TECHNICAL BRIEF

Systematic study of protein sumoylation: Development of a site-specific predictor of SUMOsp 2.0

Jian Ren[1], Xinjiao Gao[1], Changjiang Jin[1], Mei Zhu[1], Xiwei Wang[1], Andrew Shaw[2], Longping Wen[1], Xuebiao Yao[1] and Yu Xue[1]*

[1] Hefei National Laboratory for Physical Sciences at Microscale and School of Life Sciences, University of Science and Technology of China, Hefei, Anhui, P. R. China
[2] Department of Physiology and Cancer Biology Program, Morehouse School of Medicine, Atlanta, GA, USA

Received: August 5, 2008
Revised: February 8, 2009
Accepted: March 11, 2009

Protein sumoylation is an important reversible post-translational modification on proteins, and orchestrates a variety of cellular processes. Recently, computational prediction of sumoylation sites has attracted much attention for its cost-efficiency and power in genomic data mining. In this work, we developed SUMOsp 2.0, an accurate computing program with an improved group-based phosphorylation scoring algorithm. Our analysis demonstrated that SUMOsp 2.0 has greater prediction accuracy than SUMOsp 1.0 and other existing tools, with a sensitivity of 88.17% and a specificity of 92.69% under the medium threshold. Previously, several large-scale experiments have identified a list of potential sumoylated substrates in *Saccharomyces cerevisiae* and *Homo sapiens*; however, the exact sumoylation sites in most of these proteins remain elusive. We have predicted potential sumoylation sites in these proteins using SUMOsp 2.0, which provides a great resource for researchers and an outline for further mechanistic studies of sumoylation in cellular plasticity and dynamics. The online service and local packages of SUMOsp 2.0 are freely available at: http://sumosp.biocuckoo.org/.

Keywords:
Group-based Phosphorylation Scoring / Potential sumoylation peptide / SUMO / Sumoylation

The past decade has witnessed rapid progress in the functional dissection of protein sumoylation [1–5]. Proteins modified by SUMO could alter their sub-cellular localization, activity, stability, *etc.* [1–5]. In addition, protein sumoylation plays important roles in a variety of cellular processes such as transcriptional regulation and signaling transduction [1–5]. In addition, sumoylation is essential for cell plasticity as aberrant sumoylation is implicated in numerous diseases and cancer development [6, 7]. Identification of SUMO substrates with their acceptor sumoylation sites is the foundation for understanding the molecular mechanisms and regulatory roles of sumoylation. In contrast to labor-intensive and costly experimental approaches, computational prediction of sumoylation sites *in silico* also attracted much attention for its accuracy, convenience and speed [8].

Previously, we mainly used a Group-based Phosphorylation Scoring (GPS) algorithm to design a convenient online tool, SUMOsp 1.0 [8]. The process of the construction of SUMOsp 1.0 is shown in Fig. 1. By literature mining, we manually collected 239 experimentally verified sumoylation sites in 144 proteins [8]. We defined a *potential sumoylation peptide* (*PSP*)(*m*, *n*) as a lysine (K) residue flanked by *m* residues upstream and *n* residues downstream. In SUMOsp 1.0, the *PSP*(7, 7) was arbitrarily employed. Based on the

Correspondence: Professor Yu Xue, Hefei National Laboratory for Physical Sciences at Microscale and School of Life Sciences, University of Science and Technology of China, Hefei, Anhui 230027, P. R. China
E-mail: xueyu@ustc.edu.cn
Fax: +86-551-3607821

Abbreviations: AC, accuracy; **GPS,** Group-based Phosphorylation Scoring; **MaM,** matrix mutation; **MCC,** Mathew correlation coefficient; **PSP,** potential sumoylation peptide; **Sn,** sensitivity; **Sp,** specificity

*Additional corresponding author: Professor Xuebiao Yao
E-mail: yaoxb@ustc.edu.cn

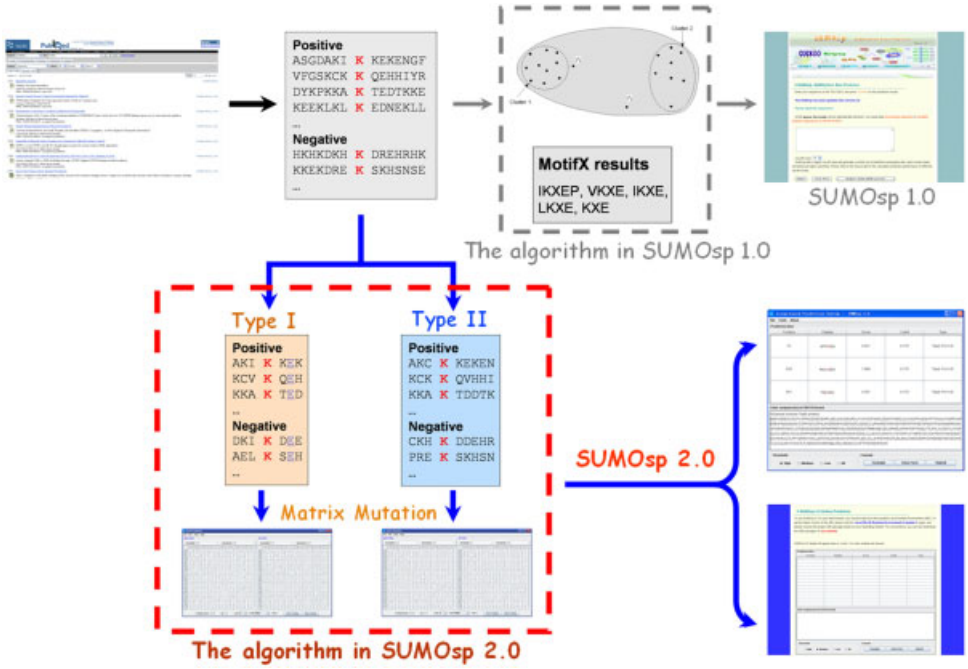

Figure 1. The procedures of constructions of SUMOsp 1.0 and SUMOsp 2.0. The process of data preparation was not changed, as the training data set was collected from PubMed and separated into positive data (+) and negative data (−). In SUMOsp 1.0, the GPS algorithm was mainly used and all experimentally verified sumoylation sites were automatically clustered into three groups by different thresholds of peptides similarity [8]. The predicted results of MotifX were also integrated in SUMOsp 1.0 [9]. In SUMOsp 2.0, we classified the training data into either Type I (consensus) or Type II sites. Then we used a simple approach of MaM to further improve the prediction performance.

hypothesis of similar peptides bearing similar biological functions, we simply used the amino acid substitution matrix BLOSUM62 to calculate the similarity between two *PSP*(*m*, *n*) peptides. Then the experimental sumoylation sites were automatically clustered into three groups according to different thresholds of peptides similarity (Fig. 1). Finally, the predicted results from MotifX were also integrated [9].

In this work, we updated our previous GPS algorithm with two major improvements. First, we used a much simpler approach to cluster sumoylation sites into groups. Based on the experimental observations, we directly classified the known sumoylation sites into two clusters, including Type I (consensus) and Type II (non-consensus) sites (Fig. 1). Type I sites followed the ψKXE (ψ is A, I, L, M, P, F, or V and X is any amino acid residue) motif [1–5], while Type II sites contained other non-canonical sites. Given a protein sequence for prediction, SUMOsp 2.0 will initially scan the sequence to separate all lysine sites into either Type I or Type II sites. Then, the Type I sites with *PSP*(3, 3) will be scored with experimental consensus sites, while the Type II sites with *PSP*(3, 5) will be scored with known non-consensus sites. During analysis of phosphorylation site prediction, we found that different amino acid matrices performed variably would generate various performances [10]. Thus, we were interested in testing whether we could find an optimal or near-optimal matrix for sumoylation in order to improve the prediction performance. In this regard, we further developed a simple method of matrix mutation (MaM) to automatically mutate BLOSUM62 into a near-optimal matrix for Type I and Type II sumoylation sites, respectively (Fig. 1). To evaluate the prediction performance, four standard measurements including accuracy (Ac), sensitivity (Sn), specificity (Sp) and the Mathew correlation coefficient (MCC) were calculated. Through exhaustive testing, we fixed the Sp at 85% to improve the Sn by Ma M for each cluster. The detailed information of the modified GPS algorithm is available in Supporting Information. Furthermore, we used the same data set to train SUMOsp 2.0 and SUMOsp 1.0, and compared their performances (Table 1). The superior performance of SUMOsp 2.0 proposed that the upgraded GPS algorithm was much better than our previous approach.

To construct the SUMOsp 2.0 software, we collected 332 non-redundant sumoylation sites in 197 proteins, by searching the research articles published before October 18, 2007. We arbitrarily took 279 sumoylation sites from 166 proteins published before February 2007 as the training data set for SUMOsp 2.0 and the remnant 53 sites in 31 proteins were not included in training as an additional data set for the performance evaluation. The evaluations of self-consistency, leave-one-out validation and four-, six-, eight-, tenfold cross-validations were performed on the training data set. By comparison, the SUMOsp 2.0 exhibited greater accuracy over other the existing tools. We chose three thresholds with high, medium and low stringencies for SUMOsp 2.0. The performance under medium threshold is 92.52% (Ac), 88.17% (Sn), 92.69% (Sp) and 0.5083 (MCC). Finally, the online service and local packages of SUMOsp 2.0 were implemented in JAVA 1.4.2 (J2SE). The detailed results on data preparation, performance evaluation and comparison are available in Supporting Information.

As applications of SUMOsp 2.0, we carried out large-scale predictions of sumoylation sites in *Saccharomyces cerevisiae*

Table 1. Comparisons of SUMOsp 2.0 with SUMOsp 1.0[a)]

	Cut-off	Ac (%)	Sn (%)	Sp (%)	MCC
SUMOsp 1.0	1.5	55.28	97.47	53.53	0.20
	4	80.43	89.12	80.07	0.32
	18	92.71	83.68	93.08	0.50
SUMOsp 2.0		83.44	97.86	82.84	0.39
		94.40	89.74	94.59	0.58
		95.79	83.76	96.29	0.62

a) The same data set containing 239 experimentally verified sumoylation sites in 144 proteins was used for training and testing [8]. We fixed Sn of SUMOsp 2.0 to be similar with that used in SUMOsp 1.0 to compare the Ac, Sn and MCC values.

and *Homo sapiens*. Owing to the great progress of MS techniques in large-scale identifications, there were six and seven proteome-scale *in vivo* and *in vitro* experiments carried out to identify 562 and 354 potential sumoylated substrates in *S. cerevisiae* [11–16] and *H. sapiens* [17–23], respectively. Although several proteins in these experiments may not be sumoylated as negative hits, a large proportion of them will be real sumoylated substrates with high confidence. However, the exact sumoylation sites in most of these proteins were still not identified. In this work, we used SUMOsp 2.0 with high threshold to predict potential sumoylation sites in these proteins. In *S. cerevisiae*, there were 403 proteins (~71%) with at least one potential sumoylation site predicted (Fig. 2A). There were only three and six proteins identified in six and five experiments, respectively, and all of these proteins were predicted with at least one potential sumoylation site (Fig. 2A). In addition, there were 21, 44, 131 and 357 potential sumoylated proteins detected in four, three, two and one experiments, respectively. Among these substrates, there were 17 (~81%), 31 (~70%), 109 (~83%) and 237 (~66%) proteins predicted with positive hits (Fig. 2A) and in *H. sapiens*, we predicted 238 (~67%) proteins with at least one potential sumoylation site. Only one protein was identified in four experiments but was missed by SUMOsp 2.0 (Fig. 2B). Only nine proteins were identified in three experiments, while six of them were predicted as positive hits with at least one site (Fig. 2B). In addition, there were 33 and 311 potential sumoylated proteins detected in two and one experiments, respectively. Among these substrates, there were 24 (~73%) and 208 (~69%) proteins predicted with positive hits (Fig. 2B). Thus, the prediction performance of SUMOsp 2.0 surpassed the expectations for large-scale predictions and our analyses generated a high-profile reservoir of potential sumoylation sites for further experimental consideration. The detailed analysis of large-scale prediction is available in Supporting Information.

Taken together, we propose that SUMOsp 2.0 will be a powerful tool for the identification of sumoylation sites. The combination of computational analyses with experimental verification will become the foundation of systematically understanding the mechanisms and the dynamics of sumoylation.

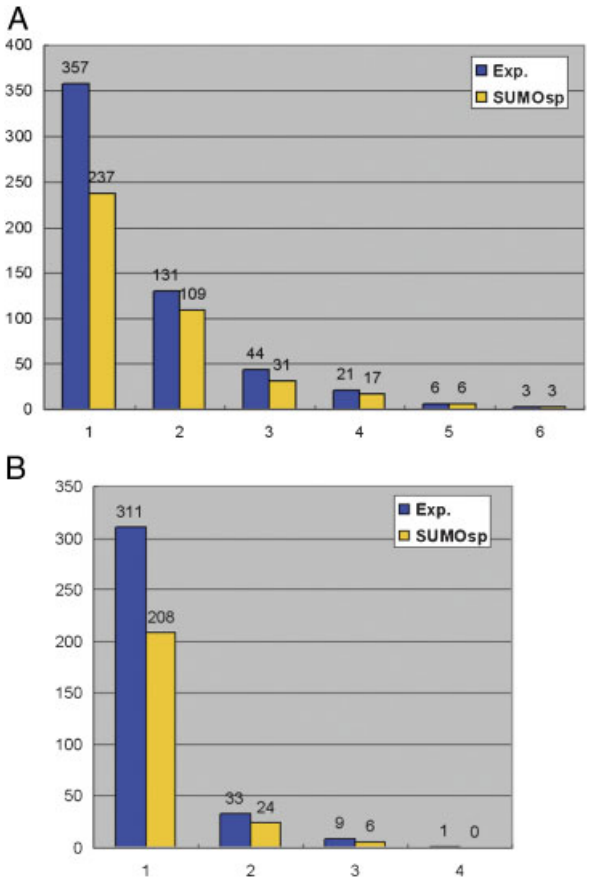

Figure 2. The potential sumoylated substrates in large-scale experiments *versus* SUMOsp 2.0 predicted his. (A) In *S. cerevisiae*, there were six high-throughput experiments carried out to identify 562 potential sumoylated substrates [11–16], while 403 (~71%) of them were predicted with at least one sumoylation site. (B) In *H. sapiens*, there were seven large-scale experiments to identify 354 potential sumoylated targets [17–23], and 238 (~67%) of them were predicted with at least one site.

The authors thank Kai Yuan, Dezhi Hou, Yu Yao, Dr. Martin Offterdinger (Innsbruck, Austria), Dr. Adi Avni (Tel-Aviv, Israel) and Dr. Yair Benita (Harvard, USA) for their constructive suggestions. The authors also thank two anonymous reviewers for their helpful comments. This work was supported by grants from the National Basic Research Program (973 project) (2006CB933300, 2007CB947401), National Natural Science Foundation of China (90919001, 30700138, 30830036, 30721002, 30871236), Chinese Academy of Sciences (KSCX2-YW-R-139, KSCX2-YW-R), the Cultivation Fund of the Ministry of Education of China (NO706035), and National Science Foundation for Post-doctoral Scientists (20080430100).

The authors have declared no conflict of interest.

References

[1] Geiss-Friedlander, R., Melchior, F., Concepts in sumoylation: a decade on. *Nat. Rev. Mol. Cell Biol.* 2007, *8*, 947–956.

[2] Gill, G., SUMO and ubiquitin in the nucleus: different functions, similar mechanisms? *Genes Dev.* 2004, *18*, 2046–2059.

[3] Seeler, J. S., Dejean, A., Nuclear and unclear functions of SUMO. *Nat. Rev. Mol. Cell Biol.* 2003, *4*, 690–699.

[4] Johnson, E. S., Protein modification by SUMO. *Annu. Rev. Biochem.* 2004, *73*, 355–382.

[5] Rodriguez, M. S., Dargemont, C., Hay, R. T., SUMO-1 conjugation in vivo requires both a consensus modification motif and nuclear targeting. *J. Biol. Chem.* 2001, *276*, 12654–12659.

[6] Dorval, V., Fraser, P. E., SUMO on the road to neurodegeneration. *Biochim. Biophys. Acta* 2007, *1773*, 694–706.

[7] Seeler, J. S., Bischof, O., Nacerddine, K., Dejean, A., SUMO, the three Rs and cancer. *Curr. Top. Microbial. Immunol.* 2007, *313*, 49–71.

[8] Xue, Y., Zhou, F., Fu, C., Xu, Y., Yao, X., SUMOsp: a web server for sumoylation site prediction. *Nucleic Acids Res.* 2006, *34*, W254–W257.

[9] Schwartz, D., Gygi, S. P., An iterative statistical approach to the identification of protein phosphorylation motifs from large-scale data sets. *Nat. Biotechnol.* 2005, *23*, 1391-1398.

[10] Xue, Y., Ren, J., Gao, X., Jin, C. *et al.*, GPS 2.0, a tool to predict kinase-specific phosphorylation sites in hierarchy. *Mol. Cell Proteomics* 2008, *7*, 1598-1608.

[11] Hannich, J. T., Lewis, A., Kroetz, M. B., Li, S. J. *et al.*, Defining the SUMO-modified proteome by multiple approaches in Saccharomyces cerevisiae. *J. Biol. Chem.* 2005, *280*, 4102-4110.

[12] Zhou, W., Ryan, J. J., Zhou, H., Global analyses of sumoylated proteins in Saccharomyces cerevisiae. Induction of protein sumoylation by cellular stresses. *J. Biol. Chem.* 2004, *279*, 32262–32268.

[13] Wohlschlegel, J. A., Johnson, E. S., Reed, S. I., Yates, J. R., III, Global analysis of protein sumoylation in Saccharomyces cerevisiae. *J. Biol. Chem.* 2004, *279*, 45662–45668.

[14] Wykoff, D. D., O'Shea, E. K., Identification of sumoylated proteins by systematic immunoprecipitation of the budding yeast proteome. *Mol. Cell Proteomics* 2005, *4*, 73–83.

[15] Panse, V. G., Hardeland, U., Werner, T., Kuster, B., Hurt, E., A proteome-wide approach identifies sumoylated substrate proteins in yeast. *J. Biol. Chem.* 2004, *279*, 41346–41351.

[16] Denison, C., Rudner, A. D., Gerber, S. A., Bakalarski, C. E. *et al.*, A Proteomic strategy for gaining insights into protein sumoylation in yeast. *Mol. Cell Proteomics* 2005, *4*, 246-254.

[17] Zhao, Y., Kwon, S. W., Anselmo, A., Kaur, K., White, M. A., Broad spectrum identification of cellular small ubiquitin-related modifier (SUMO) substrate proteins. *J. Biol. Chem.* 2004, *279*, 20999–21002.

[18] Manza, L. L., Codreanu, S. G., Stamer, S. L., Smith, D. L. *et al.*, Global shifts in protein sumoylation in response to electrophile and oxidative stress. *Chem. Res. Toxicol.* 2004, *17*, 1706–1715.

[19] Vertegaal, A. C., Ogg, S. C., Jaffray, E., Rodriguez, M. S. *et al.*, A proteomic study of SUMO-2 target proteins. *J. Biol. Chem.* 2004, *279*, 33791–33798.

[20] Li, T., Evdokimov, E., Shen, R. F., Chao, C. C. *et al.*, Sumoylation of heterogeneous nuclear ribonucleoproteins, zinc finger proteins, and nuclear pore complex proteins: a proteomic analysis. *Proc. Natl. Acad. Sci. USA* 2004, *101*, 8551–8556.

[21] Gocke, C. B., Yu, H., Kang, J., Systematic identification and analysis of mammalian small ubiquitin-like modifier substrates. *J. Biol. Chem.* 2005, *280*, 5004–5012.

[22] Rosas-Acosta, G., Russell, W. K., Deyrieux, A., Russell, D. H., Wilson, V. G., A Universal Strategy for Proteomic Studies of SUMO and Other Ubiquitin-like Modifiers. *Mol. Cell Proteomics* 2005, *4*, 56–72.

[23] Ganesan, A. K., Kho, Y., Kim, S. C., Chen, Y. *et al.*, Broad spectrum identification of SUMO substrates in melanoma cells. *Proteomics* 2007, *7*, 2216–2221.

