Molecular BioSystems

Cite this: *Mol. BioSyst.*, 2011, **7**, 2737–2740

www.rsc.org/molecularbiosystems

COMMUNICATION

GPS-PUP: computational prediction of pupylation sites in prokaryotic proteins†‡

Zexian Liu,§[ab] Qian Ma,§[a] Jun Cao,[a] Xinjiao Gao,[a] Jian Ren*[c] and Yu Xue*[ab]

Received 1st June 2011, Accepted 27th July 2011
DOI: 10.1039/c1mb05217a

Recent experiments revealed the prokaryotic ubiquitin-like protein (PUP) to be a signal for the selective degradation of proteins in *Mycobacterium tuberculosis* (Mtb). By covalently conjugating the PUP, pupylation functions as a critical post-translational modification (PTM) conserved in actinomycetes. Here, we designed a novel computational tool of GPS-PUP for the prediction of pupylation sites, which was shown to have a promising performance. From small-scale and large-scale studies we collected 238 potentially pupylated substrates for which the exact pupylation sites were still not determined. As an example application, we predicted ~85% of these proteins with at least one potential pupylation site. Furthermore, through functional analysis, we observed that pupylation can target various substrates so as to regulate a broad array of biological processes, such as the response to stress, sulfate and proton transport, and metabolism. The prediction and analysis results prove to be useful for further experimental investigation. The GPS-PUP 1.0 is freely available at: http://pup.biocuckoo.org.

The Nobel Prize in Chemistry 2004 was awarded to Aaron Ciechanover, Avram Hershko and Irwin Rose for their discovery of ubiquitin-mediated protein degradation in eukaryotes. Numerous studies subsequent to their original work showed that the ubiquitin-proteasome system (UPS) plays a critical role in regulating a variety of cellular processes such as the cell cycle and division, the immune response, inflammation, and signal transduction.[1] Recently, the PUP was identified as a tag for selective degradation of proteins in Mtb.[2] Further analysis proposed that PUP-mediated pupylation might be a ubiquitous PTM in actinomycetes, which have a conserved proteasome system.[3] In contrast with the three-step biochemical reaction of eukaryotic ubiquitination with E1, E2 and E3 ligases,[1] the prokaryotic pupylation is much simpler, having only two steps.[3–5] The PUP-GGQ C-terminal is first deamidated to -GGE by Dop/PafD (PUP deamidase/depupylase), and then conjugated to specific lysine residues of substrates catalyzed by PafA (PUP ligase, PUP-conjugating enzyme) (Fig. 1).[3–5] Since the proteasomal pathway is critical for both the virulence and persistence of Mtb, identification of the pupylated substrates along with information on the exact sites is fundamental for understanding the pathological mechanisms,[6] and can provide helpful insights into protein degradation in actinomycetes.[3]

In 2008, Pearce *et al.* experimentally identified the first pupylated substrate of FabD (Malonyl CoA-acyl carrier protein transacylase) in Mtb, with K73 being the major pupylation site.[2] Later, Festa *et al.* carried out a large-scale analysis of the Mtb pupylome with tandem mass spectrometry (MS/MS), with the result that 60 pupylation sites in 55 proteins were detected.[7] Recently, two proteome-wide analyses revealed several hundreds of potential pupylated substrates in the model organism *Mycobacterium smegmatis*, in which pupylation-mediated protein selective degradation was proposed to be highly dynamic and dependent on the culture conditions.[5,8] At present, experimental determination of pupylated substrates with exact modified sites is still a great challenge, and no canonical sequence motifs have been observed.[5] In contrast to labor-intensive and time-consuming experimental approaches, the computational prediction of putative pupylation sites can greatly narrow down the number of potential candidates, and thus rapidly

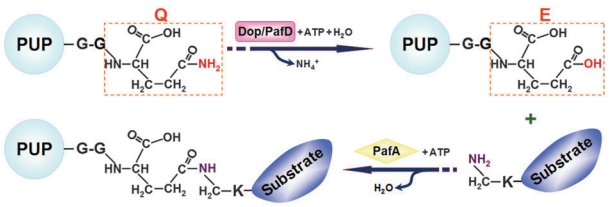

Fig. 1 The biochemical process of pupylation. The PUP is firstly deamidated at its C-terminal by Dop/PafD, and the last glutamine (Q) is changed to a glutamic acid (E). Then the activated PUP is conjugated to the specific lysine residues of the substrates catalyzed by PafA.

[a] *Hefei National Laboratory for Physical Sciences at Microscale and School of Life Sciences, University of Science & Technology of China, Hefei, Anhui 230027, China. E-mail: xueyu@mail.hust.edu.cn, xueyuhust@gmail.com; Fax: +86-27-87793172; Tel: +86-27-87793903*
[b] *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*
[c] *State Key Laboratory of Biocontrol, School of Life Sciences, Sun Yat-sen University, Guangzhou, Guangdong 510275, China. E-mail: renjian.sysu@gmail.com; Fax: +86-20-39943788; Tel: +86-20-39943788*

† Published as part of a Molecular BioSystems themed issue on Computational Biology: Guest Editor Michael Blinov.

‡ Electronic supplementary information (ESI) available. See DOI: 10.1039/c1mb05217a

§ The two authors contributed equally to this work.

provide useful information for further experimental investigation. In this regard, an accurate and convenient predictor of pupylation is urgently needed.

In this work, we manually collected 127 experimentally identified pupylation sites in 109 prokaryotic proteins from the scientific literature. Previously, we had developed the GPS 2.1 algorithm (Group-based Prediction System) for the prediction of kinase-specific phosphorylation sites, in which two sequential steps of motif length selection (MLS) and matrix mutation (MaM) were adopted for performance improvement.[9] Later, for the prediction of nitration sites, we designed the GPS 3.0 algorithm by introducing two additional approaches of *k*-means clustering and weight training (WT), with the former allowing a classification of PTM sites into several different clusters if the training data set is sufficiently large.[10] Since the number of pupylation sites was quite limited, the *k*-means clustering was not adopted in this analysis. In this regard, we developed the GPS 2.2 algorithm for the purpose of detecting pupylation, with a sequential three-step procedure of MLS, WT, and MaM. This training order cannot be changed by extensive testing. More detailed information on the data preparation and algorithm implementation are provided in Materials & Methods (ESI‡).

After the training procedures, we developed the GPS-PUP software in order to predict the pupylation sites. Although the optimal peptide from the MLS was determined to be PSP(8, 18), PSP(7, 7) was shown in the final results for practical convenience. The prediction results of Mtb hspX (alpha-crystallin, UniProt ID: P0A5B7) are presented as an example (Fig. 2). As a chaperone required for mycobacterial persistence within the macrophage,[11] hspX was determined to be pupylated on the four lysines of K64, K85, K114 and K132.[7] Utilizing the default threshold (medium), GPS-PUP predicted all of the four sites as positive hits (Fig. 2). In addition, K119 was also predicted as a positive hit, which might be useful for further experimental investigation.

To evaluate the prediction performance and robustness of GPS-PUP, the leave-one-out (LOO) validation and 4-, 6-, 8- and 10-fold cross-validations were calculated. The performance of GPS-PUP was promising, with an accuracy (Ac) under the low threshold condition of 78.85%, a precision (Pr) of 22.56%, a sensitivity (Sn) of 63.78% and a specificity (Sp) of 80.21% (Table 1). Receiver Operating Characteristic (ROC) curves were drawn, and the AROC (area under the ROC) values were calculated as 0.708 (LOO), 0.702 (4-fold), 0.731 (6-fold), 0.755 (8-fold) and 0.749 (10-fold), respectively (Fig. 3A). Since the results of the 4-, 6-, 8- and 10-fold cross-validations were similar to the LOO validation, GPS-PUP was demonstrated to be robust for the prediction of pupylation sites.

Table 1 Comparison of the GPS 2.2 with the other algorithms, including GPS 2.1 and PSSM. For the construction of the GPS-PUP software, the three thresholds, high, medium and low, were utilized. We fixed the Sp values of GPS 2.2 so as to be identical or highly similar to the other methods and compared the Sn values

Method	Threshold	Ac (%)	Sn (%)	Sp (%)	Pr (%)	MCC
GPS 2.2	High	85.44	33.07	90.18	23.33	0.1991
	Medium	82.51	44.88	85.91	22.35	0.2279
	Low	78.85	63.78	80.21	22.56	0.2864
GPS 2.1		85.25	31.50	90.11	22.35	0.1854
		81.40	37.80	85.34	18.90	0.1715
		77.35	44.09	80.36	16.87	0.1636
PSSM		84.20	14.17	90.53	11.92	0.0435
		79.96	24.41	84.98	12.81	0.0710
		76.63	35.43	80.36	14.02	0.1070

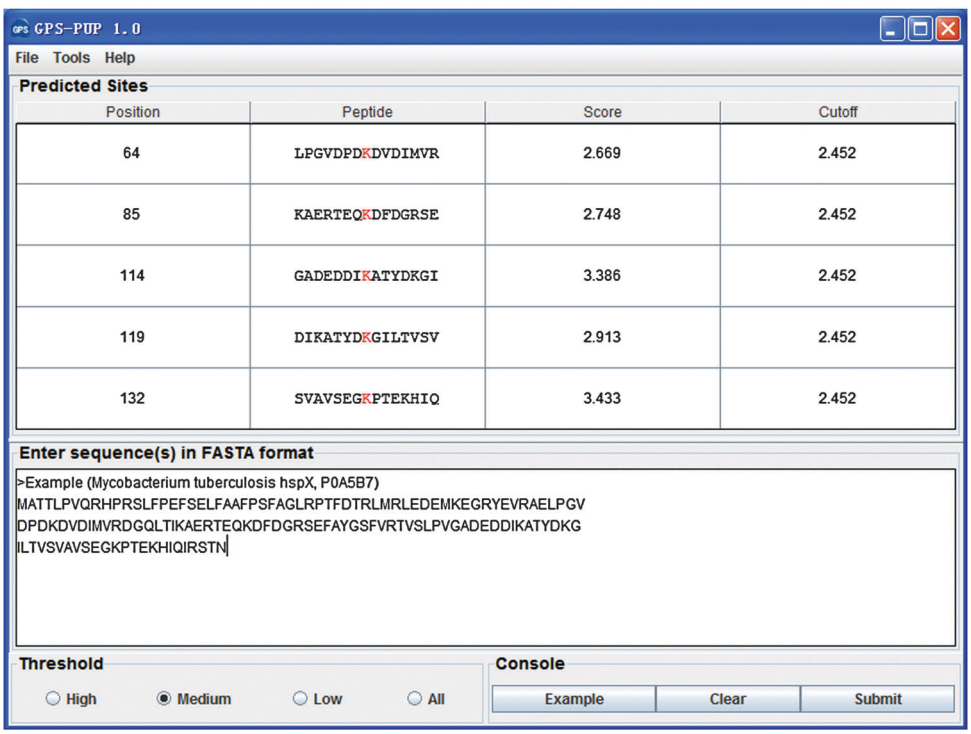

Fig. 2 Screen snapshot of the GPS-PUP 1.0 software. The default threshold (medium) was chosen. As an example, the prediction results for *Mycobacterium tuberculosis* hspX (P0A5B7) are shown.

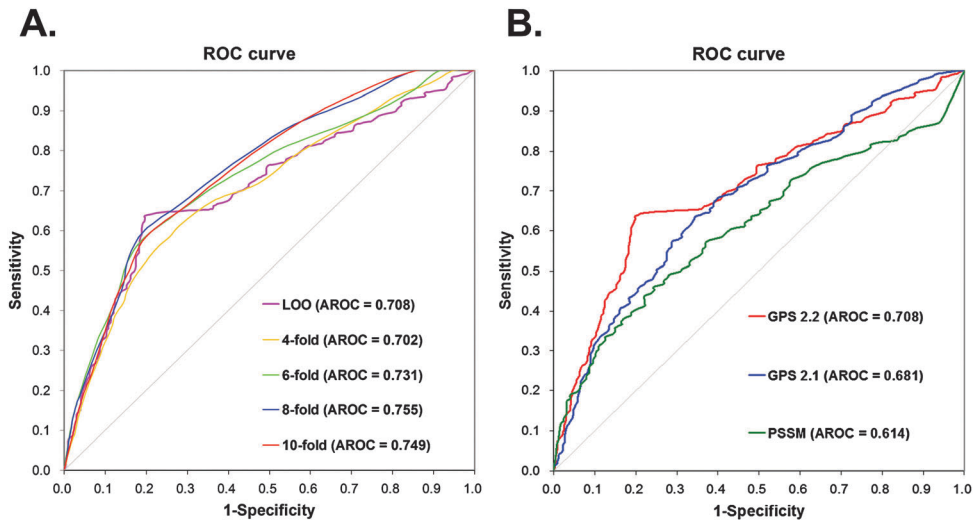

Fig. 3 The prediction performance of GPS-PUP 1.0. (A) The LOO validation and 4-, 6-, 8- and 10-fold cross-validations were performed. ROC curves were drawn, and the AROC values were calculated; (B) comparison of GPS 2.2, GPS 2.1 and PSSM using the LOO validation.

For comparison, the LOO validations for the GPS 2.2, GPS 2.1 and PSSM (Position-Specific Scoring Matrix) algorithms were calculated (more details are provided in Materials & Methods, ESI‡). Again, the ROC curves were drawn, and the AROC values were calculated as 0.708 (GPS 2.2), 0.681 (GPS 2.1) and 0.614 (PSSM), respectively (Fig. 3B). Furthermore, we fixed the Sp values of GPS 2.2 so as to be identical with the other methods, and then compared the Sn values (Table 1). The GPS 2.2 algorithm was demonstrated to be obviously much better than the other methods, especially in regions having high Sp values (Fig. 3B, Table 1).

Previously, large-scale analyses of the pupylome in *Mycobacterium smegmatis* have identified several hundreds of potentially pupylated targets, while the *bona fide* pupylation sites in most of these proteins still have yet to be elucidated.[5,8] In an application of GPS-PUP, 238 potentially pupylated proteins were collected from large-scale and small-scale studies (Table S2, ESI‡). Utilizing the medium cut-off, we predicted that 202 (~85%) of these targets had at least one potential pupylation site (Table S2, ESI‡). Several prediction results were randomly selected and are shown in Fig. 4. For example, sucC (the beta subunit of Succinyl-CoA ligase) was experimentally identified as a candidate pupylation target.[8] Here, we predicted that sucC might be pupylated at K32, K138, K207, K380 and/or K387 (Fig. 4A). Since sucC catalyzes the formation of succinyl-CoA from succinate and CoA as the critical step in the citric acid cycle,[12] we propose that pupylation might regulate the metabolism by directly targeting sucC for degradation. Also, the chaperone protein DnaJ, which plays an important role in the response to thermal stress,[13] was found to be pupylated.[8] With GPS-PUP, we predicted that K21, K131, K216 and K345 might be the major pupylation sites (Fig. 4B). Moreover, map (Proteasome-associated ATPase) and frr (Ribosome-recycling factor) were also demonstrated to be

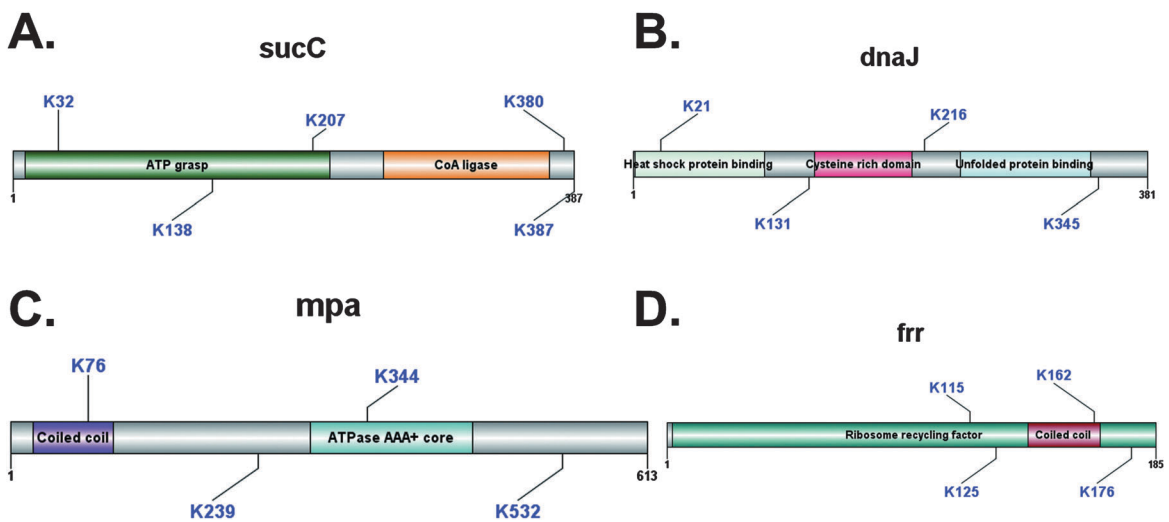

Fig. 4 Applications of GPS-PUP 1.0. We predicted potential pupylation sites in the potentially pupylated proteins of *M. smegmatis* with the default cut-off value. (A) The Succinyl-CoA ligase/sucC (A0R3M4); (B) the chaperone protein DnaJ (A0R0T8); (C) the Proteasome-associated ATPase/mpa (A0QZ54); (D) the Ribosome-recycling factor/frr (A0QVE0).

potential pupylation targets,[8] and the predicted pupylation sites for the two proteins are shown in Fig. 4C and D, respectively. Taken together, our prediction results can serve as a useful reservoir of information for further experimental consideration.

To obtain a better understanding of the functional complexity and diversity of the pupylome, we took 267 of the non-redundant pupylated substrates in *M. smegmatis* from Tables S1 and S2 (ESI‡), and statistically analyzed the over-represented biological processes, molecular functions and cellular components using gene ontology (GO) annotations (Table S3, ESI‡). Previous studies had implicated pupylated proteins in metabolic processes, such as glycolysis, the metabolism and biosynthesis of amino acids and lipids as well as translation.[5,8] In our results, the GO terms of the metabolism-related biological processes were significantly presented, such as translation (GO:0006412), cellular amino acid biosynthetic process (GO:0008652), and glycolysis (GO:0006096) (Table S3, ESI‡). Thus, this statistical analysis is consistent with previous reports. Besides, several non-metabolic GO terms were also detected, such as the response to stress (GO:0006950), sulfate transport (GO:0008272) and proton transport (GO:0015992). Moreover, by analyzing the enriched GO terms of the molecular functions and cellular components, the functions and localizations of pupylated substrates are evidently highly diverse (Table S3, ESI‡). In this regard, it is suggested that pupylation targets a variety of substrates so as to regulate a broad range of biological processes in addition to metabolism.

Taken together, these findings suggest that the novel software packages of GPS-PUP can serve as a powerful tool to help identify pupylation sites in prokaryotic proteins. Also, the large-scale prediction and functional analysis results can be used for the further investigation of molecular mechanisms of pupylation. Due to the limitation of the training data set, although the Sn and Sp values of GPS-PUP are promising, the Pr scores are only 23.33% (high threshold), 22.35% (medium threshold), and 22.56% (low threshold) (Table 1). We propose that the Pr values might be underestimated, because the negative data (−) could still contain some real pupylation sites, which are not experimentally identified. Also, we will continuously improve the software when new experimental data are available. We believe that computational prediction backed up with subsequent experimental identification can propel systematic studies of the pupylome into a new and highly productive phase.

Acknowledgements

The authors are grateful to two anonymous reviewers, whose suggestions have greatly improved the presentation of this manuscript. This work was supported by grants from the National Basic Research Program (973 project) (2010CB945400, 2011CB910400, 2011CB711000), 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; USTC: WK2340000032). Pacific Edit reviewed the manuscript prior to submission.

The authors have declared no conflict of interest.

References

1 *Protein Degradation: Cell Biology of the Ubiquitin-Proteasome System*, ed. R. John Mayer, A. J. Ciechanover and M. Rechsteiner, Wiley-VCH, 2006.
2 M. J. Pearce, J. Mintseris, J. Ferreyra, S. P. Gygi and K. H. Darwin, *Science*, 2008, **322**, 1104–1107.
3 K. H. Darwin, *Nat. Rev. Microbiol.*, 2009, **7**, 485–491.
4 F. Striebel, F. Imkamp, M. Sutter, M. Steiner, A. Mamedov and E. Weber-Ban, *Nat. Struct. Mol. Biol.*, 2009, **16**, 647–657.
5 J. Watrous, K. Burns, W. T. Liu, A. Patel, V. Hook, V. Bafna, C. E. Barry, 3rd, S. Bark and P. C. Dorrestein, *Mol. BioSyst.*, 2010, **6**, 376–385.
6 P. Salgame, *Cell Host Microbe*, 2008, **4**, 415–416.
7 R. A. Festa, F. McAllister, M. J. Pearce, J. Mintseris, K. E. Burns, S. P. Gygi and K. H. Darwin, *PLoS One*, 2010, **5**, e8589.
8 C. Poulsen, Y. Akhter, A. H. Jeon, G. Schmitt-Ulms, H. E. Meyer, A. Stefanski, K. Stuhler, M. Wilmanns and Y. H. Song, *Mol. Syst. Biol.*, 2010, **6**, 386.
9 Y. Xue, Z. Liu, J. Cao, Q. Ma, X. Gao, Q. Wang, C. Jin, Z. Zhou, L. Wen and J. Ren, *Protein Eng., Des. Sel.*, 2011, **24**, 255–260.
10 Z. Liu, J. Cao, Q. Ma, X. Gao, J. Ren and Y. Xue, *Mol. BioSyst.*, 2011, **7**, 1197–1204.
11 Y. Yuan, D. D. Crane, R. M. Simpson, Y. Q. Zhu, M. J. Hickey, D. R. Sherman and C. E. Barry, 3rd, *Proc. Natl. Acad. Sci. U. S. A.*, 1998, **95**, 9578–9583.
12 P. H. Pearson and W. A. Bridger, *J. Biol. Chem.*, 1975, **250**, 8524–8529.
13 S. P. Acebron, V. Fernandez-Saiz, S. G. Taneva, F. Moro and A. Muga, *J. Biol. Chem.*, 2007, **283**, 1381–1390.

