Supplementary materials:

GPS-PUP: Computational prediction of pupylation sites in prokaryotic proteins

Zexian Liu[1,2†], Qian Ma[1†], Jun Cao[1], Xinjiao Gao[1] Jian Ren[3‡], Yu Xue[1,2‡]

[1]Hefei National Laboratory for Physical Sciences at Microscale and School of Life Sciences, University of Science & Technology of China, Hefei, Anhui 230027, China

[1]Hubei Bioinformatics and Molecular Imaging Key Laboratory, Department of Systems Biology, College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan, Hubei 430074, China

[3]State Key Laboratory of Biocontrol, School of Life Sciences, Sun Yat-sen University, Guangzhou, Guangdong 510275, China

Running title: *Prediction of pupylation sites*

[†]The two authors contributed equally to this work.

[‡]To whom correspondence should be addressed. Tel: +86-27-87793903; Fax: +86-27-87793172; Email: xueyu@mail.hust.edu.cn or xueyuhust@gmail.com.

Correspondence may also be addressed to Jian Ren. Tel/Fax: +86-20-39943788; Email: renjian.sysu@gmail.com.

Supplementary Materials & Methods

Data preparation

We manually collected experimentally identified pupylated substrates along with their location sites by searching PubMed with the keywords of "pupylation" and "prokaryotic ubiquitin", followed by checking the scientific literature published before March 22nd, 2011. A dataset with 146 experimentally verified pupylation sites from 131 proteins was obtained. The corresponding sequences were retrieved from the UniProt database (http://www.uniprot.org/).[1]

As previously described,[2-7] we defined a *pupylation site peptide* PSP(*m, n*) as a lysine (K) residue flanked by *m* amino acids upstream and *n* amino acids downstream, while the known pupylation sites were taken as positive data (+) and all other non-pupylated lysines were regarded as negative data (-). Since the redundancy of homologous sites in the positive data (+) might lead to an overestimate, we used CD-HIT to cluster the protein sequences,[8] followed by re-alignment with BLAST packages and a manual check of the proteins with ≥40% identity.[9] If two pupylation sites from two homologous proteins were at the same position according to the sequence alignment, only one site was preserved while the other site and its corresponding sequence were discarded. Ultimately, the non-redundant training dataset contained 109 substrates with 127 positive sites and 1,405 negative sites. The 127 experimentally verified pupylation sites are shown in Table S1.

The algorithms

During the past several years, we developed the GPS (Initially defined as Group-based Phosphorylation Scoring and later renamed as Group-based Prediction System) series of algorithms mainly for the prediction of post-translational modification (PTM) sites in proteins.[2-7] Although various versions of GPS algorithm employed different approaches for performance improvement (Table S4), the fundamental hypothesis of the scoring strategy was not changed.

In the scoring strategy, we hypothesized that similar short peptides exhibit similar biochemical properties and functions.[2-7] Then we used an amino acid substitution matrix, e.g., BLOSUM62, to calculate the similarity between the two PSP(*m*, *n*) peptides of *A* and *B* as below:

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

Score(*A*[*i*], *B*[*i*]) represents the substitution score of the two amino acid of *A*[*i*] and *B*[*i*] in an amino acid substitution matrix, e.g., BLOSUM62. If S(*A*, *B*)<0, we simply redefined it as S(*A*, *B*)=0.

For performance improvement, we adopted a computational pipeline of three sequential steps of motif length selection (MLS), weight training (WT) and matrix mutation (MaM).

1) Motif length selection (MLS). In this step, the combinations of PSP(*m*, *n*) (*m* = 1, …, 30; *n* = 1, …, 30) were extensively tested, while the optimized combination of PSP(*m*, *n*) with the highest leave-one-out (LOO) performance was determined. We fixed the *Sp* at 80% to compare the *Sn* values. The PSP(8, 18) was determined in this study.

2) Weight training (WT). We updated the substitution score between two PSP(*m*, *n*) peptides *A* and *B* as:

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

The w_i is the weight of position *i*. Again, if *S'*(*A*, *B*)<0, we simply redefined it as *S'*(*A*, *B*)=0. Initially, the *w* was defined 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* value of the re-computed LOO result with the *Sp* fixed at 80% was increased. The process was repeated until convergence was reached. The weights of the PSP(8, 18) were 1, 0, 0, 3, 2, 2, 3, 1, 1 (K), 1, 1, 0, 1, 1, 2, 3, 1, 0, 1, -1, 0, 1, 0, 2, 1, 1, and 3. From the results, we proposed that the upstream amino acids are more important for the lysine residue to be pupylated.

3) Matrix mutation (MaM). As previously described,[2-5] BLOSUM62 was chosen as the initial matrix, and the leave-one-out performance was calculated. Subsequently, we fixed the *Sp* as 80% to improve the *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.

For comparison, the GPS 2.1 algorithm and PSSM algorithm were also implemented. The GPS 2.1 algorithm was carried out as previously described.[5] For the PSSM algorithm,[10] the position-specific scoring matrix was constructed with positive PSP(*m*, *n*) peptides while the background distribution was calculated from both the positive and negative PSP(*m*, *n*) peptides. $P_+[i]$ and $P_-[i]$ were defined as the probability in the position-specific scoring matrix and the background, respectively. Then the score of a given PSP(*m*, *n*) was calculated as:

$$Score[PSP(m,n)] = \sum_{-m \le i \le n} \log_2(P_+[i] / P_-[i])$$

Performance evaluation

As previously described,[2-5] we used the four measurements of accuracy (*Ac*), sensitivity (*Sn*), specificity (*Sp*), and Mathew's Correlation Coefficient (*MCC*) to evaluate the prediction performance of GPS-PUP. Also, the precision (*Pr*) was calculated. The five measurements were defined as below:

$$Ac = \frac{TP+TN}{TP+FP+TN+FN}, \quad Sn = \frac{TP}{TP+FN}, \quad Sp = \frac{TN}{TN+FP}, \quad Pr = \frac{TP}{TP+FP}, \text{ and}$$

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

In this work, the leave-one-out validation and 4-, 6-, 8- and 10-fold cross-validations were performed. The Receiver Operating Characteristic (ROC) curves and AROCs (area under ROCs) were also drawn and analyzed.

Implementation of the online service and local packages

The online service and local packages of GPS-PUP 1.0 were implemented in JAVA. For the online service, we tested the GPS-PUP 1.0 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, GPS-PUP 1.0 can be directly used without any additional packages. For convenience, we also developed local packages of GPS-PUP 1.0, which worked with the three major Operating Systems, Windows, Linux and Mac.

Statistical analysis

In order to analyze the functional abundance and diversity of pupylation, we downloaded the gene ontology (GO) (03/28/2011)[11] association files from the GOA database at the EBI (http://www.ebi.ac.uk/goa). There are 4,470 *M. smegmatis* proteins annotated with at least one GO term, with 267 annotated pupylation substrates. Here we defined:

N = number of proteins in the *M. smegmatis* proteome annotated by at least one GO term

n = number of proteins in the *M. smegmatis* proteome annotated by the GO term *t*

M = number of proteins in the *M. smegmatis* pupylated substrates annotated by at least one GO term

m = number of proteins in the *M. smegmatis* pupylated substrates annotated by the GO term *t*

Then the enrichment ratio of the GO term *t* was calculated, and the hypergeometric distribution equation[12] was used to calculate the *p*-value as below:

$$Enrichment\_ratio = \frac{\frac{m}{M}}{\frac{n}{N}}$$

$$p-value = \sum_{m'=m}^{n} \frac{\binom{M}{m'}\binom{N-M}{n-m'}}{\binom{N}{n}} \quad (Enrichment\_ratio \geq 1)\text{, or}$$

$$p-value = \sum_{m'=0}^{m} \frac{\binom{M}{m'}\binom{N-M}{n-m'}}{\binom{N}{n}} \quad (Enrichment\_ratio < 1)$$

In this work, we only consider the over-represented GO groups with an *Enrichment_ratio* ≥ 1 and p-value < 0.05.

Supplementary References

1. UniProt Consortium, *Nucleic Acids Res.*, 2010, **38**, D142-148.
2. Y. Xue, J. Ren, X. Gao, C. Jin, L. Wen and X. Yao, *Mol. Cell. Proteomics*, 2008, **7**, 1598-1608.
3. Y. Xue, Z. Liu, X. Gao, C. Jin, L. Wen, X. Yao and J. Ren, *PLoS One*, 2010, **5**, e11290.
4. Z. Liu, J. Cao, Q. Ma, X. Gao, J. Ren and Y. Xue, *Mol.Biosyst.* 2011, **7**, 1197-1204.
5. Y. Xue, Z. Liu, J. Cao, Q. Ma, X. Gao, Q. Wang, C. Jin, Y. Zhou, L. Wen and J. Ren, *Protein Eng. Des. Sel.*, 2011, **24**, 255-260.
6. Y. Xue, F. Zhou, M. Zhu, K. Ahmed, G. Chen and X. Yao, *Nucleic Acids Res.*, 2005, **33**, W184-187.
7. F. F. Zhou, Y. Xue, G. L. Chen and X. Yao, *Biochem. Biophys. Res. Commun.*, 2004, **325**, 1443-1448.
8. W. Li and A. Godzik, *Bioinformatics*, 2006, **22**, 1658-1659.
9. M. Johnson, I. Zaretskaya, Y. Raytselis, Y. Merezhuk, S. McGinnis and T. L. Madden, *Nucleic Acids Res.*, 2008, **36**, W5-9.
10. M. Gribskov, A. D. McLachlan and D. Eisenberg, *Proc. Natl. Acad. Sci. U. S. A.*, 1987, **84**, 4355-4358.
11. D. Barrell, E. Dimmer, R. P. Huntley, D. Binns, C. O'Donovan and R. Apweiler, *Nucleic Acids Res.*, 2009, **37**, D396-403.
12. F. Zhou, Y. Xue, H. Lu, G. Chen and X. Yao, *FEBS Lett.*, 2005, **579**, 3369-3375.

Supplementary Tables

Supplementary Table S1 – We manually collected 127 experimentally identified pupylation sites in 109 unique proteins from the scientific literature (PubMed). *a*. UniProt, the UniProt accession numbers of pupylation substrates; *b*. Position, the positions of the pupylation sites; *c*. PMID, the primary references for the experimentally verified pupylation sites.

UniProt[a]	Position[b]	Organism	PMID[c]
A0QNF6	147	*M. smegmatis*	20631680
A0QP32	485	*M. smegmatis*	20094657
A0QP90	51	*M. smegmatis*	20094657
A0QPN2	408	*M. smegmatis*	20094657
A0QQ65	124	*M. smegmatis*	20631680
A0QQU5	116	*M. smegmatis*	20631680
A0QQU5	132	*M. smegmatis*	20066036;20094657
A0QS98	188	*M. smegmatis*	20631680
A0QSE0	96	*M. smegmatis*	20094657
A0QSL6	66	*M. smegmatis*	20631680
A0QSU4	99	*M. smegmatis*	20094657
A0QUV7	210	*M. smegmatis*	20631680
A0QUY3	313	*M. smegmatis*	20631680
A0QUY9	380	*M. smegmatis*	20631680
A0QUZ0	61	*M. smegmatis*	20631680
A0QV10	262	*M. smegmatis*	20094657
A0QVB9	36	*M. smegmatis*	20094657;20631680
A0QVB9	111	*M. smegmatis*	20631680
A0QVB9	131	*M. smegmatis*	20631680
A0QWV9	172	*M. smegmatis*	20631680
A0QWW2	257	*M. smegmatis*	20094657;20631680
A0QWX9	132	*M. smegmatis*	20094657
A0QWX9	219	*M. smegmatis*	20631680
A0QX20	394	*M. smegmatis*	20094657
A0QX81	41	*M. smegmatis*	20631680
A0QX93	355	*M. smegmatis*	20094657
A0QXX7	149	*M. smegmatis*	20631680
A0QYD4	43	*M. smegmatis*	20094657
A0QZA1	77	*M. smegmatis*	20066036
A0QZA1	109	*M. smegmatis*	20094657;20631680
A0QZE3	217	*M. smegmatis*	20094657
A0R066	299	*M. smegmatis*	20094657
A0R079	14	*M. smegmatis*	20094657
A0R0B3	58	*M. smegmatis*	20631680

A0R0B3	79	*M. smegmatis*	20094657;20631680
A0R0B4	53	*M. smegmatis*	20631680
A0R0B5	84	*M. smegmatis*	20631680
A0R0W1	458	*M. smegmatis*	20631680
A0R0W4	218	*M. smegmatis*	20631680
A0R1B5	115	*M. smegmatis*	20094657
A0R1V9	25	*M. smegmatis*	20631680
A0R1V9	29	*M. smegmatis*	20631680
A0R1V9	41	*M. smegmatis*	20094657;20631680
A0R1Y7	187	*M. smegmatis*	20094657;20631680
A0R218	320	*M. smegmatis*	20631680
A0R2G5	299	*M. smegmatis*	20631680
A0R2V7	362	*M. smegmatis*	20094657;20631680
A0R2W6	58	*M. smegmatis*	20631680
A0R342	36	*M. smegmatis*	20631680
A0R3D2	21	*M. smegmatis*	20631680
A0R4C9	67	*M. smegmatis*	20094657;20631680
A0R4Z5	218	*M. smegmatis*	20631680
A0R518	65	*M. smegmatis*	20094657
A0R566	11	*M. smegmatis*	20094657
A0R5E1	47	*M. smegmatis*	20094657
A0R5M8	242	*M. smegmatis*	20631680
A0R5R7	339	*M. smegmatis*	20094657;20631680
A0R635	375	*M. smegmatis*	20094657
A0R647	10	*M. smegmatis*	20094657
A0R6E3	30	*M. smegmatis*	20631680
A0R6E3	121	*M. smegmatis*	20094657
A0R6Q0	76	*M. smegmatis*	20094657
A0R7F9	33	*M. smegmatis*	20094657
A0R7G6	65	*M. smegmatis*	19028679;20066036;20094657
P0CG99	310	*M. smegmatis*	20094657
P53649	38	*M. smegmatis*	20094657;20631680
P53649	90	*M. smegmatis*	20094657
O05598	528	*M. tuberculosis*	20066036;20094657
O05814	173	*M. tuberculosis*	20066036;20094657
O06188	82	*M. tuberculosis*	20066036
O06391	136	*M. tuberculosis*	20066036
O33294	502	*M. tuberculosis*	20066036
O33341	289	*M. tuberculosis*	20066036
O53176	127	*M. tuberculosis*	20066036
O53204	338	*M. tuberculosis*	20066036
O53226	12	*M. tuberculosis*	20066036
O53226	150	*M. tuberculosis*	20066036
O53442	145	*M. tuberculosis*	20066036

O53618	65	*M. tuberculosis*	20066036
O53665	168	*M. tuberculosis*	20066036
O53665	381	*M. tuberculosis*	20066036
O53871	189	*M. tuberculosis*	20066036
O69687	227	*M. tuberculosis*	20066036
O69687	231	*M. tuberculosis*	20066036;20094657;20631680
O86352	283	*M. tuberculosis*	20066036
P09621	100	*M. tuberculosis*	20066036
P0A4X0	209	*M. tuberculosis*	20066036;20094657
P0A508	322	*M. tuberculosis*	20066036
P0A556	307	*M. tuberculosis*	20066036
P0A5B7	64	*M. tuberculosis*	20066036
P0A5B7	85	*M. tuberculosis*	20066036
P0A5B7	114	*M. tuberculosis*	20066036
P0A5B7	132	*M. tuberculosis*	20066036
P0A5H3	334	*M. tuberculosis*	20066036
P0A5L2	81	*M. tuberculosis*	20066036
P0A5U4	762	*M. tuberculosis*	20066036
P0A5Z4	204	*M. tuberculosis*	20066036
P0CG95	98	*M. tuberculosis*	20066036;20094657;20631680
P17670	202	*M. tuberculosis*	20066036
P60176	474	*M. tuberculosis*	20066036;20094657
P60796	271	*M. tuberculosis*	20066036;20631680
P63345	591	*M. tuberculosis*	20066036;20094657
P63458	173	*M. tuberculosis*	18832610;20066036;20094657
P63523	355	*M. tuberculosis*	20066036
P63568	314	*M. tuberculosis*	20066036;20094657
P63673	499	*M. tuberculosis*	20066036
P64245	363	*M. tuberculosis*	20066036;20094657
P65161	44	*M. tuberculosis*	20066036;20094657
P65232	29	*M. tuberculosis*	20066036
P65277	154	*M. tuberculosis*	20066036
P65573	45	*M. tuberculosis*	20066036
P65880	292	*M. tuberculosis*	20066036;20094657
P66056	101	*M. tuberculosis*	20066036;20094657
P66902	151	*M. tuberculosis*	20066036;20094657;20631680
P69440	23	*M. tuberculosis*	20066036;20631680
P69440	94	*M. tuberculosis*	20066036
P71724	47	*M. tuberculosis*	20066036
P71973	44	*M. tuberculosis*	20066036;20094657
P77899	345	*M. tuberculosis*	20066036;20094657
P77899	400	*M. tuberculosis*	20631680
P96382	362	*M. tuberculosis*	20066036
P96825	280	*M. tuberculosis*	20066036

Q10504	346	*M. tuberculosis*	20066036;20094657
Q10530	328	*M. tuberculosis*	20066036
Q10682	47	*M. tuberculosis*	20066036
Q50685	354	*M. tuberculosis*	20066036
Q7D8W0	428	*M. tuberculosis*	20066036

Supplementary Table S2 – From both large-scale and small-scale experimental studies we also collected 238 potentially pupylation substrates for which the exact pupylation sites had still not been experimentally determined. The default threshold (medium) was adopted for GPS-PUP 1.0.

UniProt	Predicted pupylation sites	Organism	PMID
A0QNZ3	115, 216, 264	*M. smegmatis*	20631680
A0QNZ7	157	*M. smegmatis*	20631680
A0QP06	241, 556, 559	*M. smegmatis*	20094657
A0QP11	395, 398, 440, 535	*M. smegmatis*	20631680
A0QPE7	96, 146, 164, 249, 254, 272, 377	*M. smegmatis*	20631680
A0QPE8	63, 186, 242, 399, 406	*M. smegmatis*	20094657;20631680
A0QQC8	226, 228, 483, 491, 538, 546, 556, 572, 619, 622	*M. smegmatis*	20631680
A0QQF0	124, 142, 199, 301, 432, 439, 444, 477, 560, 570, 821	*M. smegmatis*	20631680
A0QQF9	302	*M. smegmatis*	20631680
A0QQI6	14	*M. smegmatis*	20631680
A0QQJ4	10, 184, 335	*M. smegmatis*	20631680
A0QQL0	241	*M. smegmatis*	20094657;20631680
A0QQU1	65, 101	*M. smegmatis*	20631680
A0QQW8	64, 216, 220	*M. smegmatis*	20094657;20631680
A0QQX6	136, 169, 255, 334	*M. smegmatis*	20094657
A0QR00	219, 246, 247	*M. smegmatis*	20094657;20631680
A0QR08	69, 75, 275	*M. smegmatis*	20631680
A0QR33	95, 340	*M. smegmatis*	20631680
A0QR89	58, 110, 167, 307, 394, 395	*M. smegmatis*	20094657;20631680
A0QRM0	11, 108, 109, 112, 136	*M. smegmatis*	20631680
A0QRU5	98, 181, 255, 266	*M. smegmatis*	20094657
A0QS07	9, 201	*M. smegmatis*	20631680
A0QS36	96	*M. smegmatis*	20631680
A0QS46	31, 151	*M. smegmatis*	20094657;20631680
A0QS62	122, 168, 169	*M. smegmatis*	20631680
A0QS66	176, 470, 634, 775, 786	*M. smegmatis*	20631680
A0QS81	180, 239	*M. smegmatis*	20631680
A0QS85	368, 377, 550	*M. smegmatis*	20094657;20631680
A0QSD0	46	*M. smegmatis*	20094657;20631680
A0QSD1	30, 213, 217	*M. smegmatis*	20631680
A0QSD2	45, 94, 153, 210	*M. smegmatis*	20094657;20631680
A0QSD4	276, 277	*M. smegmatis*	20094657;20631680
A0QSD5	89	*M. smegmatis*	20631680
A0QSD7	40, 92	*M. smegmatis*	20631680
A0QSD8	124	*M. smegmatis*	20631680
A0QSG0	94, 99, 103	*M. smegmatis*	20631680

A0QSG1	55	*M. smegmatis*	20631680
A0QSG4	139, 176, 179	*M. smegmatis*	20631680
A0QSG5	4, 27, 126	*M. smegmatis*	20631680
A0QSG8	5, 128	*M. smegmatis*	20631680
A0QSH8	160	*M. smegmatis*	20094657;20631680
A0QSK7	143, 282	*M. smegmatis*	20631680
A0QSL5	65, 111, 120, 121	*M. smegmatis*	20631680
A0QSL8	191	*M. smegmatis*	20631680
A0QSP9	143, 149	*M. smegmatis*	20631680
A0QSS3	100	*M. smegmatis*	20094657
A0QSS4	74, 243, 267, 388, 389, 402, 426, 473, 523	*M. smegmatis*	20094657;20631680
A0QSX3	3, 42, 53, 109	*M. smegmatis*	20631680
A0QSY5	281	*M. smegmatis*	20631680
A0QSZ3	151, 323, 591, 634, 688, 692, 734	*M. smegmatis*	20094657;20631680
A0QT01	52	*M. smegmatis*	20094657;20631680
A0QT04	4, 72, 75, 492	*M. smegmatis*	20094657;20631680
A0QT08	36, 88, 583	*M. smegmatis*	20094657;20631680
A0QT22	100, 192	*M. smegmatis*	20631680
A0QTE1	11, 156, 165, 176, 394, 512, 598	*M. smegmatis*	20631680
A0QTE3	171	*M. smegmatis*	20631680
A0QTE7	73, 216, 405, 417, 473	*M. smegmatis*	20094657
A0QTK6	62, 116	*M. smegmatis*	20094657;20631680
A0QU00	115, 205, 264	*M. smegmatis*	20631680
A0QU53	130, 195, 406	*M. smegmatis*	20631680
A0QU58	37, 103	*M. smegmatis*	20631680
A0QU93	45, 158, 282	*M. smegmatis*	20631680
A0QUV6	76, 150, 234	*M. smegmatis*	20094657;20631680
A0QUX1	233, 402, 408, 412, 463, 481, 484	*M. smegmatis*	20631680
A0QUX7	67, 71, 170	*M. smegmatis*	20631680
A0QUX8	57, 290, 320	*M. smegmatis*	20094657;20631680
A0QUY2	142, 332, 469	*M. smegmatis*	20631680
A0QUY6	65, 256, 258, 259	*M. smegmatis*	20631680
A0QV37	12, 94, 125	*M. smegmatis*	20631680
A0QV45	8, 77, 94	*M. smegmatis*	20631680
A0QV51	232, 301	*M. smegmatis*	20094657
A0QVB1	278, 283	*M. smegmatis*	20631680
A0QVB8	58, 109, 132, 178, 215	*M. smegmatis*	20631680
A0QVE0	115, 125, 162, 176	*M. smegmatis*	20631680
A0QVL0	74, 325, 397	*M. smegmatis*	20631680
A0QVQ3	34	*M. smegmatis*	20631680
A0QVQ5	245, 260, 709, 746, 750	*M. smegmatis*	20094657
A0QVQ8	51, 230, 231, 357	*M. smegmatis*	20631680
A0QVR8	86	*M. smegmatis*	20631680

A0QVT1	66, 100	*M. smegmatis*	20631680
A0QVX6	121, 333, 342, 385, 400, 419, 452	*M. smegmatis*	20631680
A0QVY9	40	*M. smegmatis*	20631680
A0QVZ3	67, 226, 229	*M. smegmatis*	20094657;20631680
A0QW25	44, 229	*M. smegmatis*	20631680
A0QWG8	175, 213, 275	*M. smegmatis*	20094657;20631680
A0QWQ9	198, 314, 368	*M. smegmatis*	20094657
A0QWS8	17, 29, 101	*M. smegmatis*	20094657;20631680
A0QWT2	5, 410	*M. smegmatis*	20094657
A0QWT3	341, 396	*M. smegmatis*	20631680
A0QWV0	170, 207, 230	*M. smegmatis*	20094657;20631680
A0QWW3	4, 130, 140	*M. smegmatis*	20094657;20631680
A0QWW4	193	*M. smegmatis*	20631680
A0QWY0	255, 297, 600, 652	*M. smegmatis*	20094657
A0QX01	129	*M. smegmatis*	20631680
A0QX83	180	*M. smegmatis*	20631680
A0QX96	98, 125, 382, 410	*M. smegmatis*	20094657;20631680
A0QXA3	128, 259, 317, 431	*M. smegmatis*	20094657;20631680
A0QXC8	140, 239, 283, 286	*M. smegmatis*	20631680
A0QXD8	245, 343, 356	*M. smegmatis*	20094657
A0QXH9	118, 150	*M. smegmatis*	20631680
A0QY23	250	*M. smegmatis*	20631680
A0QYD5	304	*M. smegmatis*	20631680
A0QYE0	389	*M. smegmatis*	20094657
A0QYE8	230, 407	*M. smegmatis*	20094657
A0QYF5	652	*M. smegmatis*	20094657;20631680
A0QYF7	411	*M. smegmatis*	20094657
A0QYQ7	337, 387, 423, 585	*M. smegmatis*	20094657
A0QYS6	104, 224, 394	*M. smegmatis*	20094657;20631680
A0QYT2	139, 187, 280	*M. smegmatis*	20094657
A0QYY6	93, 178, 208, 243, 426, 434, 474	*M. smegmatis*	20631680
A0QZ33	49	*M. smegmatis*	20631680
A0QZ46	52, 186, 243	*M. smegmatis*	20631680
A0QZ47	213	*M. smegmatis*	20631680
A0QZ49	398	*M. smegmatis*	20094657
A0QZ54	76, 239, 344, 532	*M. smegmatis*	20631680
A0QZ83	63, 113, 136	*M. smegmatis*	20631680
A0QZ96	52, 56	*M. smegmatis*	20094657;20631680
A0QZE4	479	*M. smegmatis*	20094657
A0QZR5	29, 227	*M. smegmatis*	20631680
A0QZZ1	281	*M. smegmatis*	20094657
A0R012	55, 83, 212	*M. smegmatis*	20631680
A0R059	16, 32, 72	*M. smegmatis*	20094657;20631680

A0R061	23, 97	*M. smegmatis*	20631680
A0R067	57	*M. smegmatis*	20631680
A0R069	103, 221, 244, 255, 283	*M. smegmatis*	20631680
A0R072	24, 157, 208, 293, 303, 312, 352	*M. smegmatis*	20094657;20631680
A0R095	2	*M. smegmatis*	20631680
A0R0B2	61, 95, 106, 157, 170	*M. smegmatis*	20631680
A0R0E9	93	*M. smegmatis*	20631680
A0R0Q9	52, 464, 596	*M. smegmatis*	20631680
A0R0T8	21, 131, 216, 345	*M. smegmatis*	20631680
A0R0W7	278, 345	*M. smegmatis*	20631680
A0R170	12	*M. smegmatis*	20631680
A0R193	249, 316, 554, 749, 757	*M. smegmatis*	20094657
A0R198	145, 203	*M. smegmatis*	20631680
A0R199	33, 125, 235, 274, 324, 352, 469	*M. smegmatis*	20631680
A0R1G3	32, 48, 202	*M. smegmatis*	20631680
A0R1H2	10	*M. smegmatis*	20631680
A0R1H5	44, 425	*M. smegmatis*	20631680
A0R1H6	62, 103	*M. smegmatis*	20094657
A0R1Y8	35	*M. smegmatis*	20631680
A0R1Z6	81	*M. smegmatis*	20094657
A0R200	7, 466, 474	*M. smegmatis*	20094657;20631680
A0R202	84, 384, 418, 490, 491, 499, 535, 539, 542, 546	*M. smegmatis*	20094657;20631680
A0R203	49, 50, 60, 66, 71, 77, 128, 211, 329	*M. smegmatis*	20631680
A0R220	151, 308	*M. smegmatis*	20631680
A0R221	88, 105, 375, 394	*M. smegmatis*	20094657
A0R278	77	*M. smegmatis*	20631680
A0R2J4	64	*M. smegmatis*	20631680
A0R2T0	78, 149, 175, 312, 367, 368	*M. smegmatis*	20631680
A0R2U7	157	*M. smegmatis*	20631680
A0R2U8	171, 421, 465	*M. smegmatis*	20094657;20631680
A0R2V3	88	*M. smegmatis*	20631680
A0R2V4	257	*M. smegmatis*	20094657;20631680
A0R2V5	210, 250, 400	*M. smegmatis*	20094657
A0R2W9	71, 114, 284	*M. smegmatis*	20631680
A0R2X8	172, 327, 400, 427, 442	*M. smegmatis*	20631680
A0R2Y1	41	*M. smegmatis*	20094657;20631680
A0R3B8	119, 258, 398, 427	*M. smegmatis*	20094657;20631680
A0R3C8	29, 125, 187, 252	*M. smegmatis*	20094657;20631680
A0R3L1	447, 451	*M. smegmatis*	20094657;20631680
A0R3L4	173, 492	*M. smegmatis*	20094657
A0R3M3	88, 206, 272, 281, 284, 291	*M. smegmatis*	20094657;20631680
A0R3M4	32, 138, 207, 380, 387	*M. smegmatis*	20631680
A0R3N8	228, 248, 308, 311, 384	*M. smegmatis*	20094657;20631680

A0R3V8	2	*M. smegmatis*	20631680
A0R3Y5	179, 235	*M. smegmatis*	20094657;20631680
A0R417	284, 332, 336	*M. smegmatis*	20094657;20631680
A0R452	111	*M. smegmatis*	20631680
A0R461	247, 250	*M. smegmatis*	20631680
A0R472	34	*M. smegmatis*	20094657;20631680
A0R478	73, 116, 179, 225, 246	*M. smegmatis*	20094657;20631680
A0R4B3	317, 325	*M. smegmatis*	20631680
A0R4D0	9, 88	*M. smegmatis*	20631680
A0R4G4	318	*M. smegmatis*	20094657;20631680
A0R4S6	*103, 153, 251, 299, 340, 349*	*M. smegmatis*	20094657
A0R574	47, 126, 209, 309, 360, 446, 516, 526, 530, 540, 613, 788	*M. smegmatis*	20631680
A0R597	136, 158	*M. smegmatis*	20631680
A0R5C5	117, 305	*M. smegmatis*	20094657;20631680
A0R5H1	24, 152, 209	*M. smegmatis*	20631680
A0R5L6	153	*M. smegmatis*	20631680
A0R5M3	82, 139	*M. smegmatis*	20094657
A0R5N7	99, 141, 235	*M. smegmatis*	20631680
A0R5P4	108, 315	*M. smegmatis*	20094657
A0R5Q2	187, 188, 473, 514, 573	*M. smegmatis*	20094657
A0R5R5	104, 298	*M. smegmatis*	20631680
A0R5X8	68	*M. smegmatis*	20094657
A0R609	417, 483, 524, 555	*M. smegmatis*	20631680
A0R618	143, 270, 340, 561, 601	*M. smegmatis*	20094657;20631680
A0R652	38, 200	*M. smegmatis*	20631680
A0R683	385	*M. smegmatis*	20094657
A0R6Q7	38, 200	*M. smegmatis*	20094657;20631680
A0R710	229	*M. smegmatis*	20094657
A0R716	175, 197, 285	*M. smegmatis*	20631680
A0R7G8	138	*M. smegmatis*	20631680
A0R7I9	32, 186, 214	*M. smegmatis*	20094657;20631680
A4ZHU4	359, 434, 651, 692, 697	*M. smegmatis*	20094657
O33246	61	*M. tuberculosis*	20631680
O68447	9, 183, 334	*M. smegmatis*	20094657
P42829	56	*M. smegmatis*	20631680
Q9AFI5	84, 95, 119	*M. smegmatis*	20631680
Q9ZHC5	94, 116, 117, 127, 136, 145, 154, 163, 172, 181, 186, 199, 200, 205	*M. smegmatis*	20094657
A0QP01		*M. smegmatis*	20631680
A0QQC1		*M. smegmatis*	20631680
A0QQX4		*M. smegmatis*	20094657;20631680
A0QRA5		*M. smegmatis*	20631680
A0QRA6		*M. smegmatis*	20631680

A0QS97	*M. smegmatis*	20631680
A0QSG6	*M. smegmatis*	20631680
A0QSZ1	*M. smegmatis*	20094657;20631680
A0QTD7	*M. smegmatis*	20631680
A0QTK2	*M. smegmatis*	20094657
A0QU45	*M. smegmatis*	20631680
A0QV09	*M. smegmatis*	20631680
A0QVL2	*M. smegmatis*	20631680
A0QWY3	*M. smegmatis*	20631680
A0QXZ4	*M. smegmatis*	20631680
A0QYW6	*M. smegmatis*	20094657;20631680
A0QZ34	*M. smegmatis*	20094657;20631680
A0QZ58	*M. smegmatis*	20631680
A0QZA2	*M. smegmatis*	20094657
A0R033	*M. smegmatis*	20094657
A0R090	*M. smegmatis*	20094657
A0R0A1	*M. smegmatis*	20631680
A0R0B0	*M. smegmatis*	20631680
A0R0I8	*M. smegmatis*	20631680
A0R0R1	*M. smegmatis*	20631680
A0R0S1	*M. smegmatis*	20631680
A0R1Z9	*M. smegmatis*	20631680
A0R2E9	*M. smegmatis*	20631680
A0R2V1	*M. smegmatis*	20631680
A0R343	*M. smegmatis*	20631680
A0R3E3	*M. smegmatis*	20631680
A0R4B7	*M. smegmatis*	20631680
A0R4H0	*M. smegmatis*	20631680
A0R4H2	*M. smegmatis*	20631680
A0R623	*M. smegmatis*	20631680
A0R729	*M. smegmatis*	20094657;20631680

Supplementary Table S3 – The top 15 most enriched biological processes, molecular functions and cellular components of the pupylated substrates in *M. smegmatis*, respectively. *a*. the number of proteins annotated; *b*. the proportion of proteins annotated; *c*. E-ratio, enrichment ratio, the pupylation proportion in relation to the proteomic proportion.

Description of GO term	Pupylation		Proteome		E-ratio[c]	*P*-value
	Num.[a]	Per.[b]	Num.	Per.		
The top 15 most enriched biological processes						
Translation (GO:0006412)	31	11.61%	101	2.26%	5.14	4.84E-15
Cellular amino acid biosynthetic process (GO:0008652)	20	7.49%	63	1.41%	5.31	2.19E-10
Branched chain family amino acid biosynthetic process (GO:0009082)	8	3.00%	12	0.27%	11.16	5.88E-08
Tricarboxylic acid cycle (GO:0006099)	9	3.37%	18	0.40%	8.37	2.56E-07
Glycolysis (GO:0006096)	8	3.00%	16	0.36%	8.37	1.24E-06
Response to stress (GO:0006950)	11	4.12%	36	0.81%	5.12	4.43E-06
Protein folding (GO:0006457)	7	2.62%	16	0.36%	7.32	1.80E-05
Sulfate transport (GO:0008272)	4	1.50%	7	0.16%	9.57	3.77E-04
Cellular amino acid metabolic process (GO:0006520)	6	2.25%	18	0.40%	5.58	4.33E-04
Threonine biosynthetic process (GO:0009088)	3	1.12%	4	0.09%	12.56	8.06E-04
ATP synthesis coupled proton transport (GO:0015986)	4	1.50%	9	0.20%	7.44	1.23E-03
Proteasomal protein catabolic process (GO:0010498)	3	1.12%	5	0.11%	10.04	1.93E-03
One-carbon metabolic process (GO:0006730)	3	1.12%	5	0.11%	10.04	1.93E-03
Lipid biosynthetic process (GO:0008610)	5	1.87%	17	0.38%	4.92	2.50E-03
Proton transport (GO:0015992)	4	1.50%	11	0.25%	6.09	2.94E-03
The top 15 most enriched molecular functions						
Structural constituent of ribosome (GO:0003735)	26	9.74%	58	1.30%	7.50	1.84E-17
RRNA binding (GO:0019843)	20	7.49%	37	0.83%	9.05	1.06E-15
RNA binding (GO:0003723)	24	8.99%	78	1.75%	5.15	6.65E-12
Lyase activity (GO:0016829)	24	8.99%	153	3.42%	2.63	9.23E-06
Hydrogen ion transporting ATP synthase activity, rotational mechanism (GO:0046933)	4	1.50%	7	0.16%	9.57	3.77E-04
TRNA binding (GO:0000049)	5	1.87%	12	0.27%	6.98	4.10E-04
Pyridoxal phosphate binding (GO:0030170)	12	4.49%	69	1.54%	2.91	6.63E-04
Proton-transporting ATPase activity, rotational mechanism (GO:0046961)	3	1.12%	4	0.09%	12.56	8.06E-04
Acyltransferase activity (GO:0008415)	12	4.49%	73	1.63%	2.75	1.12E-03
Oxidoreductase activity, acting on the aldehyde or oxo group of donors, NAD or NADP as acceptor (GO:0016620)	3	1.12%	5	0.11%	10.04	1.93E-03
Threonine-type endopeptidase activity (GO:0004298)	2	0.75%	2	0.04%	16.74	3.56E-03
Succinate-CoA ligase (ADP-forming) activity (GO:0004775)	2	0.75%	2	0.04%	16.74	3.56E-03

NAD or NADH binding (GO:0051287)	7	2.62%	36	0.81%	3.26	4.68E-03
Thiosulfate sulfurtransferase activity (GO:0004792)	2	0.75%	3	0.07%	11.16	1.02E-02
Oxidoreductase activity, acting on the CH-NH2 group of donors, NAD or NADP as acceptor (GO:0016639)	2	0.75%	3	0.07%	11.16	1.02E-02
The top 15 most enriched cellular components						
Ribosome (GO:0005840)	26	9.74%	59	1.32%	7.38	3.12E-17
Ribonucleoprotein complex (GO:0030529)	25	9.36%	58	1.30%	7.22	2.51E-16
Cytoplasm (GO:0005737)	49	18.35%	268	6.00%	3.06	2.84E-13
Small ribosomal subunit (GO:0015935)	6	2.25%	8	0.18%	12.56	1.09E-06
Proton-transporting ATP synthase complex, catalytic core F(1) (GO:0045261)	4	1.50%	5	0.11%	13.39	5.94E-05
Proteasome complex (GO:0000502)	3	1.12%	3	0.07%	16.74	2.11E-04
Intracellular (GO:0005622)	31	11.61%	282	6.31%	1.84	5.61E-04
Proton-transporting two-sector ATPase complex, catalytic domain (GO:0033178)	2	0.75%	2	0.04%	16.74	3.56E-03
Proton-transporting two-sector ATPase complex (GO:0016469)	2	0.75%	2	0.04%	16.74	3.56E-03
Proteasome core complex (GO:0005839)	2	0.75%	2	0.04%	16.74	3.56E-03
Large ribosomal subunit (GO:0015934)	3	1.12%	7	0.16%	7.17	6.16E-03
Peroxisome (GO:0005777)	1	0.37%	1	0.02%	16.74	5.97E-02
Tricarboxylic acid cycle enzyme complex (GO:0045239)	1	0.37%	1	0.02%	16.74	5.97E-02
Proteasome core complex, alpha-subunit complex (GO:0019773)	1	0.37%	1	0.02%	16.74	5.97E-02
Protein complex (GO:0043234)	1	0.37%	1	0.02%	16.74	5.97E-02

Supplementary Table S4 – The differences among various versions of GPS series algorithms. First, the scoring strategy was reserved in any release of GPS algorithms. For performance improvement, the Markov Cluster Algorithm (MCL for short) was adopted in GPS 1.0 & 1.10 to classify known phosphorylation sites into several clusters.[6-7] This method was not used in later versions for its low efficiency. In the latest 3.0 version, the *k*-means clustering was adopted to cluster known PTM sites if the data set is large.[3-4] However, due to the data limitation, this approach was not included in this study, while the GPS 2.2 algorithm contains a sequential three-step procedure of MLS, WT and MaM for performance improvement.

Algorithm	Performance improvement	Ref.
GPS 1.0 & 1.10	MCL	[6-7]
GPS 2.0	MaM	[2]
GPS 2.1	MLS & MaM	[5]
GPS 2.2	MLS, WT & MaM	In this study
GPS 3.0	*k*-means clustering, MLS, WT & MaM	[3-4]