SysPTM: A Systematic Resource for Proteomic Research on Post-translational Modifications

Hong Li¹,²,³†, Xiaobin Xing¹,²,³†, Guohui Ding¹, Qingrun Li¹,³, Chuan Wang¹, Lu Xie²*, Rong Zeng¹*, Yixue Li¹,²*

1. Key Lab of Systems Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, 320 Yueyang Road, Shanghai 200031, P.R.China
2. Shanghai Center for Bioinformation Technology, 100 Qinzhou Road, Shanghai 200235, P.R.China
3. Graduate School of the Chinese Academy of Sciences, 19 Yuquan Road, Beijing 100039, P.R.China

* Correspondence should be addressed to xielu@scbit.org, zr@sibs.ac.cn, and yxli@sibs.ac.cn.
† These authors contributed equally to this work.

RUNNING TITLE: SysPTM: a data repository for PTM research
**Abbreviations:**

PTM, post-translational modification

PTM protein, protein experimentally identified as having one or more post-translational modifications

MS/MS, tandem mass spectrometry

MDLC, multiple-dimensional liquid chromatography

GO, Gene Ontology

SysPTM-A, PTM sites culled from public resources

SysPTM-B, PTM sites culled from peer-reviewed tandem mass spectrometry literature

PTMBlast, a tool for comparing user PTM datasets with data in SysPTM

PTMPathway, a tool for mapping PTM proteins to KEGG pathways

PTMPhylog, a tool for analyzing potentially conserved PTM sites

PTMCluster, a tool for finding clusters of multi-site modifications
SUMMARY

With the rapid expansion of protein post-translational modification (PTM) research based on large-scale proteomic work, there is an increasing demand for a suitable repository to analyze PTM data. Here we present a curated, web-accessible PTM database-SysPTM. SysPTM provides a systematic and sophisticated platform for proteomic PTM research, equipped not only with a knowledge base of manually curated multi-type modification data, but also with four fully developed, in-depth data mining tools. Currently, SysPTM contains data detailing 117349 experimentally determined PTM sites on 33421 proteins involving nearly 50 PTM types, curated from public resources including five databases and four webservers and more than one hundred peer-reviewed mass spectrometry papers. Protein annotations including Pfam domains, KEGG pathways, GO functional classification, and ortholog groups are integrated into the database. Four online tools have been developed and incorporated, including: PTMBlast, to compare a user’s PTM dataset with PTM data in SysPTM; PTMPathway, to map PTM proteins to KEGG pathways; PTMPhylog, to discover potentially conserved PTM sites; and PTMCluster, to find clusters of multi-site modifications. The workflow of SysPTM was demonstrated by analyzing an in-house phosphorylation dataset identified by MS/MS. It is shown that, in SysPTM, the roles of single-type and multi-type modifications can be systematically investigated in a full biological context. SysPTM could be an important contribution to modificomics research. SysPTM is freely available at: http://www.sysbio.ac.cn/SysPTM or http://lifecenter.sgst.cn/SysPTM.
INTRODUCTION

Post-translational modifications (PTMs) are various processing events that change the maturity, activity, and/or turnover of proteins. More than 200 different types of PTMs have been found, with new ones still being reported (1). PTMs not only change the physicochemical properties of proteins (2), but also dynamically regulate various biological events such as protein degradation, subcellular localization, conformational change, protein-protein interaction, and signal transduction (3-5). Previous studies have revealed the central roles of PTMs in human health and disease. For example, phosphorylation of pRB1 has been associated with tumorigenesis through controlling cell division (6); S-nitrosylation of parkin stimulates its E3 ligase activity, resulting in protein accumulation in sporadic Parkinson's disease (7); and defects in protein glycosylation have been related to several forms of congenital muscular dystrophy (8). Given this important role in health and disease, PTMs have been regarded as potential disease biomarkers or therapeutic targets. For example, Erlotinib (Tarceva), an inhibitor of epidermal growth factor receptor (EGFR) tyrosine kinase, has been approved by the FDA to treat non-small cell lung cancer (9); and histone deacetylase (HDAC) inhibitors have been demonstrated to have a potential therapeutic role in Huntington's disease (10). The broad range of important roles played by PTMs in physiological and pathological processes has made PTM research an active field in recent years. Yet we remain limited in our knowledge of the full scope of PTM distribution on proteins and the precise location of PTM sites.

There are two major kinds of experimental methods to identify PTMs: 1) traditional biological experiments such as radiolabeling PTM proteins (11), western analysis with antibodies against specific modifications (12), and site-directed mutagenesis of potential modification sites (13) and 2) large-scale proteomic experiments, especially multiple-dimensional liquid chromatography tandem mass spectrometry (MDLC-MS/MS). Traditional experiments are laborious and time-consuming, resulting in slow data accumulation. By contrast, more recent MS/MS experiments have led to the discovery of thousands of new phosphorylation (14), glycosylation (15), acetylation (16), sumoylation (17), S-nitrosylation (18), and other modification sites. For example, based on MS/MS data, more than 6000 phosphopeptides have been reported in HeLa cells (14), and 159 candidate sumoylated proteins have been found in yeast (17). Although advanced technologies have allowed PTM data to accumulate rapidly, it is impossible to identify all PTM sites for a set of proteins in
one experiment, due to biased modification enrichment related to experimental protocol, limited sensitivity of mass spectrometer instrumentation, and failures in spectrum matching. Databases are needed to amass PTM data from various experiments for comprehensive understanding of PTMs.

Most databases for storing PTM information have fallen into two general classes. One class focuses on a single modification type, such as Phospho.ELM (19) for phosphorylation or O-GlyBase (20) for glycosylation. Although these databases have been widely used, they are limited in utility due to recording only a single modification type. The other class of PTM database is the primary protein database; these databases collect PTM information with multiple modification types, but are more broadly focused on providing diverse information about proteins, rather than PTM information specifically. Swiss-Prot (21) and HPRD (22) are examples of such databases. As compared to either of the above two types of database, integrated PTM databases are more desirable. One example is dbPTM (23), which integrated experimentally determined PTM information from four external databases. PhosphoSite started the harvesting of phosphorylation sites from published literature with a focus on in vivo mammalian phosphorylation data (24), but recently has expanded to integrate nine other modification types (PhosphoSitePlus: www.phosphosite.org). Even integrated databases, however, have not taken into full consideration the aforementioned quickly accumulating PTM data from MS/MS experiments. These data, many of which are reported in the published literature but not collected in any database, continue to increase rapidly due to new experiments. Such a wealth of information should be incorporated more comprehensively into the current PTM knowledge domain.

At the same time, the high-throughput nature and complexity of MS/MS data pose computational challenges for proteome-scale PTM analyses in a biological context. A pure data repository is insufficient for such tasks. Powerful computational tools must accompany data repositories to allow knowledge extraction.

To address these needs, we developed a systematic resource for PTM research, SysPTM, consisting of a PTM database and four analysis tools. The SysPTM database incorporates the existing features of numerous previous databases, with an emphasis on collecting modification datasets from MS/MS experiments reported in the literature. The current release of SysPTM (v1.1) contains data detailing 117349 PTM sites on 33421 proteins involving nearly 50 modification types. The four analysis
tools are PTMBlast, PTMPathway, PTMPhylog and PTMCluster, which, respectively, can compare user PTM datasets with PTM data stored in SysPTM, map PTM proteins to KEGG pathways, discover potentially conserved PTM sites, and find significant clusters of multi-site modifications.

In this work, an in-house MS/MS phosphorylation dataset from mouse embryonic stem cells was analyzed to demonstrate the SysPTM workflow. SysPTM can be accessed via http://www.sysbio.ac.cn/SysPTM or http://lifecenter.sgst.cn/SysPTM.

EXPERIMENTAL PROCEDURES

System Configuration

SysPTM consists of a relational database and a dynamic web interface. A simplified entity-relationship diagram of the SysPTM database is shown in Supplementary Fig. 1. The SysPTM database is implemented using Mysql Server Edition 5.0 (http://www.mysql.com) and is configured on a running RedHat Linux Server. The SysPTM website is publicly available at http://www.sysbio.ac.cn/SysPTM or http://lifecenter.sgst.cn/SysPTM. The web interface is implemented with JavaServer Pages Technology (http://java.sun.com/products/jsp/) using the Apache Tomcat 5.5 server (http://tomcat.apache.org/). All functions are programmed in Java and Perl language.

Data Collection

PTM Data Collection. For comprehensive PTM data coverage and timely updates, semi-automatic methods were used to collect PTM sites from public data resources and peer-reviewed MS/MS literature (see Supplementary Methods). In the current version of SysPTM, modification information was automatically retrieved from five databases [Swiss-Prot version 56.2 (21), Phospho.ELM version 8.0 (19), HPRD release 7 (22), O-GlyBase version 6.0 (20), and Ubiprot version 1.0 (25)] and four web servers [SUMOsp version 1.0 (26), Memo version 2.0 (27), NetAcet version 1.0 (28), and LysAcet version 1.1 (paper in print)]. These data were integrated and stored as SysPTM-A (see Supplementary Methods). SysPTM-A will be updated every time a new major database version is released. In addition, numerous modification sites scattered in the MS/MS literature and rarely collected by existing databases were integrated into our database as SysPTM-B. A Perl program was used to search PubMed with the following limits: MS-related keywords (mass spectrometry,
proteomics), seven modification types (phosphorylation, acetylation, methylation, sumoylation, ubiquitination, glycosylation, S-nitrosylation), and a time duration of January 2005 to October 2008. This search retrieved 1118 MS/MS papers. Data quality was controlled by manually checking whether the original datasets were validated through manual check, score cutoff, false discovery rate, or other methods (see Supplementary Methods), which resulted in stringent literature filtering. Of the 1118 papers retrieved, 104 were selected for further data collection. (Data quality control for these 104 papers is shown in Supplementary Table 1.) Data on unambiguous PTM sites and peptides were manually extracted from the supplemental materials referenced in these candidate papers. PTM sites on peptides were automatically mapped to protein sequences. Information on modification type, site location, peptide sequence, and peptide score were then stored in SysPTM-B (Supplementary Fig. 1). SysPTM-B will be updated every six months.

**PTM Data Integration.** As mentioned above, SysPTM data were collected from diverse resources and literature, resulting in various protein identifiers from different databases. To integrate heterogeneous data and avoid redundancy, a two-step integrating process was performed (see Supplementary Methods): 1) protein identifiers in the same database were mapped to identifiers in the newest database version and corresponding protein sequences were retrieved; and 2) proteins with different identifiers but with the same sequence (100% identity) and in the same species were regarded as the same protein.

**PTM Protein Annotation.** In addition to PTM-related information, detailed functional annotations from external databases were integrated into SysPTM to help users access other protein-related knowledge, including domains identified by HMMER (29) based on the models in the Pfam database (30), pathways from KEGG (31), gene ontology from the GO database (32), relationships between genes and disease from OMIM (33), and ortholog groups from HomoloGene (34).

**Analysis Tools**

1) PTMBlast is a PTM site similarity search program designed to explore similar PTM sites between a user dataset (query) and SysPTM data (subject). The program provides three sequence alignment methods with different sensitivity and specificity: in method 1, protein sequences between query and subject must be identical; in method 2, protein sequences between query and subject are aligned by the
protein-protein BLAST (BLASTP) program (http://www.ncbi.nlm.nih.gov/blast/Blast.cgi) (35); and in method 3, modified peptides in the query are aligned with subject protein sequences by the Smith-Waterman algorithm, which is more sensitive than BLAST for short peptides (http://www.ebi.ac.uk/Tools/fasta33/index.html) (36). After sequence alignment, residues modified at the same position in both query and subject are identified as overlapped PTM sites. Otherwise, modified residues are considered unique to the query or subject. Overlapped and unique PTM sites are highlighted in vivid color on the aligned sequences, which can be downloaded.

2) PTMPathway is a tool to associate PTM proteins with KEGG pathways. The Perl SOAP::Lite package (http://search.cpan.org/~mkutter/SOAP-Lite-0.71/) and KEGG API (http://www.genome.jp/kegg/soap/) are used to access the KEGG pathway database (31).

3) PTMPhylog maps and colors PTM sites on multiple sequence alignments of orthologous proteins. Ortholog groups are retrieved from the NCBI HomoloGene database (http://www.ncbi.nlm.nih.gov/sites/entrez?db=homologene, Build 58) (34), and ClustalW (http://www.ebi.ac.uk/clustalw/index.html) (37) is used for multiple sequence alignment.

4) PTMCluster is used to find statistically significant PTM site clusters on the same protein. The PTMCluster algorithm is an improvement on a neighborhood model proposed by Li et al. (38). PTM sites are searched for candidate PTM clusters along a positional distance tree, and a p-value P is calculated for each cluster to qualify whether PTM sites in the cluster are close enough in space than randomly distributed sites (see Supplementary Fig. 2 for a detailed illustration of this method). Significant PTM clusters are selected if they satisfy p-value and site number cut-offs (P ≤ 0.01; N ≥ 3).

Case Study
An in-house phosphorylation dataset generated by an MS/MS experiment on mouse embryonic stem (mES) cell proteins was employed as a case to demonstrate the use of SysPTM (see Supplementary Methods). mES cells were lysed in lysis buffer mixed with phosphatase inhibitors. The peptide mixtures were separated by strong anion exchange - reversed-phase high performance liquid chromatography (SAX-RP HPLC). LC-MS/MS was performed using an LTQ-Orbitrap mass spectrometer. SEQUEST
was used to match MS/MS spectra against the International Protein Index (IPI) mouse database (version 3.22). All output results were merged by in-house software (BuildSummary) and the false discovery rate (FDR) was controlled to below 1%. Finally 1152 phosphosites on 526 distinct phosphoproteins were determined (Supplementary Table 5). The spectra of phosphopeptides used for the case study were manually checked (Supplementary Methods).

RESULTS

Contents and Statistics
SysPTM is a systematic resource integrating a PTM database and four analysis tools. The structure of SysPTM is depicted in Fig. 1A. The SysPTM database currently houses information relating to 117349 experimentally determined PTM sites, with nearly 50 modification types, on 33421 proteins. Most modification types are amino-acid-specific; different modification types may target the same amino acid (Supplementary Fig. 3). The number of PTM sites collected in SysPTM for several common modification types, along with the number of correlated proteins, is shown in Fig. 1B (Supplementary Table 4). Phosphorylation is the most frequent PTM, with 87068 sites on 24705 proteins; other abundant modifications include glycosylation (7564 sites) and acetylation (3001 sites).

Modification sites in SysPTM are categorized into two groups based on data source: SysPTM-A, with information culled from public resources, contains data on 75047 unique PTM sites on 21971 proteins (Supplementary Table 2); and SysPTM-B, with information culled from 104 peer-reviewed MS/MS papers, contains information on 67596 unique PTM sites on 20675 proteins (Supplementary Table 3). SysPTM-A makes up the major portion of the SysPTM database and covers nearly 50 modification types. These data include “golden standard” sites culled from public databases (Swiss-Prot, Phospho.ELM, HPRD, O-GlyBase, Ubiprot), as well as sites modeled as positive training sets for prediction webservers (SUMOsp, Memo, NetAcet, and LysAcet). Thus, the data were fully integrated into SysPTM-A after removing ambiguous items (e.g. sites listed as “potential”, “probable”, or “by similarity” in Swiss-Prot), or citing “HTP” for curated data from high-throughput technologies (e.g. “HTP” data from Phospho.ELM). SysPTM-B makes up a significant portion of SysPTM and covers seven modification types: phosphorylation, acetylation, methylation, sumoylation, ubiquitination, glycosylation, and
S-nitrosylation, with phosphorylation (64025 sites) and glycosylation (1970 sites) the most abundant modification types. SysPTM-B was collected from peer-reviewed mass spectrometry literature providing intact PTM datasets and detailed PTM identification procedures. Diverse data-quality control for modified peptides were carried out in each paper, such as: spectra manually checked, filtered by software score threshold or false discovery rate (FDR), validated by an algorithm (e.g. Ascore), single protein extensively identified with modifications, etc. For inclusion in SysPTM-B, we collected only PTM sites with at least one data-quality check measure. Ambiguous PTM sites that could not be assigned to a specific amino acid were not added to SysPTM-B. Of the PTM sites collected in SysPTM-B, 62.6% are not covered by SysPTM-A. For example, SysPTM-B contains 115 S-nitrosylation sites, whereas only 48 S-nitrosylation sites are collected in SysPTM-A.

In addition to functioning alone, modifications can “crosstalk,” working together to regulate biological functions (39). To investigate this phenomenon, the number of modification sites and types on all proteins was analyzed, with results shown in Fig. 1C and D. Fifty-nine percent of proteins collected in SysPTM have more than one modification sites (Fig. 1C), and 13% of proteins have more than one modification types (Fig. 1D).

Moreover, abundant modification data and diverse functional annotations in SysPTM provide multi-faceted views of PTM-associated features. Supplementary Table 6 lists functions associated with six frequent PTMs, shown to be significantly enriched (P<=0.01) under the chi-square test. Interestingly, proteins modified by phosphorylation and sumoylation are enriched in similar GO-defined functions, such as nucleus (GO category: cellular component), transcription (GO category: biological process), and protein binding (GO category: molecular function), which might support the hypothesis of crosstalk between these two modification types (40).

Data Accessibility
An online version of SysPTM is available at http://www.sysbio.ac.cn/SysPTM or http://lifecenter.sgst.cn/SysPTM, where users can query protein and PTM information through search and browse pages, retrieve batch files through a statistics page, upload their own modification datasets through a submit page, and acquire help through a FAQ page. Supplementary Fig. 4A shows the search page interface. Users may query the database by gene name, protein description, Swiss-Prot ID/AC, IPI ID, or NCBI
protein GI. The BLAST search engine also is included to allow users to perform sequence similarity searches. Supplementary Fig. 4B illustrates the browse page. Through this page, users can browse PTM proteins by PTM type, data source, or KEGG pathway. Search and browse results are returned as SysPTM entries; each entry contains eight sections (Supplementary Fig. 4C), each of which can be expanded by clicking to show details (Supplementary Fig. 5). The sections provide the following information: **Summary** gives basic protein information, such as species, protein description, gene name, protein identifier, and 3D structure; **PTMsite-Statistics** lists the number of PTM sites of each modification type, along with the SysPTM data source (SysPTM-A and/or B) for these sites; **PTMsite-Map** is an interactive interface to view PTM sites, Pfam domains, and PTM position clusters along the protein sequence; **PTMsite-Table** lists each PTM residue, position, type, and data source (SysPTM-A and/or B); **PTMsite-Source** provides more detailed information on data source (protein database or literature reference) and original experimental evidence; **PTMsite-Cluster** gives predicted PTM position clusters with their statistical p-value; **PTMprotein-Sequence** highlights PTM sites on the protein sequence; and **PTMprotein-Annotation** shows protein annotation information and links to public databases.

**Accessibility and Application of Four Modification Data Mining Tools**

**PTMBlast.** PTMBlast accepts PTM target sites/peptides and protein sequences as input, and provides three methods for comparing these sequences with different target datasets in SysPTM. Flexible parameters such as comparison method, similarity cutoff, species, modification type, and data source can be selected by the user, and the results of PTMBlast with aligned modification sites and statistics tables can be directly viewed on a webpage or returned by email (Fig. 2A). The three PTMBlast comparison methods vary in sensitivity and specificity. Method 1 has greatest stringency and lowest sensitivity, requiring the compared protein sequences to be exactly identical. Method 2 has medium stringency and sensitivity, and is suitable for general requirements, while method 3 has least stringency and greatest sensitivity because it considers only the similarity of modified peptides without regard to protein sequence. It is recommended for users to retrieve results by e-mail because method 2 and 3 may take long computational time. Species is another important parameter of PTMBlast. When data are compared within the same species, PTMBlast returns
modification sites that overlap with previously discovered sites, as well as potentially novel modification sites. When data are compared across species, the results of PTMBlast indicate conserved modification sites among species.

**PTMPathway.** PTMPathway implements two main functions: 1) to browse PTM proteins based on a given KEGG pathway name and PTM type and 2) to map user-defined proteins to KEGG pathways and highlight all PTM proteins. Results are returned in the form of a pathway graph: rectangles with yellow background area denote PTM proteins found in SysPTM; pink texts in rectangles denote user-defined proteins (Fig. 2C). Clicking a protein within a yellow rectangle returns the protein’s SysPTM entry page, with a detailed view of all information. Human Wnt signaling pathway was queried as an example and 406 modification sites on 64 proteins involving about 10 modification types were found by PTMPathway. Out of which five frequent modification types were manually labeled with different shapes and are shown in Fig. 3A.

**PTMPhylog.** PTMPhylog was developed to study the conservation of PTM sites and associated proteins. It searches by protein description, identifier, or sequence, and returns ortholog protein sequences aligned across multiple species, with PTM residues highlighted in color (Fig. 2B). The evolutionary conservation of human phosphoproteins was estimated based on the percent identities with ortholog proteins from 19 species in the HomoloGene database (41), and the results are shown in Fig. 3B. Generally phosphoproteins are significantly more conserved than their non-phosphorylated counterparts, suggesting functional importance of phosphoproteins. Given that this tendency may be true for other modifications as well, PTMPhylog provides an important tool for identifying conserved modification sites among different species.

**PTMCluster.** Genes with positional proximity along the genome have been reported to co-express (42) or exhibit tissue-specific features (43); similar to this, multiple modification sites on a protein may cluster within a small region, with physical proximity mediating protein biological activity (44). PTMCluster was developed to mine proteome-scale PTM position clusters. Statistical significance of clustering predictions is given by p-value, and regions of predicted clusters are plotted as yellow rectangles along the protein sequence (Fig. 2D). Almost 10000 clusters were found in protein sequences stored in SysPTM; these clusters covered most PTM types and appeared in many proteins. Fig. 3C shows an example of a PTM cluster in the
C-terminal domain (CTD) of human RNA polymerase II largest subunit (RPB1). This cluster contains 18 phosphosites, most located on the S2 or S5 residue of the repeated consensus heptapeptide YSPTSP. A previous study reported that both S2 and S5 in the YSPTSP consensus sequence were required to be simultaneously phosphorylated during M phase of the cell cycle (45). Therefore phosphosites in this PTM position cluster may work together as a functional region.

**Workflow and Case Study**

Here we propose a workflow for analyzing an experimental PTM dataset with SysPTM, using the example of an MS-identified phosphorylation dataset from mouse embryonic stem (mES) cells. The experimental aim was to identify phosphoproteins in stem cells and explore their functions. At the data overview level of the workflow (see Fig. 4A), PTMBlast was performed to compare phosphosites in mES cells with phosphosites recorded in SysPTM. The number of unique phosphosites identified in our experiment was quite large (59.2% using method 1, 48.9% using method 2, and 47.6% using method 3). These unique phosphosites might reveal potential stem-cell-specific characteristics of mES cells. GO functional analysis of newly identified and overlapped phosphoproteins based on PTMBlast method 2 was performed, and the results are shown in Fig. 4B. A significant number of the unique phosphoproteins are involved in DNA binding, transcription factor activity, regulation of transcription, and multi-cellular organism development, supporting the idea that phosphorylation is an important regulatory process for ES cells (46).

We next generated an overview of phosphorylation at the network level (see Fig. 4C) by mapping phosphoproteins recorded in SysPTM and/or identified in our experiment to a published ES-related protein interaction network (47). In total, 16 (47.1%) of the proteins in the network have experimentally identified phosphosites; 10 of these 16 proteins were identified in our experiment. As seen in Fig. 4C, these phosphoproteins interact with mES cell markers (Nanog, Oct4, and Rex1), indicating potential important roles of phosphorylation in the stem-cell-specific characteristics of mES cells.

Next, the association between biological pathways and phosphorylation was explored using the PTMPathway tool. Focal adhesion kinase (FAK) signaling has been reported as an important pathway in mES cells, regulating cardiogenesis (48). Using PTMPathway, we developed a new picture of the FAK pathway, highlighting 25
potential phosphoproteins (Fig. 4D). These phosphorylated components may help researchers explore further the association between the FAK pathway and stem cells. In addition to global analysis, SysPTM also is excellent for the analysis of an individual modified proteins. PTMCluster and PTMPhylog were used, respectively, to analyze PTM site clusters and evolutionary conservation of phosphosites on the potential stem cell marker α-catenin (CTNA1_MOUSE). PTMCluster found one PTM position cluster in α-catenin, shown in Fig. 4E. For phosphosite S641, PTMPhylog found the aligned sites were conserved across multiple orthologs and it could be phosphorylated in human (Fig. 4F).

In short, the multi-level analysis run by SysPTM on an in-house proteomic phosphorylation dataset from mouse embryonic stem cells demonstrates that the roles of post-translational modification in a biological organism can be analyzed systematically by combining high-throughput experiments and powerful computational analysis.

DISCUSSION AND CONCLUSION

The progress of post-translational modification research has accelerated since the introduction of mass spectrometry as a primary technology in the field of proteomics. In 2007, modificomics was advocated as a prominent and independent extension of proteomics (49), with the aim of exploring the language of post-translational modifications at the “omics” level and interpreting phenotype in the context of PTMs (39). To aid modificomics research, we have developed SysPTM, which integrates a data repository with powerful information extraction tools.

The SysPTM database (v1.1) is made up of SysPTM-A and SysPTM-B, together containing data on 117349 experimentally determined PTM sites on 33421 proteins covering nearly 50 PTM types. SysPTM-A contains 75047 unique PTM sites on 21971 proteins integrated from five databases (Swiss-Prot, Phospho.ELM, HPRD, O-GlyBase, and Ubiprot) and four web servers (SUMOsp, Memo, NetAcet, and LysAcet); this portion of the database will be updated every time a new major database version is released. SysPTM-B contains 67596 unique PTM sites on 20675 proteins manually collected from 104 papers selected from 1118 papers published from January 2005 to October 2008; this portion of the database will be updated every six months with data from new publications. SysPTM-B is unique in its emphasis on collecting modification datasets from MS/MS experiments, an emphasis lacking in
previous PTM databases (19-22, 25). SysPTM is the most comprehensive PTM data repository available today, with the capacity to store and analyze PTM site/protein data for multiple modification types, from multiple data sources and multiple organisms.

Data quality is critical for reliable utility of SysPTM. Currently we rely on a data quality check performed primarily on original data sources, as described in Methods, Results and Supplemental Methods. In addition, we collect as much original identification information for each dataset as possible to record in SysPTM-B, including modification sites, peptide sequences, peptide scores, search engines, and mass spectrometry equipments. Experienced users can perform specific quality filtering by setting their own score thresholds. For each PTM site in SysPTM-B, the confidence also can be checked to some degree based on the number of papers listed that have identified this site. In the future, we plan to develop a statistical score function to integrate these naïve measures (e.g., modified peptide score, FDR cutoff, and count of publications) and assign a quality score for each PTM site to allow users a more ready quality check. On the other hand, when more and more proteomics datasets comprising raw mass spectra are stored in public databases like PRIDE (50), analogous to the storage of microarray data in GEO (51), a uniform algorithm for peptide identification (e.g., PeptideProphet) and PTM site localization (e.g., Ascore) may be applied or developed; a consistent quality check system for all PTM sites will then be available.

Rich data in SysPTM make it a good resource for examining general PTM features. Abundant PTM sites provide valuable information for detecting PTM sequence properties such as flanking sequence motifs and developing PTM prediction algorithms. Schwartz et al. have proposed a new general strategy to predict organism-specific PTMs (52); however, due to limited data, their strategy has been applied only to phosphorylation in yeast, fly, mouse, and human, and acetylation in human. Using the abundant data in SysPTM, their method could now be generalized to other modification types or species.

The multi-organism feature of SysPTM makes it extremely useful for comparative modificomics. Boekhorst et al. compared phosphoproteomics datasets of six eukaryotes and revealed evolutionary conservation of phosphorylation (53). In SysPTM, comparison across multiple species also showed that phosphoproteins are significantly more conserved than their non-phosphorylated counterparts, suggesting
functional importance of phosphorylation. Conservation of other modifications also can be analyzed using SysPTM.

Additionally, with SysPTM, multi-site and multi-type modification can be investigated. Our results show that 59% of proteins in SysPTM have multiple modification sites and 13% of proteins have multiple modification types. For instance, human epidermal growth factor receptor (EGFR) has 83 modification sites and 4 modification types: phosphorylation, ubiquitination, glycosylation, and disulfide bridges (see http://www.biosino.org/SysPTM/protein.jsp?proteinID=M16847). Such multi-site and multi-type modification may indicate complicated regulations that enable EGFR to exert its broad range of functions in biological processes (54). In this regard, SysPTM is also a valuable resource for studying the potential co-regulation of multiple modification sites and crosstalk between different PTM types.

A vast data source without proper data mining tools may be data rich, but information poor. Experimental PTM researchers need tools to access background knowledge, compare their data with existing discoveries, and anticipate future research directions. Taking into consideration of these needs, four in-depth analysis tools were developed in affiliation with the SysPTM database. PTMBlast is a similarity comparison tool supporting multi-type and multi-organism modification datasets. PTMBlast takes advantage of rich modification information in SysPTM and allows a researcher to conveniently compare his/her own dataset, to distinguish between confirmations of previously discovered PTM sites and novel site identifications. The second tool, PTMPathway, takes the data analysis one step further. Many proteins in biological pathways may be modified; modifications may trigger, mediate, or terminate pathway activity. To elucidate the complexity and diversity of PTMs in the context of pathways, however, is a formidable task. PTMPathway provides a quick view of all modified proteins in a certain pathway; in combination with PTMBlast, PTMPathway further allows for visualization of newly identified modification proteins on pathway. Therefore, PTMPathway provides a very useful tool to extend our understanding of modifications to biological pathways at a systematic level.

As stated earlier, highly conserved protein residues tend to correlate with structural or functional importance, and multi-site modification has been reported to be a common regulation mechanism. To enable analysis of residue conservation tendency for all PTM types, the PTMPhylog tool was developed. PTMCluster is the first algorithm to search proteins for multi-site PTM position clusters. Almost 10000 clusters were
found in protein sequences stored in SysPTM, suggesting a new method for studying co-regulation of multi-site modifications.

The overall process of using SysPTM was demonstrated through a case study. By analyzing an in-house phosphorylation dataset identified by MS/MS, we showed that, in SysPTM, a modificomics dataset can be mapped to other sequences, to biological pathways, and to existing networks for data overview; in addition, the conservation status and clustering status of individual residues of important modified proteins can be studied. In general, the roles of single-type and multi-type modifications can be investigated in a full biological context. This is the contribution of SysPTM to modificomics research. Thus, SysPTM may become an important aid to both experimental and computational post-translational modification researchers to enhance proteomics progress in this important and challenging field.

Future Development. Plans are underway to: 1) develop a statistical tool for scoring the confidence of MS/MS-identified modifications; 2) allow users to submit datasets via a temporary database and store them in SysPTM after manual check; and 3) design an algorithm to predict modification sites.

Acknowledgments
We acknowledge Lynne Berry from the Vanderbilt Cancer Biostatistics Center for her great editing work. Funding for this research was provided by National High-Tech R&D Program (863) 2006AA02Z334, National Basic Research Program of China (2006CB910700, 2009CB918404, to Yixue Li); National Natural Science Foundation of China 30621091, Ministry of Science and Technology (2006CB943900, 2007CB947904, 2007CB947100, 2007CB948000), Key Research Program (CAS) KSCX2-YW-R-112 (to Rong Zeng); Key project for prevention and treatment of major infectious diseases (2008ZX10002-021) and Shanghai Natural Science Fund 08ZR1415800 (to Lu Xie).

References
1. de Hoog, C. L., and Mann, M. (2004) Proteomics. Annu Rev Genomics Hum Genet 5, 267-293.
2. Mann, M., and Jensen, O. N. (2003) Proteomic analysis of post-translational modifications. Nat Biotechnol 21, 255-261.
3. Virshup, D. M., Eide, E. J., Forger, D. B., Gallego, M., and Harnish, E. V. (2007) Reversible
Protein Phosphorylation Regulates Circadian Rhythms. Cold Spring Harb Symp Quant Biol 72, 413-420.
4. Yang, X. J. (2004) Lysine acetylation and the bromodomain: a new partnership for signaling. Bioessays 26, 1076-1087.
5. Gill, G. (2004) SUMO and ubiquitin in the nucleus: different functions, similar mechanisms? Genes Dev 18, 2046-2059.
6. Krueger, K. E., and Srivastava, S. (2006) Posttranslational protein modifications: current implications for cancer detection, prevention, and therapeutics. Mol Cell Proteomics 5, 1799-1810.
7. Yao, D., Gu, Z., Nakamura, T., Shi, Z. Q., Ma, Y., Gaston, B., Palmer, L. A., Rockenstein, E. M., Zhang, Z., Masliah, E., Uehara, T., and Lipton, S. A. (2004) Nitrosative stress linked to sporadic Parkinson's disease: S-nitrosylation of parkin regulates its E3 ubiquitin ligase activity. Proc Natl Acad Sci U S A 101, 10810-10814.
8. Martin-Rendon, E., and Blake, D. J. (2003) Protein glycosylation in disease: new insights into the congenital muscular dystrophies. Trends Pharmacol Sci 24, 178-183.
9. Johnson, B. E., and Janne, P. A. (2005) Epidermal growth factor receptor mutations in patients with non-small cell lung cancer. Cancer Res 65, 7525-7529.
10. Sadri-Vakili, G., and Cha, J. H. (2006) Mechanisms of disease: Histone modifications in Huntington's disease. Nat Clin Pract Neurol 2, 330-338.
11. Zhi, Y., and Sandri-Goldin, R. M. (1999) Analysis of the phosphorylation sites of herpes simple virus type 1 regulatory protein ICP27. J Virol 73, 3246-3257.
12. Brunet, A., Sweeney, L. B., Sturgill, J. F., Chua, K. F., Greer, P. L., Lin, Y., Tran, H., Ross, S. E., Mostoslavsky, R., Cohen, H. Y., Hu, L. S., Cheng, H. L., Jedrychowski, M. P., Gygi, S. P., Sinclair, D. A., Alt, F. W., and Greenberg, M. E. (2004) Stress-dependent regulation of FOXO transcription factors by the SIRT1 deacetylase. Science 303, 2011-2015.
13. Yuan, Z. L., Guan, Y. J., Chatterjee, D., and Chin, Y. E. (2005) Stat3 dimerization regulated by reversible acetylation of a single lysine residue. Science 307, 269-273.
14. Olsen, J. V., Blagoev, B., Gnad, F., Macek, B., Kumar, C., Mortensen, P., and Mann, M. (2006) Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. Cell 127, 635-648.
15. Elortza, F., Nuhe, T. S., Foster, L. J., Stensballe, A., Peck, S. C., and Jensen, O. N. (2003) Proteomic analysis of glycosylphosphatidylinositol-anchored membrane proteins. Mol Cell Proteomics 2, 1261-1270.
16. Kim, S. C., Sprung, R., Chen, Y., Xu, Y., Ball, H., Pei, J., Cheng, T., Kho, Y., Xiao, H., Xiao, L., Grishin, N. V., White, M., Yang, X. J., and Zhao, Y. (2006) Substrate and functional diversity of lysine acetylation revealed by a proteomics survey. Mol Cell 23, 607-618.
17. Denison, C., Rudner, A. D., Gerber, S. A., Bakalarski, C. E., Moazed, D., and Gygi, S. P. (2005) A proteomic strategy for gaining insights into protein sumoylation in yeast. Mol Cell Proteomics 4, 246-254.
18. Martinez-Ruiz, A., and Lamas, S. (2007) Proteomic identification of S-nitrosylated proteins in endothelial cells. Methods Mol Biol 357, 215-223.
19. Diella, F., Cameron, S., Gemund, C., Linding, R., Via, A., Kuster, B., Sicheritz-Ponten, T., Blom, N., and Gibson, T. J. (2004) Phospho.ELM: a database of experimentally verified phosphorylation sites in eukaryotic proteins. BMC Bioinformatics 5, 79.
20. Gupta, R., Birch, H., Rapacki, K., Brunak, S., and Hansen, J. E. (1999) O-GLYCBASE version 4.0: a revised database of O-glycosylated proteins. Nucleic Acids Res 27, 370-372.
21. Watanabe, K., and Harayama, S. (2001) [SWISS-PROT: the curated protein sequence database on Internet]. Tanpakushitsu Kakusan Koso 46, 80-86.

22. Mishra, G. R., Suresh, M., Kumaran, K., Kannabiran, N., Suresh, S., Bala, P., Shivakumar, K., Anuradha, N., Reddy, R., Raghavan, T. M., Menon, S., Hanumanthu, G., Gupta, M., Upendra, S., Gupta, S., Mahesh, M., Jacob, B., Mathew, P., Chatterjee, P., Arun, K. S., Sharma, S., Chandrika, K. N., Deshpande, N., Palvankar, K., Raghavnath, R., Krishnakanth, R., Karathia, H., Rekha, B., Nayak, R., Vishnupriya, G., Kumar, H. G., Nagini, M., Kumar, G. S., Jose, R., Deepthi, P., Mohan, S. S., Gandhi, T. K., Harsha, H. C., Deshpande, K. S., Sarker, M., Prasad, T. S., and Pandey, A. (2006) Human protein reference database--2006 update. Nucleic Acids Res 34, D411-414.

23. Lee, T. Y., Huang, H. D., Hung, J. H., Huang, H. Y., Yang, Y. S., and Wang, T. H. (2006) dbPTM: an information repository of protein post-translational modification. Nucleic Acids Res 34, D622-627.

24. Hornbeck, P. V., Chabra, I., Kornhauser, J. M., Skrzypek, E., and Zhang, B. (2004) PhosphoSite: A bioinformatics resource dedicated to physiological protein phosphorylation. Proteomics 4, 1551-1561.

25. Chernorudskiy, A. L., Garcia, A., Eremin, E. V., Shorina, A. S., Kondratieva, E. V., and Gainullin, M. R. (2007) Ubiprot: a database of ubiquitylated proteins. BMC Bioinformatics 8, 126.

26. Xue, Y., Zhou, F., Fu, C., Xu, Y., and Yao, X. (2006) SUMOsp: a web server for sumoylation site prediction. Nucleic Acids Res 34, W254-257.

27. Chen, H., Xue, Y., Huang, N., Yao, X., and Sun, Z. (2006) MeMo: a web tool for prediction of protein methylation modifications. Nucleic Acids Res 34, W249-253.

28. Kiemer, L., Bendtsen, J. D., and Blom, N. (2005) NetAcet: prediction of N-terminal acetylation sites. Bioinformatics 21, 1269-1270.

29. Eddy, S. R. (1998) Profile hidden Markov models. Bioinformatics 14, 755-763.

30. Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L., Studholme, D. J., Yeats, C., and Eddy, S. R. (2004) The Pfam protein families database. Nucleic Acids Res 32, D138-141.

31. Kanehisa, M., and Goto, S. (2000) KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28, D514-517.

32. Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25, 25-29.

33. Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A., and McKusick, V. A. (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res 33, D514-517.

34. Wheeler, D. L., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., DiCuccio, M., Edgar, R., Federhen, S., Gerl, L. Y., Kapustin, Y., Khovayko, O., Landsman, D., Lipman, D. J., Madden, T. L., Maglott, D. R., Ostell, J., Miller, V., Pruitt, K. D., Schuler, G. D., Sequeira, E., Sherry, S. T., Sirotkin, K., Souvorov, A., Starchenko, G., Tatusov, R. L., Tatusova, T. A., Wagner, L., and Yashchenko, E. (2007) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 35, D514-517.

35. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25, 3389-3402.
20. Smith, T. F., and Waterman, M. S. (1981) Identification of common molecular subsequences. J Mol Biol 147, 195-197.

37. Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22, 4673-4680.

38. Li, Q., Lee, B. T., and Zhang, L. (2005) Genome-scale analysis of positional clustering of mouse testis-specific genes. BMC Genomics 6, 7.

39. Jensen, O. N. (2006) Interpreting the protein language using proteomics. Nat Rev Mol Cell Biol 7, 391-403.

40. Yang, X. J., and Gregoire, S. (2006) A recurrent phospho-sumoyl switch in transcriptional repression and beyond. Mol Cell 23, 779-786.

41. Macek, B., Gnad, F., Soufi, B., Kumar, C., Olsen, J. V., Mijakovic, I., and Mann, M. (2008) Phosphoproteome Analysis of E. coli Reveals Evolutionary Conservation of Bacterial Ser/Thr/Tyr Phosphorylation. Mol Cell Proteomics 7, 299-307.

42. Boutanaev, A. M., Kalmykova, A. I., Shevelyov, Y. Y., and Nurminsky, D. I. (2002) Large clusters of co-expressed genes in the Drosophila genome. Nature 420, 666-669.

43. Caron, H., van Schaik, B., van der Mee, M., Baas, F., Riggins, G., van Sluis, P., Hermus, M. C., van Asperen, R., Boon, K., Voute, P. A., Heisterkamp, S., van Kampen, A., and Versteeg, R. (2001) The human transcriptome map: clustering of highly expressed genes in chromosomal domains. Science 291, 1289-1292.

44. Yang, X. J. (2005) Multisite protein modification and intramolecular signaling. Oncogene 24, 1653-1662.

45. Xu, Y. X., Hirose, Y., Zhou, X. Z., Lu, K. P., and Manley, J. L. (2003) Pin1 modulates the structure and function of human RNA polymerase II. Genes Dev 17, 2765-2776.

46. Prudhomme, W., Daley, G. Q., Zandstra, P., and Lauffenburger, D. A. (2004) Multivariate proteomic analysis of murine embryonic stem cell self-renewal versus differentiation signaling. Proc Natl Acad Sci U S A 101, 2900-2905.

47. Wang, J., Rao, S., Chu, J., Shen, X., Levasseur, D. N., Theunissen, T. W., and Orkin, S. H. (2006) A protein interaction network for pluripotency of embryonic stem cells. Nature 444, 364-368.

48. Hakuno, D., Takahashi, T., Lammerding, J., and Lee, R. T. (2005) Focal adhesion kinase signaling regulates cardiogenesis of embryonic stem cells. J Biol Chem 280, 39534-39544.

49. Reinders, J., and Sickmann, A. (2007) Modificomics: posttranslational modifications beyond protein phosphorylation and glycosylation. Biomol Eng 24, 169-177.

50. Jones, P., Cote, R. G., Cho, S. Y., Klé, S., Martens, L., Quinn, A. F., Thornycroft, D., and Hermjakob, H. (2008) PRIDE: new developments and new datasets. Nucleic Acids Res 36, D878-883.

51. Barrett, T., Troup, D. B., Wilhite, S. E., Ledoux, P., Rudnev, D., Evangelista, C., Kim, I. F., Soboleva, A., Tomashovskiy, M., and Edgar, R. (2007) NCBI GEO: mining tens of millions of expression profiles--database and tools update. Nucleic Acids Res 35, D760-765.

52. Schwartz, D., Chou, M. F., and Church, G. M. (2008) Predicting protein post-translational modifications using meta-analysis of proteome-scale data sets. Mol Cell Proteomics.

53. Boekhorst, J., van Breukelen, B., Heck, A. J., and Snel, B. (2008) Comparative phosphoproteomics reveals evolutionary and functional conservation of phosphorylation across eukaryotes. Genome Biol 9, R144.

54. Seet, B. T., Dikic, I., Zhou, M. M., and Pawson, T. (2006) Reading protein modifications with
interaction domains. *Nat Rev Mol Cell Biol* 7, 473-483.
FIGURE LEGENDS:

**Figure 1:** SysPTM content and statistics. A. SysPTM road map, with overview of database construction and online tools. B. Number of unique proteins and unique PTM sites stored in SysPTM, in total and for seven common modification types. C. Distribution of PTM site number: 59% of proteins have more than one modification site. D. Distribution of PTM type number: 13% of proteins have more than one modification type.

**Figure 2.** Sample pages showing the interface for SysPTM tools. A. PTMBlast results. Unique (aqua for SysPTM data and orange for query data) and overlapped PTM sites (magenta) are highlighted in color on the protein sequence for display and download. B. PTMPhylog results. PTM sites are mapped to aligned sequences of ortholog proteins in different species. Red residues are from SysPTM-A; blue residues are uniquely from SysPTM-B. C. PTMPathway results. Rectangles with yellow background area denote PTM proteins found in SysPTM; pink texts in rectangles denote user-defined proteins. D. PTMCluster results. The protein is plotted as a horizontal box. In this box, predicted PTM clusters are represented as yellow bars and Pfam domains are shown as gray bars. PTM sites from SysPTM-A and SysPTM-B, respectively, are labeled above and below the protein box.

**Figure 3:** Application of SysPTM tools. A. PTM Pathway diagram showing five types of frequent modifications on proteins in the human Wnt signaling pathway. B. Evolutionary conservation of human phosphoproteins and non-phosphoproteins, based on percent identities with orthologs from 19 other species. C. PTM position cluster on human RPB1 predicted by PTMCluster. This cluster ranges from amino acid 1836 to 1924, with red characters representing phosphosites recorded in SysPTM-A and blue characters representing phosphosites uniquely recorded in SysPTM-B. Consensus heptapeptide YSPTSP repeats are underlined. Note that many phosphosites occur at S2 or S5 of these repeats.

**Figure 4:** Example workflow outlining analysis of an mES cell protein phosphorylation dataset. Identified phosphopeptides and phosphoproteins were taken as input for multi-level analysis in SysPTM. A. PTMBlast results comparing phosphosites in mES cell proteins and phosphosites in SysPTM by three methods. B. GO functional analysis of newly identified and overlapped phosphoproteins based on PTMBlast method 2. C, F, and P are abbreviations for GO categories: C = cell component, F = molecular function, P = biological process. C. Phosphoproteins in an
ES-related protein interaction network. D. PTMPathway results showing all phosphoproteins in mouse focal adhesion pathway. E. PTMCluster results for the protein CTNA1_MOUSE. The yellow bar is a PTM position cluster predicted by PTMCluster. F. PTMPhylog results for the protein CTNA1_MOUSE. Residues shown in red are recorded in SysPTM-A; residues in blue are uniquely from SysPTM-B. S641 and S652 are identified phosphosites in the mES case.

SUPPLEMENTARY MATERIAL:
(Supplementary material is freely available at: http://www.biosino.org/papers/SysPTM)

Supplementary Figure 1: A simplified entity-relationship (ER) diagram for the SysPTM database.

Supplementary Figure 2: A neighborhood model to find PTM position clusters.

Supplementary Figure 3: Amino acid bias of different modification types. For a given modification type, the percentage of experimentally identified modifications that occur at each type of amino acid is shown by heatmap.

Supplementary Figure 4: Sample pages showing SysPTM database interface. A. Search interface. Users can query a protein by sequence or keyword. B. Browse interface. Users can browse proteins by modification type, data source, or KEGG pathway. C. Browse/search results for the example protein entry M17929. Each section of the protein entry can be opened by clicking to reveal detailed information (Supplementary Fig. 5).

Supplementary Figure 5: Detailed view of example SysPTM entry result, showing eight expanded sections: Summary, PTMsite-Statistics, PTMsite-Map, PTMsite-Table, PTMsite-Source, PTMsite-Cluster, PTMprotein-Sequence, and PTMprotein-Annotation.

Supplementary Table 1: Quality control and summary of MS/MS literature providing data for SysPTM-B.

Supplementary Table 2: Versions and statistics for public resources providing data for SysPTM-A.

Supplementary Table 3: Statistics for MS/MS literature providing data for SysPTM-B.

Supplementary Table 4: Number of PTM sites and proteins for each PTM type
covered in SysPTM.

**Supplementary Table 5:** List of phosphopeptides identified by SEQUEST in mouse embryonic stem cells. The symbols * and # indicate phosphorylation and oxidation, respectively. Clicking the hyperlink on the peptide sequence will link automatically to the mass spectrum.

**Supplementary Table 6:** Significantly enriched (P<= 0.01) functional annotations of six frequent modifications (acetylation, glycosylation, methylation, phosphorylation, sumoylation and ubiquitination), in the categories of GO cellular component, GO biological process, GO molecular function, KEGG pathway, and Pfam domain. P-value was calculated by the chi-square test.
Results:

SLGPRPTGGSSPEIR  SLESVLSLGRPTGGSSPEIR
100.0 %; Homo sapiens)

Download

| Species         | NCBI_Protein_GI | Multiple_Sequence_Alignment |
|-----------------|-----------------|-----------------------------|
| Homo sapiens    | NP_003528       | MARTQKAKRLQGKAPNLAKAARKSAPATOGVKKPHRYSRPTOTALREIRYQRSTE N S |
| Pan troglodytes | XP_518289       | MARTQKAKRLQGKAPNLAKAARKSAPATOGVKKPHRYSRPTOTALREIRYQRSTE N S |
| Bos taurus      | XP_874006       | MARTQKAKRLQGKAPNLAKAARKSAPATOGVKKPHRYSRPTOTALREIRYQRSTE N S |
| Mus musculus    | NP_038578       | MARTQKAKRLQGKAPNLAKAARKSAPATOGVKKPHRYSRPTOTALREIRYQRSTE N S |
| Mus musculus    | NP_835513       | MARTQKAKRLQGKAPNLAKAARKSAPATOGVKKPHRYSRPTOTALREIRYQRSTE N S |
| Mus musculus    | NP_835514       | MARTQKAKRLQGKAPNLAKAARKSAPATOGVKKPHRYSRPTOTALREIRYQRSTE N S |
| Mus musculus    | NP_659539       | MARTQKAKRLQGKAPNLAKAARKSAPATOGVKKPHRYSRPTOTALREIRYQRSTE N S |
| Rattus norvegicus | XP_001057941   | MARTQKAKRLQGKAPNLAKAARKSAPATOGVKKPHRYSRPTOTALREIRYQRSTE N S |
| Danio rerio     | XP_692985       | MARTQKAKRLQGKAPNLAKAARKSAPATOGVKKPHRYSRPTOTALREIRYQRSTE N S |
| Arabidopsis thaliana | NP_177690 | MARTQKAKRLQGKAPNLAKAARKSAPATOGVKKPHRYSRPTOTALREIRYQRSTE N S |

**********   *   **********;**********;**********;**********;**********;**********;

**REDUCTIVE CARBOXYLATE CYCLE IN PHOTOSYNTHETIC BACTERIA**
**Figure 4**

**A**

| PTMBlast | All_Sites | Overlap_Sites | Unique_Sites |
|----------|-----------|---------------|--------------|
| Method1  | 1152      | 469           | 683          |
| Method2  | 1152      | 588           | 564          |
| Method3  | 1324      | 694           | 630          |

**B**

- transcription factor complex (C)
- chromosome (C)
- DNA binding (F)
- transcription factor activity (F)
- regulation of transcription (P)
- multicellular organismal development (P)

**C**

- Caveolin
- ITGA
- ITGB
- PKC
- ECM
- ECM-receptor interaction
- Calpain
- Vinculin
- Zyxin
- VASP
- Vinculin
- Parvin
- Actinin
- Filamin
- Talin
- Paxillin
- Actin
- RhoA
- RhoGAP
- RhoGEF
- MLCP
- MLCK
- MLC
- ROCK
- mDia
- PAK
- Cdc42
- ILK
- Src
- FAK
- Cavolin

**D**

- Cell motility

**E**

- CTNA1_MOUSE
- T658
- S655
- S654
- S652
- S641
- Y177
- Y619

**F**

- Homo sapiens
- Pan troglodytes
- Mus musculus
- Rattus norvegicus
- Gallus gallus
- Danio rerio

**Legend**

- Diamond: Unidentified protein
- Circle: Identified protein
- Blue: Identified phosphoprotein
- Green: Phosphoprotein in SysPTM
- Red: Overlapped phosphoprotein