A genome-wide analysis of sumoylation-related biological processes and functions in human nucleus

Fengfeng Zhou[a,1], Yu Xue[b,1], Hualei Lu[b], Guoliang Chen[a], Xuebiao Yao[b,c,*]

[a] *National High-Performance Computing Center at Hefei, University of Science and Technology of China, Hefei 230027, China*
[b] *Laboratory of Cell Dynamics, University of Science and Technology of China, Hefei 230027, China*
[c] *Department of Physiology, Morehouse School of Medicine, Atlanta, GA 30310, USA*

Received 30 March 2005; accepted 26 April 2005

Available online 23 May 2005

Edited by Takashi Gojobori

Abstract Protein sumoylation is an important reversible post-translational modification of proteins in the nucleus, and it orchestrates a variety of the cellular processes. Genome-wide analysis of functional abundance and distribution of Small Ubiquitin-related MOdifier (SUMO) substrates may shed a light on how sumoylation is involved in nuclear biological processes and functions. Two interesting questions about sumoylation have emerged: (1) how many SUMO substrates exist in mammalian proteomes, such as human and mouse, (2) and what are their functions and how are they involved in a variety of biological processes? To address these two questions,we present an in silico genome-scale analysis for SUMO substrates in human. Based on the pattern recognition and phylogenetic conservation, we retrieved a list of 2683 potential SUMO substrates conserved in both human and mouse. Then, by functional enrichment analysis, we surveyed the over-represented GO terms and functional domains of them against the whole human proteome. Besides the consistence between our analyses and in vivo or in vitro work, the in silico predicted candidates also point to several potential roles of sumoylation, e.g., perception of sound. These potential SUMO substrates in human are of great value for further in vivo or in vitro experimental analysis.

Keywords: SUMO; Sumoylation; Transcription factor; Signal transduction; Perception of sound

1. Introduction

Small Ubiquitin-related MOdifier (SUMO) proteins are ubiquitously expressed in eukaryotic cells [1–4]. They are reversiblylinked to specific lysine residues of numerous substrates by sumoylation, and are implicated in various intracellular processes, such as nucleocytoplasmic signal transduction [5], transcription [6–8], stress response [9] and mitosis/cell-cycle progression [10,11], etc. SUMO proteins belong to the superfamily of ubiquitin-like modifiers (UBLs) [12], and consist of three components in mammalian cells: SUMO-1, SUMO-2, and SUMO-3 [13]. Only recently was another component SUMO-4 discovered in human [14]. SUMO proteins are highly conserved from yeast to human.

Conventional experimental approaches are employed to identify SUMO substrates with their sites in vivo or in vitro, although labor-intensive and time-consuming. Before millennium, there were only 12 experimentally verified SUMO substrates [4]. Recently, several genomic/proteomic-wide analyses of SUMO substrates have been deployed by mass spectrometry (MS) approaches in budding yeast [15–20]. Approximately, ~500 potential SUMO substrates in these large-scale experiments were found. These results are excellent candidates for further experimental consideration. Moreover, it is of great interest to identify novel SUMO substrates in mammals, especially in human given the recent completion of human genome project [21–25]. Due to the complexity of human proteome, only about two hundred candidates were found so far, and the exact sumoylation sites on most of these substrates remain elusive. In order to provide a more comprehensive view on sumoylation in human and on how they are involved in all kinds of intracellular biochemical processes, we developed a program SSP (SUMO substrates prediction) and conducted an in silico genome-wide analysis for nuclear SUMO substrates in human, based on pattern recognition and phylogenetic conservation approaches.

The majority of the SUMO substrates have a consensus motif with four amino acids. There are several motifs reported in the literatures: such as ψ-K-X-E (ψ is a hydrophobic amino acid) [2,4,23] and [VILMAFP]K.E (http://www.elm.eu.org/elmPages/MOD_SUMO.html) [26], etc. And a nuclear localization signal (NLS) suffices for SUMO conjugation in vivo [27], with only a few exceptions [28]. So we follow the ψ-K-X-E motif with a NLS as the consensus pattern for SUMO substrates prediction. In addition, the potential false positive hits are greatly reduced by phylogenetic conservation. For the prediction of sumoylation sites, SSP is nearly as sensitive as the existing tool SUMOplot (http://www.abgent.com/doc/sumoplot), with significantly improved specificity (see in Table 2).

We have generated a list of 2683 potential SUMO substrates conserved between human and mouse. We adopted the functional enrichment analysis to search for

[*] Corresponding author. Fax: +86 551 3607141.
E-mail address: yaoxb@ustc.edu.cn (X. Yao)

[1] The authors contributed equally to this work.

0014-5793/$30.00 © 2005 Federation of European Biochemical Societies. Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.febslet.2005.04.076

Table 1
The prediction results of known SUMO substrates

Protein name	Sumoylation sites	References	Prediction		
			SSP	SUMOplot	
				High	All
AP-2α	K10	12072434	Yes	Yes	Yes
AP-2β	K10	12072434	Yes	Yes	Yes
AP-2γ	K10	12072434	Yes	Yes	Yes
AR (androgen receptor)	K386, K520	11121022	Yes	Yes	Yes
ARNT (aryl hydrocarbon receptor nuclear transporter)	K245	12354770	Yes	Yes	Yes
Axin		12223491	Yes	Yes	Yes
Bach2		15060166	Yes	Yes	Yes
C/EBPβ-1	K173	12810706	Yes	Yes	Yes
C/EBPα (CCAAT/enhancer-binding protein alpha)	K159	12511558	Yes	Yes	Yes
c-Jun	K229	10788439	Yes	Yes	Yes
c-Myb	K503, K527	12631292	Yes	Yes	Yes
CREB (cAMP-response element-binding protein)	*K285, K304*	*12552083*	*No*[a]	*No*	*Yes*
CtBP1	*K428*	*12769861*	*No*[b]	*Yes*	*Yes*
Daxx	*K630, K631*	*12150977*	*No*[a]	*No*	*Yes*
Dnmt3a		*14752048*	*No*[b]	*Yes*	*Yes*
Dnmt3b		*14752048*	Yes	Yes	Yes
Dynamin-1		15123615	Yes	Yes	Yes
Dynamin-2		15123615	Yes	Yes	Yes
Dynamin-3		15123615	Yes	Yes	Yes
Elk-1	K230, K249	14992729	Yes	Yes	Yes
FAK (focal adhesion kinase)	K152	14500712	Yes	Yes	Yes
GATA-2		12750312	Yes	Yes	Yes
GLUT1		*10655495*	*No*[b]	Yes	Yes
GLUT4		*11842083*	*No*[b]	Yes	Yes
GR (glucocorticoid receptor)	K277, K293	12144530	Yes	Yes	Yes
GRIP1	K239, K731, and K788	12060666	Yes	Yes	Yes
HDAC1	*K444, K476*	*11960997*	*No*[b]	Yes	Yes
HDAC4	*K559*	*12032081*	*No*[c]	Yes	Yes
HIPK2	K1182	10535925	Yes	Yes	Yes
Histone H4		*14578449*	*No*[a]	*No*	Yes
hnRNP C	K237	15082759	Yes	Yes	Yes
hnRNP M		*15082759*	*No*[b]	Yes	Yes
HSF1 (heat shock transcription factor 1)	K298	11514557	Yes	Yes	Yes
HSF2 (heat shock transcription factor 2)	K82	11278381	Yes	Yes	Yes
IκBα	K21	9734360	Yes	Yes	Yes
IRF-1 (interferon regulatory factor-1)		*12387893*	*No*[a]	*No*	Yes
LEF1	K27, K269	11731474	Yes	Yes	Yes
Mdm2		11384992	Yes	Yes	Yes
MR (mineralocorticoid receptor)		14500761	Yes	Yes	Yes
NEMO/IKKγ	K277, K309	14651848	Yes	Yes	Yes
NFAT1		15117942	Yes	Yes	Yes
Nurr1 (NR4A2, RNR-1,TINUR, HZF-3)	K91, K577	14559918	Yes	Yes	Yes
p300/CBP	K1017, K1029	12718889	Yes	Yes	Yes
p53	*K386*	*10788439*	*No*[c]	Yes	Yes
p73α	K627	10961991	Yes	Yes	Yes
PC2		12679040	Yes	Yes	Yes
Pdx1 (pancreatic duodenal homeobox-1)		12488243	Yes	Yes	Yes
PIAS1		*12077349*	*No*[c]	Yes	Yes
PIASx-β		*12077349*	*No*[a]	Yes	Yes
PLZF (promyelocytic leukemia zinc finger)	*K242*	*14527952*	*No*[c]	Yes	Yes
PML (promyelocytic leukaemia protein)	K65, K160, and K490	10525530	Yes	Yes	Yes
PPAR-γ	*K107*	15123625;15229330	*No*[c]	Yes	Yes
PR (progesterone receptor)	K388	12529333	Yes	Yes	Yes
RanBP2/NUP358		15037602	Yes	Yes	Yes
RanGAP1	*K526*	*9442102*	*No*[b]	Yes	Yes
SALL1	K1086	12200128	Yes	Yes	Yes
SATB2		14701874	Yes	Yes	Yes
SENP1		14563852	Yes	Yes	Yes
Smad4	K113, K159	12621041	Yes	Yes	Yes
Sox6		A	Yes	Yes	Yes
Sox9		A	Yes	Yes	Yes
Sp100	*K297*	*10212234*	*No*[c]	Yes	Yes
Sp3	K539	12419227	Yes	Yes	Yes
SREBP-1a	K123, K418	12615929	Yes	Yes	Yes
SREBP-2	*K464*	*12615929*	*No*[b]	Yes	Yes
SRF (serum response factor)	K147	12788062	Yes	Yes	Yes
STAT1	*K703*	*12764129*	*No*[b]	Yes	Yes

Table 1 (*continued*)

Protein name	Sumoylation sites	References	Prediction		
			SSP	SUMOplot	
				High	All
Steroid receptor coactivator SRC-1/NCoA-1	K732, K774	12529333	Yes	Yes	Yes
Tcf-4	K297	12727872	Yes	Yes	Yes
TDG	K330	11889051	Yes	Yes	Yes
TEL	K99	12626745	Yes	Yes	Yes
TFII-I		15016812	Yes	Yes	Yes
TIF1α	K690, K708	11313457	Yes	Yes	Yes
TOPO I	K117, K153	12439742	Yes	Yes	Yes
Topoisomerase II α		14597774	Yes	Yes	Yes
Topoisomerase II β		12832072	Yes	Yes	Yes
Topors	K560	14516784	Yes	Yes	Yes
WRN		10806190	Yes	Yes	Yes
GATA4	K366	15337742	Yes	Yes	Yes
ZNF76	K411	15280358	Yes	Yes	Yes
PLAG1	K244, K263	15208321	Yes	Yes	Yes
Steroidogenic factor 1	K199, K194	15192092; 15192080	Yes	Yes	Yes
GATA1	K137	15173587	Yes	Yes	Yes
NFAT	K684, K897	15117942	Yes	Yes	Yes
Zinc finger protein APA-1		*12370286*	*No*[a]	*No*	Yes

85 experiment-verified SUMO substrates are listed. Our method can predict 64 of them correctly (~75%).

A. Fernandez-Lloris R. et al. (2002) Post-translational Sox6 protein modification by SUMO-1. In: 28th Meeting of the Federation of European Biochemical Societies, Istanbul, Turkey, pp. 20–25.

[a] No consensus motif (6 proteins).

[b] Not "nuc" (nuclear) hit by PSORT II prediction (9 proteins).

[c] Excluded by orthlogy information (6 proteins).

the over-represented GO terms and functional domains (Interpro) of the potential SUMO substrates against the whole human proteome. Our analyses of these potential substrates support the previous prediction of the functional relevance of sumoylation. For example, transcription factors and protein kinases are abundant in SUMO substrates, playing important roles in transcriptional regulation and gene expression [6–8] and signal transduction [1,2,29]. However, surprisingly, newly identified sumoylation candidates also point to several potential roles of sumoylation, e.g., perception of sound. Further analyses of these candidates in vivo or in vitro will provide insights into the function of sumoylation in mammalians, especially human.

2. Materials and methods

2.1. Identification of SUMO substrates with their sites in human and mouse

We took the orthology-relationship data of mouse and human with the corresponding sequences from the InParanoid database (Version 2.6, 30/03/2004) [30]. For the 34 499 mouse sequences and 36 379 human sequences in InParanoid, we firstly scanned the sequences for the consensus motif ψ-K-X-E in mouse and human, respectively. Sequences without such motif were excluded. Then we got 13 026 sequences in mouse and human, respectively. By PSORT II [31], we predicted the sub-cellular localization of the retained sequences. Only proteins with predicted nuclear localization were retained. After this step, there were 6662 sequences in mouse and 7649 in human, respectively.

In order to eliminate the potential false positive results, we followed a simple rule below: for the pairwise orthologs between the retained mouse and human proteins, there must be at least one consensus SUMO substrate motif at the same position after sequence alignment. Thus, proteins without such orthologs were excluded. After the sequence alignment, orthologs sharing no consensus motif at the same position were also excluded, resulting a final 2683 orthologous proteins in both mouse and human proteomes.

2.2. Statistical analysis for SUMO substrates

We downloaded the GO (08/10/2004) and Interpro (23/06/2004) [32] association files from EBI (ftp://ftp.ebi.ac.uk/pub/) and searched for the GO and Interpro annotations of human proteins. Among 36 379 human proteins of InParanoid, there are 24 090 and 26 873 annotated with at least one GO and Interpro term, respectively, and there are 1956 and 2264 proteins of our 2683 potential SUMO substrates annotated, separately. Following a statistical approach described before [33], we compared the group S (predicted SUMO substrates of human) against the group W (whole human proteome) to find a GO/Interpro term *t* that occurred more frequently in group S than in group W. Here we define:

N	total number of proteins in group W annotated by GO/Interpro
n	number of proteins in group W annotated by GO/Interpro term t
M	total number of proteins in group S annotated by GO/Interpro
m	number of proteins in group S annotated by GO/Interpro term t

Then we calculate the enrichment ratio of GO/Interpro term *t* in group S, and with the equation of the hypergeometric distribution, we can also calculate its *P*-value:

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

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

or

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

In this work, we only consider the over-representation of GO/Interpro groups with Enrichment_ratio $\geqslant$ 1.

3. Results

3.1. Accuracy of SSP1.0 program

It is reported that evolutionary stable sites can be used to improve the prediction specificity for functional sites/motifs [34], based on the hypothesis that functional sites/motifs should be more conserved than random pseudo-sites/motifs. For our phylogenetic conservation analysis, we chose distance near specie mouse for human rather than other too distant species such as budding yeast or fly, to avoid missing too many real sumoylation sites. Too near species such as primates are not used, because these proteomes are too similar with human and cannot reduce the potential false positives much. So we adopted the phylogenetic conservation between human and mouse to reduce the potential false positives. Curated from the published work, we got 85 experimental verified SUMO substrates (see Table 1). The SSP 1.0 can predict 64 (75%) of them correctly.

For precise sumoylation site prediction, we compare our computational results with the existing tool SUMOplot (see Table 2). For the 63 known sumoylation sites, our method could recover 51 of them with sensitivity $S_{\mathrm{n}} \sim 81\%$, which is similar to the SUMOplot results ~81% (motifs with high probability) or ~84% (all). Yet the specificity S_{p} of our approach is significantly improved to ~60% (51 in a total of 86), compared to SUMOplot ~31% (motifs with high probability) or ~15%

Table 2
The comparison of the sumoylation site prediction against SUMOplot

Protein name	Sumoylation sites		
	Verified	SSP 1.0	SUMOplot
AP-2α	K10	**IKYE**[a]	1/1[b]; (1/4)[c]
AP-2β	K10	**IKYE**	1/1; (1/5)
AP-2γ	K10	**IKYE**	1/1; (1/4)
AR (androgen receptor)	K386, K520	**IKLE**, **VKSE**	2/3; (2/6)
ARNT (aryl hydrocarbon receptor nuclear transporter)	K245	**VKKE**	0/2; (0/5)
C/EBPβ-1	K173	**LKAE**	1/2; (1/4)
C/EBPα (CCAAT/enhancer-binding protein alpha)	K159	**LKAE**	1/1; (1/1)
c-Jun	K229	AKME, **LKEE**, IKAE	1/3; (1/4)
c-Myb	K503, K527	**IKQE**, **IKQE**	2/5; (2/10)
Elk-1	K230, K249	**VKVE**	2/2; (2/4)
FAK (focal adhesion kinase)	K152	WKYE	1/5; (1/17)
GR (glucocorticoid receptor)	K277, K293	**VKTE**, **IKQE**, VKRE	0/1; (0/5)
GRIP1	K239, K731, K788	**VKLE**, **MKQE**	2/7; (2/17)
HIPK2	K1182	LKIE, LKPE	0/3; (0/7)
hnRNP C	K237	IKKE,**VKME**	0/4; (1/12)
HSF1 (heat shock transcription factor 1)	K298	VKPE, LKSE, MKHE, **VKEE**	1/6; (1/9)
HSF2 (heat shock transcription factor 2)	K82	**VKQE**, IKQE, LKSE	1/6; (1/9)
IκBα	K21	**LKKE**, MKDE	1/4; (1/4)
LEF1	K27, K269	**FKDE**, **VKQE**	2/3; (2/7)
NEMO/IKKγ	K277, K309	**AKQE**, LKEE	1/8; (1/13)
Nurr1 (NR4A2, RNR-1,TINUR, HZF-3)	K91, K577	**IKVE**, **LKLE**	2/4; (2/10)
p300/CBP	K1017, K1029	MKTE, VKEE, VKVE, VKEE, FKPE	2/11; (2/22)
p73α	K627	**IKEE**	1/3; (1/7)
PML (promyelocytic leukaemia protein)	K65, K160, K490	**LKHE**, **IKME**	3/6; (3/7)
PR (progesterone receptor)	K388	**IKEE**	1/3; (1/6)
SALL1	K1086	IKTE, **IKTE**	1/7; (1/15)
Smad4	K113, K159	**VKDE**	1/3; (1/7)
Sp3	K539	IKDE, **IKEE**	1/3; (1/5)
SREBP-1a	K123, K418	**IKEE**, LKQE, **VKTE**	2/7; (2/11)
SRF (serum response factor)	K147	**IKME**	1/1; (1/3)
Steroid receptor coactivator SRC-1/ NCoA-1	K732, K774	AKAE, **IKLE**, **VKVE**, IKLE, IKSE	1/7; (1/13)
Tcf-4	K297	FKDE, **VKQE**	1/6; (1/10)
TDG	K330	**VKEE**	0/0; (0/8)
TEL	K99	IKQE	0/2; (0/3)
TIF1α	K690, K708	**IKQE**, **VKQE**, IKLE	2/5; (2/11)
TOPO I	K117, K153	IKKE, **IKTE**, IKEE, FKIE, IKGE, MKLE	2/12; (2/29)
Topors	K560	LKRE	0/4; (1/10)
GATA4	K366	**IKTE**	1/1; (1/2)
ZNF67	K411	VKGE, **VKEE**	1/2; (1/5)
PLAG1	K244, K263	FKCE,**VKTE**, **IKDE**, LKGE	2/5; (2/11)
Steroidogenic factor 1	K199, K194	**FKLE**, **IKSE**	2/2; (2;3)
GATA1	K137	**LKTE**	1/1; (1/4)
NFAT	K684, K897	**IKTE**, **IKQE**	2/3; (2/6)
Total sites	63	51/86[d]	51/166; (53/355)

63 verified sumoylation sites of 43 known SUMO substrates are chosen.

[a] SSP 1.0 hits are in bold character font.

[b] SUMOplot hits (motifs with high probability)/total predicted sites (motifs with high probability).

[c] SUMOplot hits (all)/total predicted sites (all).

[d] SSP 1.0 hits/total predicted sites.

(all). So our method greatly reduces the number of potential false positives while still keeps a satisfying sensitivity.

3.2. Functional abundance and distribution of SUMO substrates

SUMO substrates are implicated in many intracellular processes. However, the systems biology of sumoylation remains unclear. Thus, it is of great interest to illustrate in which functions the nuclear SUMO substrates of human are significantly abundant. Here we perform a statistical analysis to predict such significance. Our analytical outcomes (see Table 3) are consistent with several widely held, but yet to be systematically examined assumptions on the sumoylation involved processes.

For example, sumoylation was proposed to play a role in transcriptional regulation and gene expression [6–8], where many SUMO substrates are transcription factors [1,2]. Are there any strong correlations between transcriptional regulation and sumoylation? From our analysis, we found that the transcription factor and transcriptional regulation are both among the top of the list of significantly enriched functions or processes (Table 3). In the human proteome, 2255, and 1102 proteins are annotated with functions of DNA binding (GO:0003677) and transcription factor activity (GO:0003700), respectively. And there are also 2174 proteins annotated with process of regulation of transcription, DNA-dependent (GO:0006355). In our data set, there are 530, 304, and 510 proteins with the above three annotations, respectively. So it could be estimated that about 1/4–1/3 of the transcription factors are downregulated by sumoylation. Interestingly, we found that the processes of transcription from Pol II promoter (GO:0006366) and regulation of transcription from Pol II promoter (GO:0006357) are highly correlated with SUMO substrates. This supports the hypothesis that sumoylation may play a role at the promoter by modifying transcription factors as chromatin-bound complexes, but not by regulating transcription directly [1,2], and the functions of transcription coactivator activity (GO:0003713) (see Table 3) and transcription corepressor activity (GO:0003714) ($P < 10^{-7}$) are also significantly represented. This finding is consistent with the recent observations that sumoylation can repress or activate transcription [1,2].

Several SUMO substrates were summarized to be essential in signal transduction [1,2,29]. In Drosophila brain, the functional dynamics of neuronal calcium/calmodulin-dependent protein kinase II was regulated by sumoylation, which is important for the differentiated nervous system [35]. In NF-κB signaling pathways, the regulatory subunit of the IκB kinase (IKK) complex NEMO/IKKγ will be sumoylated to release NF-κB from its inhibitor IκBα, inducing a survival response against genotoxic stress [5,9]. In our data set, the processes

Table 3
The top 15 most enriched processes and functions in SUMO substrates

Description of GO term	Number of proteins annotated in group S[a]	Number of proteins annotated in group W[b]	Enrichment ratio	*P*-value
The top 15 most enriched processes in SUMO substrates				
Regulation of transcription, DNA-dependent (GO:0006355)	26.1% (510)	9.0% (2174)	2.89	6.12 E − 121
Transcription from Pol II promoter (GO:0006366)	3.5% (69)	0.8% (204)	4.17	1.00 E − 25
Development (GO:0007275)	5.8% (114)	2.6% (631)	2.23	2.96 E − 16
Signal transduction (GO:0007165)	9.1% (178)	5.0% (1207)	1.82	2.06 E − 15
Regulation of transcription from Pol II promoter (GO:0006357)	2.7% (52)	0.8% (192)	3.34	4.13 E − 15
Protein amino acid phosphorylation (GO:0006468)	6.7% (131)	3.5% (850)	1.90	5.45 E − 13
Cell growth and/or maintenance (GO:0008151)	3.4% (67)	1.4% (341)	2.42	9.45 E − 12
Cell cycle (GO:0007049)	2.5% (49)	1.0% (240)	2.51	1.49 E − 09
Intracellular signaling cascade (GO:0007242)	4.6% (90)	2.5% (609)	1.82	2.00 E − 08
Endocytosis (GO:0006897)	1.4% (27)	0.4% (108)	3.08	9.71 E − 08
Mitosis (GO:0007067)	1.3% (26)	0.4% (103)	3.11	1.35 E − 07
Perception of sound (GO:0007605)	1.2% (23)	0.4% (87)	3.26	2.87 E − 07
Morphogenesis (GO:0009653)	1.2% (23)	0.4% (107)	2.65	1.31 E − 05
Frizzled signaling pathway (GO:0007222)	0.5% (10)	0.1% (26)	4.74	1.92 E − 05
Negative regulation of transcription from Pol II promoter (GO:0000122)	0.9% (18)	0.3% (74)	3.00	1.93 E − 05
The top 15 most enriched functions in SUMO substrates				
DNA binding (GO:0003677)	27.1% (530)	9.4% (2255)	2.89	1.00 E − 126
Transcription factor activity (GO:0003700)	15.5% (304)	4.6% (1102)	3.40	3.64 E − 87
Nucleic acid binding (GO:0003676)	14.2% (277)	7.6% (1823)	1.87	7.89 E − 26
Zinc ion binding (GO:0008270)	14.6% (285)	8.2% (1968)	1.78	2.80 E − 23
Protein serine/threonine kinase activity (GO:0004674)	6.1% (119)	2.3% (559)	2.62	7.18 E − 23
Actin binding (GO:0003779)	3.7% (72)	1.1% (259)	3.42	4.25 E − 21
ATP binding (GO:0005524)	13.3% (260)	8.0% (1925)	1.66	3.69 E − 17
Protein kinase activity (GO:0004672)	6.5% (128)	3.2% (776)	2.03	6.38 E − 15
RNA polymerase II transcription factor activity (GO:0003702)	2.1% (41)	0.6% (138)	3.66	1.12 E − 13
Steroid hormone receptor activity (GO:0003707)	1.5% (29)	0.3% (75)	4.76	2.47 E − 13
GTPase activator activity (GO:0005096)	1.8% (35)	0.5% (110)	3.92	7.49 E − 13
Transcription coactivator activity (GO:0003713)	2.2% (43)	0.7% (158)	3.35	8.04 E − 13
Ligand-dependent nuclear receptor activity (GO:0004879)	1.5% (29)	0.3% (79)	4.52	1.17 E − 12
Protein binding (GO:0005515)	11.9% (233)	8.0% (1907)	1.50	7.58 E − 11
Calmodulin binding (GO:0005516)	1.8% (35)	0.5% (132)	3.27	2.31 E − 10

We list the top 15 of the over-represented functions and processes for further discussion.
[a] Group S, the SUMO substrates.
[b] Group W, whole human proteome.

Table 4
The top 10 most over-represented protein domains in SUMO substrates

Description of Interpro term	Number of proteins annotated in group S[a]	Number of proteins annotated in group W[b]	Enrichment ratio	P-value
Serine/threonine protein kinase, active site (IPR008271)	5.1% (115)	1.8% (486)	2.81	8.69 E − 25
Pleckstrin-homology-related (IPR011036)	5.0% (113)	2.0% (538)	2.49	5.71 E − 20
Pleckstrin-like (IPR001849)	4.2% (94)	1.6% (421)	2.65	1.16 E − 18
Serine/threonine protein kinase (IPR002290)	3.4% (78)	1.2% (328)	2.82	2.34 E − 17
Zn-finger, C2H2 type (IPR007087)	8.2% (185)	4.4% (1177)	1.87	4.24 E − 17
Zn-finger-like, PHD finger (IPR001965)	2.0% (46)	0.5% (139)	3.93	1.46 E − 16
Homeodomain-like (IPR009057)	3.6% (81)	1.3% (362)	2.66	2.53 E − 16
Protein kinase (IPR000719)	5.8% (132)	3.0% (796)	1.97	2.96 E − 14
Protein kinase-like (IPR011009)	5.8% (131)	2.9% (790)	1.97	3.73 E − 14
Winged helix DNA-binding (IPR009058)	2.6% (59)	0.9% (244)	2.87	8.25 E − 14

We list the top 10 of the over-represented protein domains for further discussion.
[a]Group S, the SUMO substrates.
[b]Group W, whole human proteome.

of signal transduction (GO:0007165) and intracellular signaling cascade (GO:0007242) are much enriched ($P < 10^{-7}$), which implies that sumoylation may be involved in signal transduction extensively. We also find that the GO groups of protein serine/threonine kinase activity (GO:0004674), and protein kinase activity (GO:0004672) are significantly over-represented ($P < 10^{-14}$). In the human proteome, there are 559 and 776 proteins annotated with the two GO terms respectively, while there are 119 and 128 of them are among our data set. Thus, it could be estimated that ~1/5 serine/threonine kinases could be sumoylated. This result is in accordance with the hypothesis that crosstalk between sumoylation and phosphorylation may be fundamental and essential in signal transduction [8].

Another interesting cellular process identified to be highly relevant to sumoylation is the process of perception of sound (GO:0007605) ($P < 10^{-6}$). Thus, we propose that sumoylation may play an important role in the perception of sound pathways, a novel finding that was never reported.

3.3. Significantly represented protein domains in the data set

To provide further insight into the functional enrichment of SUMO substrates, we also perform the statistical analysis to obtain additional evidence of what types of protein domains are more frequently encoded in them. Since sumoylation may be mainly implicated in transcription regulation and signal transduction by sumoylating transcription factors and protein serine/threonine kinases, respectively, it could be anticipated that some specific protein domains, such as DNA binding or kinase, should be abundant in our data set.

The analysis on the InterPro annotations [32] satisfyingly confirms with the above results. The top 10 most enriched protein domains in SUMO substrates are listed (Table 4). It is not surprising that the protein domains such as Serine/threonine protein kinase, active site (IPR008271), Serine/threonine protein kinase (IPR002290), Zn-finger, C2H2 type (IPR007087), and Zn-finger-like, PHD finger (IPR001965) are significantly abundant in the data set ($P < 10^{-15}$). Unexpectedly, we notice that Pleckstrin-homology-related (IPR011036) and Pleckstrin-like (IPR001849) are also in our top list. Pleckstrin homology (PH) domains are small modular domains with ~100 amino-acid residues that occur once, or occasionally several times, in a large variety of proteins involved in intracellular signaling or as constituents of the cytoskeleton [36]. This observation may propose that there are some specific protein domains abundant in SUMO substrates to form links between sumoylation and signaling related pathways. Although most of the protein domains are focused on all kinds of DNA-binding domain, Zn-finger, C2H2 type (IPR007087) domains can bind both DNA and RNA. And we found that the domain of RNA-binding region RNP-1 (RNA recognition motif) (IPR000504) is significant ($P < 10^{-7}$). So our analysis also supports the hypothesis that sumoylation may play a role in RNA metabolism [37].

4. Discussion

In this paper, we provide a genome-scale analysis of sumoylation-related biological processes and functions. The results show that sumoylation may be strongly correlated with the transcription regulation and signal transduction, which is consistent with the experimental observations. Our analysis also provides several other interesting hints, e.g., sumoylation may be involved in the perception of sound, offering insights for further experimental manipulation. Taken together, our data set establishes a good resource for potential SUMO substrates with high specificity.

5. Supplementary materials

Supplementary materials and the software SSP (SUMO Substrates Prediction) implemented in Delphi are available from: http://973-proteinweb.ustc.edu.cn/sumo/.

Acknowledgments: We thank Yi Xing (UCLA) for helpful discussions and Andrew Shaw for critical reading of this manuscript. This work is supported by grants from Chinese Natural Science Foundation (39925018 and 30121001), Chinese Academy of Science (KSCX2-2-01), Chinese 973 project (2002CB713700), Beijing Office for Science (H020220020220) and American Cancer Society (RPG-99-173-01) to X. Yao. X. Yao is a GCC Distinguished Cancer Research Scholar.

References

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

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

[3] Melchior, F., Schergaut, M. and Pichler, A. (2003) SUMO: ligases, isopeptidases and nuclear pores. Trends Biochem. Sci. 28, 612–618.

[4] Melchior, F. (2000) SUMO – nonclassical ubiquitin. Annu. Rev. Cell Dev. Biol. 16, 591–626.

[5] Hay, R.T., Vuillard, L., Desterro, J.M. and Rodriguez, M.S. (1999) Control of NF-kappa B transcriptional activation by signal induced proteolysis of I kappa B alpha. Philos. Trans. R. Soc. Lond. B: Biol. Sci. 354, 1601–1609.

[6] Verger, A., Perdomo, J. and Crossley, M. (2003) Modification with SUMO. A role in transcriptional regulation. EMBO Rep. 4, 137–142.

[7] Schmidt, D. and Muller, S. (2003) PIAS/SUMO: new partners in transcriptional regulation. Cell. Mol. Life Sci. 60, 2561–2574.

[8] Gill, G. (2003) Post-translational modification by the small ubiquitin-related modifier SUMO has big effects on transcription factor activity. Curr. Opin. Genet. Dev. 13, 108–113.

[9] Huang, T.T., Wuerzberger-Davis, S.M., Wu, Z.H. and Miyamoto, S. (2003) Sequential modification of NEMO/IKKgamma by SUMO-1 and ubiquitin mediates NF-kappaB activation by genotoxic stress. Cell 115, 565–576.

[10] Pinsky, B.A. and Biggins, S. (2002) Top-SUMO wrestles centromeric cohesion. Dev. Cell 3, 4–6.

[11] Muller, S., Hoege, C., Pyrowolakis, G. and Jentsch, S. (2001) SUMO, ubiquitin's mysterious cousin. Nat. Rev. Mol. Cell. Biol. 2, 202–210.

[12] Schwartz, D.C. and Hochstrasser, M. (2003) A superfamily of protein tags: ubiquitin, SUMO and related modifiers. Trends Biochem. Sci. 28, 321–328.

[13] Saitoh, H. and Hinchey, J. (2000) Functional heterogeneity of small ubiquitin-related protein modifiers SUMO-1 versus SUMO-2/3. J. Biol. Chem. 275, 6252–6258.

[14] Bohren, K.M., Nadkarni, V., Song, J.H., Gabbay, K.H. and Owerbach, D. (2004) A M55V polymorphism in a novel SUMO gene (SUMO-4) differentially activates heat shock transcription factors and is associated with susceptibility to type I diabetes mellitus. J. Biol. Chem. 279, 27233–27238.

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

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

[17] Hannich, J.T., Lewis, A., Kroetz, M.B., Li, S.J., Heide, H., Emili, A. and Hochstrasser, M. (2005) Defining the SUMO-modified proteome by multiple approaches in *Saccharomyces cerevisiae*. J. Biol. Chem. 280, 4102–4110.

[18] 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.

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

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

[21] Rosas-Acosta, G., Russell, W.K., Deyrieux, A., Russell, D.H. and Wilson, V.G. (2005) A universal strategy for proteomic studies of SUMO and other ubiquitin-like modifiers. Mol. Cell. Proteomics 4, 56–72.

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

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

[24] Vertegaal, A.C., Ogg, S.C., Jaffray, E., Rodriguez, M.S., Hay, R.T., Andersen, J.S., Mann, M. and Lamond, A.I. (2004) A proteomic study of SUMO-2 target proteins. J. Biol. Chem. 279, 33791–33798.

[25] Manza, L.L., Codreanu, S.G., Stamer, S.L., Smith, D.L., Wells, K.S., Roberts, R.L. and Liebler, D.C. (2004) Global shifts in protein sumoylation in response to electrophile and oxidative stress. Chem. Res. Toxicol. 17, 1706–1715.

[26] Puntervoll, P., Linding, R., Gemund, C., Chabanis-Davidson, S., Mattingsdal, M., Cameron, S., Martin, D.M., Ausiello, G., Brannetti, B. and Costantini, A., et al. (2003) ELM server: a new resource for investigating short functional sites in modular eukaryotic proteins. Nucleic Acids Res. 31, 3625–3630.

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

[28] Watts, F.Z. (2004) SUMO modification of proteins other than transcription factors. Semin. Cell Dev. Biol. 15, 211–220.

[29] Johnson, E.S. (2004) Protein modification by SUMO. Annu. Rev. Biochem. 73, 355–382.

[30] Remm, M., Storm, C.E. and Sonnhammer, E.L. (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. J. Mol. Biol. 314, 1041–1052.

[31] Nakai, K. and Horton, P. (1999) PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. Trends Biochem. Sci. 24, 34–36.

[32] Mulder, N.J., Apweiler, R., Attwood, T.K., Bairoch, A., Barrell, D., Bateman, A., Binns, D., Biswas, M., Bradley, P. and Bork, P., et al. (2003) The InterPro Database, 2003 brings increased coverage and new features. Nucleic Acids Res. 31, 315–318.

[33] Xing, Y., Xu, Q. and Lee, C. (2003) Widespread production of novel soluble protein isoforms by alternative splicing removal of transmembrane anchoring domains. FEBS Lett. 555, 572–578.

[34] Blom, N., Sicheritz-Ponten, T., Gupta, R., Gammeltoft, S. and Brunak, S. (2004) Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. Proteomics 4, 1633–1649.

[35] Long, X. and Griffith, L.C. (2000) Identification and characterization of a SUMO-1 conjugation system that modifies neuronal calcium/calmodulin-dependent protein kinase II in *Drosophila melanogaster*. J. Biol. Chem. 275, 40765–40776.

[36] Rebecchi, M.J. and Scarlata, S. (1998) Pleckstrin homology domains: a common fold with diverse functions. Annu. Rev. Biophys. Biomol. Struct. 27, 503–528.

[37] Li, T., Evdokimov, E., Shen, R.F., Chao, C.C., Tekle, E., Wang, T., Stadtman, E.R., Yang, D.C. and Chock, P.B. (2004) Sumoylation of heterogeneous nuclear ribonucleoproteins, zinc finger proteins, and nuclear pore complex proteins: a proteomic analysis. Proc. Natl. Acad. Sci. USA 101, 8551–8556.