are defined as false positives (*FP*). Among the data with negative predictions by CSS-Palm 3.0, the real positives are defined as false negatives (*FN*), while the others are defined as true negatives (*TN*). The performance measurements of *Ac*, *Sn*, *Sp* and *MCC* are defined as below:

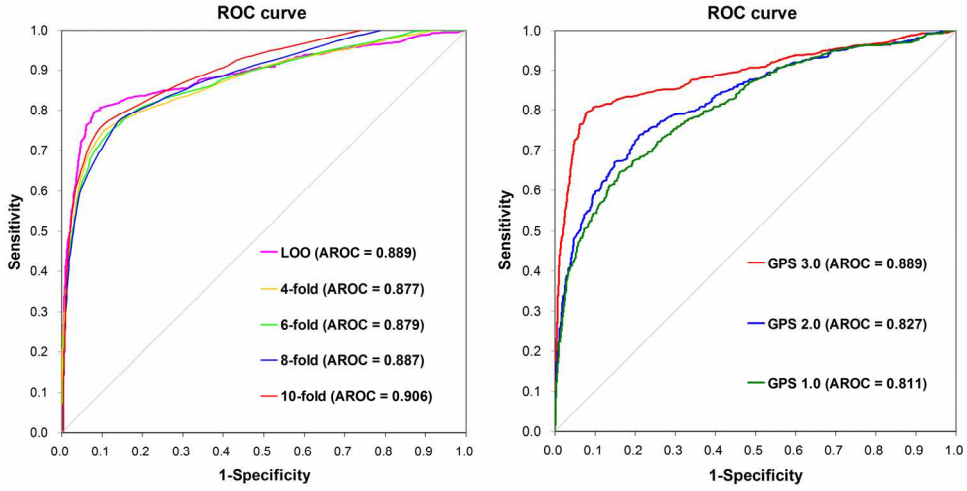
$$Sn = \frac{TP}{TP + FN} \tag{3}$$

$$Sp = \frac{TN}{TN + FP} \tag{4}$$

$$Ac = \frac{TP + TN}{TP + FP + TN + FN} \tag{5}$$

$$MCC = \frac{(TP \times TN) - (FN \times FP)}{\sqrt{(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)}} \tag{6}$$

To avoid overestimation, the LOO validation and 4-, 6-, 8-, 10-fold cross-validations were performed to evaluate the prediction robustness and performance of CSS-Palm 3.0. Receiver Operating Characteristic (ROC) curves are presented in Fig. 5A, while the AROCs (area under ROCs) were calculated as 0.889 (leave-one-out), 0.877 (4-fold), 0.879 (6-fold), 0.887 (8-fold) and 0.906 (10-fold), respectively (Fig. 5A). Since the 4-, 6-, 8-, 10-fold cross-validations were close to the leave-one-out validation, it was demonstrated that CSS-Palm 3.0 is a robust predictor of palmitoylation sites with promising performance.



(A) (B)

Fig. 5. Performance evaluation and comparison. (A) The LOO validation and 4-, 6-, 8-, 10-fold cross-validations were performed to evaluate the prediction robustness and performance of CSS-Palm 3.0. (B) Comparison of LOO results of GPS 1.0, GPS 1.0 and GPS 3.0 algorithms.

For comparison, we also calculated the performance of GPS 1.0 and 2.0 algorithms. To avoid any bias, the same training data set used in GPS 3.0 was also employed in the two methods. The LOO validations were carried out for GPS 1.0 and 2.0 algorithms, while the ROC curves were drawn (Fig. 5B). The AROC values were calculated as 0.811 (GPS 1.0), 0.827 (GPS 2.0) and 0.889 (GPS 3.0), respectively. In addition, we fixed the *Sp* values and compared *Sn* scores (Table 2). When the *Sp* value was ~85%, the *Sn* values of GPS 3.0, GPS 2.0 and GPS 1.0 were 72.44%, 48.29% and 43.05%, respectively (Table 2). Also, when the *Sp* value was ~90%, the *Sn* values of GPS 3.0, GPS 2.0 and GPS 1.0 were 80.64%, 59.68% and 54.21%, separately (Table 2). In addition, when the *Sp* value was ~95%, the *Sn* of GPS 3.0 (82.23%) was still much better than GPS 2.0 (67.20%) and GPS 1.0 (62.41%) (Table 2). Taken together, our results exhibited that the performance of GPS 3.0 is better than GPS 1.0 and 2.0. Finally, the

CSS-Palm 3.0 software was constructed with three thresholds of High, Medium and Low, with the *Sp* values of ~95%, ~90% and ~85%, respectively.

Method	Threshold	*Ac*	*Sn*	*Sp*	*MCC*
GPS 3.0	High	91.30%	72.44%	95.12%	0.6850
	Medium	88.74%	80.64%	90.37%	0.6458
	Low	84.75%	82.23%	85.26%	0.5749
GPS 2.0		87.24%	48.29%	95.12%	49.64%
		84.94%	59.68%	90.05%	48.09%
		82.03%	67.20%	85.03%	45.90%
GPS 1.0		86.28%	43.05%	95.03%	0.4485
		84.21%	54.21%	90.28%	0.4410
		81.23%	62.41%	85.03%	0.4220

Table 2. For comparison, we fixed the *Sp* values of GPS 3.0 algorithm so as to be similar or identical to GPS 1.0 and 2.0 algorithms, and compared the *Sn* values.

4. Conclusion

During the past several decades, accumulated experimental studies have made slow but steady contributions toward understanding molecular mechanisms and regulatory roles of various PTMs (Mann and Jensen, 2003; Walsh, 2005; Walsh and Jefferis, 2006). Recently, rapid progresses in the state-of-the-art HTP-MS techniques have boomed an explosion of modification data for systematically analyzing PTM regulation in a proteomic level (Choudhary and Mann, 2010). However, the biological functions of PTMs are still far from fully elucidated. In this regard, more efforts remain to be carried out.

In contrast with expensive and error-prone experimental methods, *in silico* prediction of PTM-specific substrates with their sites has emerged as a popular alternative approach. In this field, two questions should be addressed: 1) How to predict modification sites in a given protein sequence? 2) How to predict regulatory enzyme information of modification sites in a given protein sequence? The importance of the two questions is different for distinct types of PTMs. For example, a phosphoproteomics analysis can detect thousands of phosphorylation sites in a single experiment (Olsen et al., 2006; Villen et al., 2007). In this regard, the prediction of general or non-specific phosphorylation sites is not much useful at the current stage. However, there are only ~3,500 phosphorylation sites with known upstream PK information in the public databases (Xue et al., 2008; Xue et al., 2011). In this regard, the prediction of kinase-specific phosphorylation sites is still a great challenge, while the results can be a help for further experimental consideration. For sumoylation and palmitoylation, accurately large-scale identification of their substrates and sites is not easy to be performed. In this regard, the prediction of general sumoylation and palmitoylation in proteins is useful for guiding further experimental verifications. Also, since the experimentally identified enzyme-specific information for both sumoylation and palmitoylation is quite limited, the prediction of enzyme-specific sumoylation or palmitoylation is still not available due to data limitation.

Intuitively, the prediction of PTM sites seems to be a trivial job. Assume that one may easily obtain experimentally identified PTMs from one or two review articles as the training data set, casually select a machine learning algorithm such as PSSM, SVMs or ANN, carry out

several validations to evaluate the performance, and develop a web server for the prediction. Then the manuscript can be written with a cup of coffee in hand. Previously, a number of researchers asked us by personal communications that why we did not used a simple existed algorithm to develop an integrate tool for the prediction of all types of PTMs sites. Is the prediction of PTMs sites really simple? From our research experience, the answer is "not at all". First, most of widely-used machine learning algorithms are derived from the fields of informatics or statistics and originally designed for general propose, but not specifically for PTMs sites prediction. Second, different types of PTMs can have distinct sequence features. One algorithm can generate promising performance for a specific PTM but exhibit poor accuracy for other types of PTMs. For example, the prediction of PKA-specific phosphorylation sites with any algorithm can generate satisfying performance (Xue et al., 2008). However, for PTMs with strong motifs, the scenario is different. For example, the sumoylation has a strong consensus motif of ψKXE, which about 77% of all known sumoylation sites follow this pattern (Xue et al., 2006b). Since the simple strong motif can generate great accuracy, development of new algorithms will not be necessary if the performance can not be significantly improved. For palmitoylation, two weak motifs can be obtained (Ren et al., 2008). Prediction of palmitoylation with weak motifs will generate poor performance. But it's also difficult for computational algorithm to retrieve informative features for prediction. In addition, for S-nitrosylation and nitration, even weak motifs are not available (Liu et al., 2011; Xue et al., 2010b). In this regard, development a novel and useful algorithm specifically for PTMs site prediction is an urgent demand. Also, great attention needs to be paid since different PTMs have different properties.

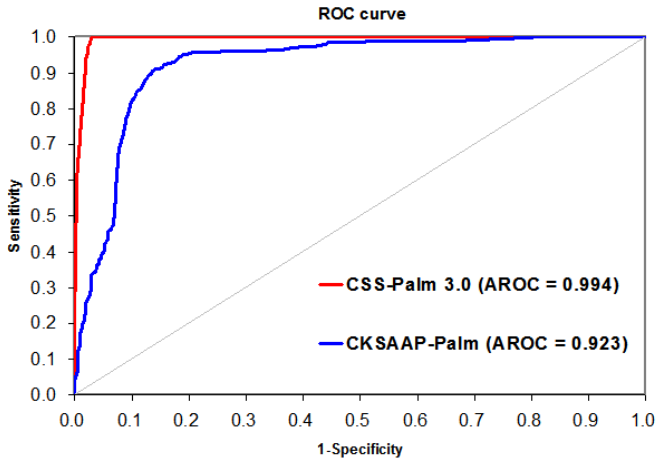


Fig. 6. Comparison of CSS-Palm 3.0 and CKSAAP-Palm (Wang et al., 2009).

During the past several years, our group take great efforts on designing and improve the GPS series algorithms, which were specifically designed for the prediction of PTM sites in proteins. Although the preliminary GPS 1.x algorithms could be only comparative to other approaches, the GPS 2.x exhibited superior performance against analogous predictors. The latest version of GPS 3.0 algorithm has been further improved and much better than our GPS 2.x algorithms. As an application, we used this algorithm to construct CSS-Palm 3.0 for

the prediction of palmitoylation sites. By comparison of a recently released predictor of CKSAAP-Palm (Wang et al., 2009), the performance of CSS-Palm 3.0 is significantly better (Fig. 6).

Finally, we do not propose that the GPS 3.0 will be the final version, while more strategies will be developed and included in GPS series algorithms. We anticipated that the combination of computational predictions and experimental verifications will become the foundation of systematically understanding the mechanisms and the dynamics of PTMs.

5. Acknowledgment

This work was supported by grants from the National Basic Research Program (973 project) (2010CB945400, 2011CB910400), National Natural Science Foundation of China (90919001, 31071154, 30900835, 30830036, 91019020, 21075045), and Fundamental Research Funds for the Central Universities (HUST: 2010JC049, 2010ZD018; SYSU: 11lgzd11, 11lgjc09).

6. References

Ackermann, B.L., & Berna, M.J. (2007). Coupling immunoaffinity techniques with MS for quantitative analysis of low-abundance protein biomarkers. *Expert Review of Proteomics*, Vol.4, No.2, (April 2007), pp. 175-186, ISSN 1744-8387

Balter, M. & Vogel, G. (2001). Nobel prize in physiology or medicine. Cycling toward Stockholm. *Science*, Vol.294, No.5542, (October 2001), pp. 502-503, ISSN 0036-8075

Blom, N., Sicheritz-Ponten, T., Gupta, R., Gammeltoft, S. & Brunak, S. (2004). Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. *Proteomics*, Vol.4, No.6, (June 2004), pp. 1633-1649, ISSN 1615-9853

Boschetti, E. & Righetti, P.G. (2009). The art of observing rare protein species in proteomes with peptide ligand libraries. *Proteomics*, Vol.9, No.6, (March 2009), pp. 1492-1510, ISSN 1615-9853

Caenepeel, S., Charydczak, G., Sudarsanam, S., Hunter, T. & Manning, G. (2004). The mouse kinome: discovery and comparative genomics of all mouse protein kinases. *Proceedings of the National Academy of Sciences of the United States of America*, Vol.101, No.32, (August 2004), pp. 11707-11712, ISSN 0027-8424

Choudhary, C., Kumar, C., Gnad, F., Nielsen, M.L., Rehman, M., Walther, T.C., Olsen, J.V. & Mann, M. (2009). Lysine acetylation targets protein complexes and co-regulates major cellular functions. *Science*, Vol.325, No.5942, (August 2009), pp. 834-840, ISSN 0036-8075

Choudhary, C. & Mann, M. (2010). Decoding signalling networks by mass spectrometry-based proteomics. *Nature Reviews Molecular Cell Biology*, Vol.11, No.6, (June 2010), pp. 427-439, ISSN 1471-0080

Dang, T.H., Van Leemput, K., Verschoren, A. & Laukens, K. (2008). Prediction of kinase-specific phosphorylation sites using conditional random fields. *Bioinformatics*, Vol.24, No.24, (December 2008), pp. 2857-2864, ISSN 1367-4811

Fukata, Y. & Fukata, M. (2010). Protein palmitoylation in neuronal development and synaptic plasticity. *Nature Reviews Neuroscience*, Vol.11, No.3, (March 2010), pp. 161-175, ISSN 1471-0048

Herwig, R., Poustka, A.J., Muller, C., Bull, C., Lehrach, H. & O'Brien, J. (1999). Large-scale clustering of cDNA-fingerprinting data. *Genome Research*, Vol.9, No.11, (November 1999), pp. 1093-1105, ISSN 1088-9051

Huang, H.D., Lee, T.Y., Tzeng, S.W. & Horng, J.T. (2005). KinasePhos: a web tool for identifying protein kinase-specific phosphorylation sites. *Nucleic Acid Research*, Vol.33, Web Server issue, (July 2005), pp. W226- W229, ISSN 1362-4962

Hunter, T. (2007). The age of crosstalk: phosphorylation, ubiquitination & beyond. *Molecular Cell*, Vol.28, No.5, (December 2007), pp. 730-738, ISSN 1097-2765

Jenuwein, T. & Allis, C.D. (2001). Translating the histone code. *Science*, Vol.293, No.5532, (August 2001), pp. 1074-1080, ISSN 0036-8075

Johnson, E.S. (2004). Protein modification by SUMO. *Annual Review of Biochemistry*, Vol.73, (June 2004), pp. 355-382, ISSN 0066-4154

Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., McGinnis, S. & Madden, T.L. (2008). NCBI BLAST: a better web interface. *Nucleic Acid Research*, Vol.36, Web Server issue, (July 2008), pp. W5-W9, ISSN 1362-4962

Kim, J.H., Lee, J., Oh, B., Kimm, K. & Koh, I. (2004). Prediction of phosphorylation sites using SVMs. *Bioinformatics*, Vol.20, No.17, (November 2004), pp. 3179-3184, ISSN 1367-4803

Kresge, N., Simoni, R.D. & Hill, R.L. (2011). The process of reversible phosphorylation: the work of Edmond H. Fischer. *The Journal of Biological Chemistry*, Vol.286, No.3, (January 2011), pp. e1-e2, ISSN 0021-9258

Kumar, N. & Mohanty, D. (2010). Identification of substrates for Ser/Thr kinases using residue-based statistical pair potentials. *Bioinformatics*, Vol.26, No.2, (January 2010), pp. 189-197, ISSN 1367-4811

Lahiry, P., Torkamani, A., Schork, N.J. & Hegele, R.A. (2010). Kinase mutations in human disease: interpreting genotype-phenotype relationships. *Nature Reviews Genetics*, Vol.11, No.1, (January 2010), pp. 60-74, ISSN 1471-0064

Li, W. & Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, Vol.22, No.13, (July 2006), pp. 1658-1659, ISSN 1367-4811

Linder, M.E. & Deschenes, R.J. (2007). Palmitoylation: policing protein stability and traffic. *Nature Reviews Molecular Cell Biology*, Vol.8, No.1, (January 2007), pp. 74-84, ISSN 1471-0080

Liu, Z., Cao, J., Ma, Q., Gao, X., Ren, J. & Xue, Y. (2011). GPS-YNO2: computational prediction of tyrosine nitration sites in proteins. *Molecular BioSystems*, Vol.7, No.4, (January 2011), pp. 1197-1204, ISSN 1742-2051

Mann, M. & Jensen, O.N. (2003). Proteomic analysis of post-translational modifications. *Nature Biotechnology*, Vol.21, No.3, (March 2003), pp. 255-261, ISSN 1087-0156

Manning, G., Whyte, D.B., Martinez, R., Hunter, T. & Sudarsanam, S. (2002). The protein kinase complement of the human genome. *Science*, Vol.298, No.5600, (December 2002), pp. 1912-1934, ISSN 1095-9203

Matsuoka, S., Ballif, B.A., Smogorzewska, A., McDonald, E.R., 3rd, Hurov, K.E., Luo, J., Bakalarski, C.E., Zhao, Z., Solimini, N., Lerenthal, Y., Shiloh, Y., Gygi, S. P. & Elledge, S. J. (2007). ATM and ATR substrate analysis reveals extensive protein networks responsive to DNA damage. *Science*, Vol.316, No.5828, (May 2007), pp. 1160-1166, ISSN 1095-9203

Norvell, A. & McMahon, S.B. (2010). Cell biology. Rise of the rival. *Science*, Vol.327, No.5968, (February 2010), pp. 964-965, ISSN 1095-9203

Obenauer, J.C., Cantley, L.C. & Yaffe, M.B. (2003). Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic Acid Research*, Vol.31, No.13, (July 2003), pp. 3635- 3641, ISSN 1362-4962

Olsen, J.V., Blagoev, B., Gnad, F., Macek, B., Kumar, C., Mortensen, P. & Mann, M. (2006). Global, in vivo & site-specific phosphorylation dynamics in signaling networks. *Cell*, Vol.127, No.3, (November 2006), pp. 635-648, ISSN 0092-8674

Ren, J., Gao, X., Jin, C., Zhu, M., Wang, X., Shaw, A., Wen, L., Yao, X. & Xue, Y. (2009). Systematic study of protein sumoylation: Development of a site-specific predictor of SUMOsp 2.0. *Proteomics*, Vol.9, No.12, (June 2009), pp. 3409-3412, ISSN 1615-9853

Ren, J., Wen, L., Gao, X., Jin, C., Xue, Y. & Yao, X. (2008). CSS-Palm 2.0: an updated software for palmitoylation sites prediction. *Protein Engineering, Design and Selection*, Vol.21, No.11, (November 2008), pp. 639-644, ISSN 1741-0134

Schwartz, D. & Gygi, S.P. (2005). An iterative statistical approach to the identification of protein phosphorylation motifs from large-scale data sets. *Nature Biotechnology*, Vol.23, No.11, (November 2005), pp. 1391-1398, ISSN 1087-0156

The UniProt Consortium (2010). The Universal Protein Resource (UniProt) in 2010. *Nucleic Acid Research*, Vol.38, Database issue, (January 2010), pp. D142-D148, ISSN 1362-4962

Villen, J., Beausoleil, S.A., Gerber, S.A. & Gygi, S.P. (2007). Large-scale phosphorylation analysis of mouse liver. *Proceedings of the National Academy of Sciences of the United States of America*, Vol.104, No.5, (January 2007), pp. 1488-1493, ISSN 0027-8424

Vogel, G. (2004). Nobel Prizes. Gold medal from cellular trash. *Science*, Vol.306, No.5695, (October 2004), pp. 400-401, ISSN 1095-9203

Walsh, C. (2005). *Posttranslational Modification of Proteins: Expanding Nature's Inventory*, Roberts and Co. Publishers, ISBN 978-097-4707-73-0, Colorado, USA

Walsh, G. & Jefferis, R. (2006). Post-translational modifications in the context of therapeutic proteins. *Nature Biotechnology*, Vol.24, No.10, (October 2006), pp. 1241-1252, ISSN 1087-0156

Wang, Q., Zhang, Y., Yang, C., Xiong, H., Lin, Y., Yao, J., Li, H., Xie, L., Zhao, W., Yao, Y. Ning, Z. B. Zeng, R. Xiong, Y. Guan, K. L. Zhao, S. & Zhao, G. P. (2010). Acetylation of metabolic enzymes coordinates carbon source utilization and metabolic flux. *Science*, Vol.327, No.5968, (February 2010), pp. 1004-1007, ISSN 1095-9203

Wang, X.B., Wu, L.Y., Wang, Y.C. & Deng, N.Y. (2009). Prediction of palmitoylation sites using the composition of k-spaced amino acid pairs. *Protein Engineering, Design and Selection*, Vol.22, No.11, (November 2009), pp. 707-712, ISSN 1741-0134

Xue, Y., Gao, X., Cao, J., Liu, Z., Jin, C., Wen, L., Yao, X. & Ren, J. (2010a). A summary of computational resources for protein phosphorylation. *Current Protein & Peptide Science*, Vol.11, No.6, (September 2010), pp. 485-496, ISSN 1875-5550

Xue, Y., Zhou, F., Zhu, M., Ahmed, K., Chen, G. & Yao, X. (2005). GPS: a comprehensive www server for phosphorylation sites prediction. *Nucleic Acid Research*, Vol.33, Web Server issue, (July 2005), pp. W184-W187, ISSN 1362-4962

Xue, Y., Li, A., Wang, L., Feng, H. & Yao, X. (2006a). PPSP: prediction of PK-specific phosphorylation site with Bayesian decision theory. *BMC Bioinformatics*, Vol.7, (March 2006), pp. 163, ISSN 1471-2105

Xue, Y., Zhou, F., Fu, C., Xu, Y. & Yao, X. (2006b). SUMOsp: a web server for sumoylation site prediction. *Nucleic Acid Research*, Vol.34, Web Server issue, (July 2006), pp. W254-W257, ISSN 1362-4962

Xue, Y., Ren, J., Gao, X., Jin, C., Wen, L. & Yao, X. (2008). GPS 2.0, a tool to predict kinase-specific phosphorylation sites in hierarchy. *Molecular & Cellular Proteomics*, Vol.7, No.9, (September 2008), pp. 1598-1608, ISSN 1535-9484

Xue, Y., Liu, Z., Gao, X., Jin, C., Wen, L., Yao, X. & Ren, J. (2010b). GPS-SNO: Computational Prediction of Protein S-Nitrosylation Sites with a Modified GPS Algorithm. *PLoS ONE*, Vol.5, No.6, (June 2010), pp. e11290, ISSN 1932-6203

Xue, Y., Liu, Z., Cao, J., Ma, Q., Gao, X., Wang, Q., Jin, C., Zhou, Y., Wen, L. & Ren, J. (2011). GPS 2.1: enhanced prediction of kinase-specific phosphorylation sites with an algorithm of motif length selection. *Protein Engineering, Design and Selection*, Vol.24, No.3, (March 2011), pp. 255-260, ISSN 1741-0134

Yates, J.R., Ruse, C.I. & Nakorchevsky, A. (2009). Proteomics by mass spectrometry: approaches, advances & applications. *Annual Review of Biomedical Engineering*, Vol.11, (April 2009), pp. 49-79, ISSN 1545-4274

Yoon, S., Ebert, J.C., Chung, E.Y., De Micheli, G. & Altman, R.B. (2007). Clustering protein environments for function prediction: finding PROSITE motifs in 3D. *BMC Bioinformatics*, Vol. 8, Suppl. 4, (June 2007), pp. S10, ISSN 1471-2105

Young, N.L., Plazas-Mayorca, M.D. & Garcia, B.A. (2010). Systems-wide proteomic characterization of combinatorial post-translational modification patterns. *Expert Review of Proteomics*, Vol.7, No.1, (February 2010), pp. 79-92, ISSN 1744-8387

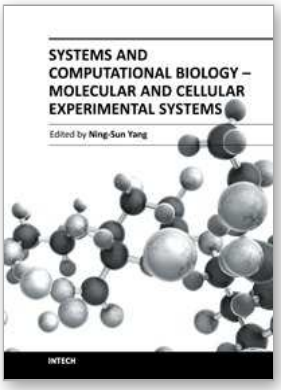
Zhao, S., Xu, W., Jiang, W., Yu, W., Lin, Y., Zhang, T., Yao, J., Zhou, L., Zeng, Y., Li, H., Li, Y., Shi, J., An, W., Hancock, S. M., He, F., Qin, L., Chin, J., Yang, P., Chen, X., Lei, Q., Xiong, Y. & Guan, K. L. (2010). Regulation of cellular metabolism by protein lysine acetylation. *Science*, Vol.327, No.5968, (February 2010), pp. 1000-1004, ISSN 1095-9203

Zhou, F., Xue, Y., Chen, G.L. & Yao, X. (2004). GPS: a novel group-based phosphorylation predicting and scoring method. *Biochemical and Biophysical Research Communications*, Vol.325, No.4, (December 2004), pp. 1443-1448, ISSN 0006-291X

Zhou, F., Xue, Y., Yao, X. & Xu, Y. (2006a). CSS-Palm: palmitoylation site prediction with a clustering and scoring strategy (CSS). *Bioinformatics*, Vol.22, No.7, (April 2006), pp. 894-896, ISSN 1367-4811

Zhou, F., Xue, Y., Yao, X. & Xu, Y. (2006b). A general user interface for prediction servers of proteins' post-translational modification sites. *Nature Protocols*, Vol.1, No.3, (April 2007), pp. 1318-1321, ISSN 1750-2799

Zielinska, D.F., Gnad, F., Wisniewski, J.R. & Mann, M. (2010). Precision mapping of an in vivo N-glycoproteome reveals rigid topological and sequence constraints. *Cell*, Vol.141, No.5, (May 2010), pp. 897-907, ISSN 1097-2765



Systems and Computational Biology - Molecular and Cellular Experimental Systems

Edited by Prof. Ning-Sun Yang

ISBN 978-953-307-280-7
Hard cover, 332 pages
Publisher InTech
Published online 15, September, 2011
Published in print edition September, 2011

Whereas some â€œmicroarrayâ€ or â€œbioinformaticsâ€ scientists among us may have been criticized as doing â€œcataloging researchâ€, the majority of us believe that we are sincerely exploring new scientific and technological systems to benefit human health, human food and animal feed production, and environmental protections. Indeed, we are humbled by the complexity, extent and beauty of cross-talks in various biological systems; on the other hand, we are becoming more educated and are able to start addressing honestly and skillfully the various important issues concerning translational medicine, global agriculture, and the environment. The two volumes of this book presents a series of high-quality research or review articles in a timely fashion to this emerging research field of our scientific community.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Yu Xue, Zexian Liu, Jun Cao and Jian Ren (2011). Computational Prediction of Post-Translational Modification Sites in Proteins, Systems and Computational Biology - Molecular and Cellular Experimental Systems, Prof. Ning-Sun Yang (Ed.), ISBN: 978-953-307-280-7, InTech, Available from: http://www.intechopen.com/books/systems-and-computational-biology-molecular-and-cellular-experimental-systems/computational-prediction-of-post-translational-modification-sites-in-proteins

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821