Bioinformatics, 31(21), 2015, 3483–3491
doi: 10.1093/bioinformatics/btv403
Advance Access Publication Date: 2 July 2015
Original Paper

Sequence analysis

JASSA: a comprehensive tool for prediction of SUMOylation sites and SIMs

Guillaume Beauclair[1,2,3,*,‡], **Antoine Bridier-Nahmias**[1,2,3,4], **Jean-François Zagury**[5], **Ali Saïb**[1,2,3,4,†] **and Alessia Zamborlini**[1,2,3,4,†]

[1]CNRS UMR7212, Hôpital St Louis, [2]Inserm U944, Institut Universitaire d'Hématologie, Hôpital St Louis, [3]Université Paris Diderot, Sorbonne Paris Cité, Hôpital St Louis, [4]Laboratoire PVM, Conservatoire national des arts et métiers (Cnam) and [5]Laboratoire Génomique, Bioinformatique, et Applications, EA4627, Chaire de bioinformatique, Conservatoire national des arts et métiers (Cnam), Paris, France

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the last two authors should be regarded as Joint Last Authors.
‡Present address: Institute of Virology, Hannover Medical School, Carl-Neuberg-Strasse 1, 30625, Hannover, Germany.

Associate Editor: John Hancock

Received on April 14, 2015; revised on June 5, 2015; accepted on June 25, 2015

Abstract

Motivation: Post-translational modification by the Small Ubiquitin-like Modifier (SUMO) proteins, a process termed SUMOylation, is involved in many fundamental cellular processes. SUMO proteins are conjugated to a protein substrate, creating an interface for the recruitment of cofactors harboring SUMO-interacting motifs (SIMs). Mapping both SUMO-conjugation sites and SIMs is required to study the functional consequence of SUMOylation. To define the best candidate sites for experimental validation we designed JASSA, a Joint Analyzer of SUMOylation site and SIMs.

Results: JASSA is a predictor that uses a scoring system based on a Position Frequency Matrix derived from the alignment of experimental SUMOylation sites or SIMs. Compared with existing web-tools, JASSA displays on par or better performances. Novel features were implemented towards a better evaluation of the prediction, including identification of database hits matching the query sequence and representation of candidate sites within the secondary structural elements and/or the 3D fold of the protein of interest, retrievable from deposited PDB files.

Availability and Implementation: JASSA is freely accessible at http://www.jassa.fr/. Website is implemented in PHP and MySQL, with all major browsers supported.

Contact: guillaume.beauclair@inserm.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

SUMOylation is a eukaryotic post-translational modification which consists in the reversible attachment of members of the Small Ubiquitin-like Modifier (SUMO) protein family on a protein substrate resulting in the dynamic regulation of its biochemical properties. Proteins involved in many fundamental cellular processes like DNA repair, transcription control, chromatin organization, macromolecular assembly and signal transduction are SUMOylated [for a review see (Flotho and Melchior, 2013)]. Thus, it is not surprising that deregulation of SUMOylation is associated to various pathological conditions like neurological disorders, cancers and pathogen proliferation [for reviews (Droescher *et al.*, 2013; Sarge and Park-Sarge, 2009; Wilson, 2012; Wimmer *et al.*, 2012)].

SUMO proteins are conjugated to a lysine (K) residue of the substrate through the sequential action of SUMO-specific activating (E1), conjugating (E2) and ligating (E3) enzymes and de-conjugation relies on SUMO-specific proteases. Virtually all eukaryotes express SUMO proteins, mammals and plants harboring several paralogues, and the components of the conjugation pathway are highly conserved across eukaryote proteomes (Flotho and Melchior, 2013; Gareau and Lima, 2011).

Analysis of SUMO targets shows that the modified K is often embedded within the consensus sequence ΨKxE (where Ψ is a hydrophobic residue and x any amino acid), which is a binding site for Ubc9, the single E2 conjugating enzyme (Sampson *et al.*, 2001). Extended variants of this motif were described, including negatively charged amino acid-dependent SUMOylation motifs (NDSM) (Yang *et al.*, 2006), phosphorylation-dependent SUMOylation motifs (PDSM) (Hietakangas *et al.*, 2006) and phosphorylation SUMOylation motif (Picard *et al.*, 2012), where a cluster of negatively charged and/or phosphorylatable residues downstream of the core motif promotes SUMO conjugation by strengthening the interaction between the substrate and Ubc9 (Mohideen *et al.*, 2009). SUMOylation of the inverted consensus motif ([E/D]xKΨ) was also reported (Ivanov *et al.*, 2007; Matic *et al.*, 2010). However, not every sequence conforming to the consensus motifs is modified, likely because the environment of the target K must adopt a favorable conformation to be accessible to the SUMO machinery (Pichler *et al.*, 2005). Notably, modification by SUMO occurs at non-consensus sites. This is the case for ~50% of the SUMOylated substrates identified by the Vertegaal's group (Hendriks *et al.*, 2014).

SUMO creates an interface for the recruitment of protein cofactors that harbor short peptide sequences known as SUMO-interacting motifs (SIMs). Most SIMs feature a loose consensus sequence composed of 3–4 aliphatic residues often flanked by acidic and/or phosphorylatable amino acids (Kerscher, 2007). The hydrophobic core adopts a β-strand conformation that can accommodate in a pocket formed by the α_1-helix and the β_2-strand of SUMO (Hecker *et al.*, 2006; Song *et al.*, 2005). Adjacent negatively charged residues control the affinity, the polarity and the paralogue-specificity of the SIM/SUMO interaction (Chang *et al.*, 2011; Hecker *et al.*, 2006; Meulmeester *et al.*, 2008).

Mapping SUMO-conjugation sites and SIMs is mandatory to fully characterize the biological consequences of SUMOylation. Large-scale mass spectrometry (MS)-based proteomic studies allowed the identification of hundreds of proteins harboring SUMOylation sites and/or SIMs (Blomster *et al.*, 2009; Hendriks *et al.*, 2014; Impens *et al.*, 2014; Matic *et al.*, 2010; Tammsalu *et al.*, 2014; Tatham *et al.*, 2011). However, the use of this approach is limited by the transient nature of the modification, the small fraction of a protein that is SUMOylated and the difficulty to identify branched peptides resulting for the tryptic digestion of SUMOylated proteins. When MS data are not available, computer-aided prediction of SUMOylation sites and SIMs by *in silico* analysis represents a promising strategy to reduce the number of potential targets for experimental verification.

We developed JASSA (Joint analyzer of SUMOylation site and SIMs) to provide a comprehensive overview of potential SUMOylation sites and SIMs. The prediction relies on a scoring system based on a Position Frequency Matrix (PFM) derived from the alignment of experimentally validated sequences. When compared with existing bioinformatics tools, JASSA displays on par or better predictive performances. To increase the reliability of the prediction, JASSA offers additional features such as identification of database (DB) hits matching the query sequence, systematic pattern search against extended motifs, analysis of the physico-chemical properties of adjacent residues and, when a PDB file is available, the possibility to represent candidate sites within the secondary structural elements and the 3D fold of the query protein. We believe that JASSA will provide a valuable support for selecting the best candidate sites for experimental studies.

2 System and methods

2.1 Databases of SUMOylation sites and SIMs

The training DB of SUMOylation sites was generated by collecting data from PhosphoSite (Hornbeck *et al.*, 2004), Teng *et al.* (Teng *et al.*, 2012) and NCBI (publication until January 2011) using 'SUMO' and 'SUMOylation' as keywords. It encompasses 877 unique SUMOylated experimentally defined K residues from 505 proteins. The sequence of the 21-mer centered on the modified K was retrieved from UniProt (Apweiler *et al.*, 2004). JASSA operates using either of three clusters of experimental SUMOylation sites: ALL, DIRECT or INVERTED, which include all the sequences of the DB or the sequences following the direct ΨKxα or the inverted αxKΨ consensus, respectively (where Ψ is an hydrophobic residue; α is E or D and x is any amino acid) (see below).

The collection of SIMs consists of 102 non-redundant motifs from 66 proteins obtained from NCBI using '(SIM or SBD or SBM) and SUMO' as keywords (publication until January 2014). Putative SIMs from proteins interacting with SUMO in yeast two-hybrid screens, but not validated further, were not included.

2.2 Characterization of the SUMOylation sites DB and motifs clustering strategy

Most of the SUMOylation sites retrieved from the scientific literature and included in our training DB are from *Homo sapiens* (71.9%) and rodents (14.7%). The remaining motifs are from yeast, viruses and plants (5.6, 3.1 and 2.5%, respectively) (Supplementary Fig. S1A). A pattern search against known SUMOylation consensus motifs showed that 598 sites (68.2%) of the DB fit with the direct ΨKxα motif (where Ψ = A, F, G, I, L, M, P, V or Y; α = D or E; x = any amino acid). Among these 26.3% are NSDM, 3.6% PSDM, 12% HSCM and 20% synergy control motif (Table 1). We also found that 80 sites of the DB (9.1%) follow the inverted αxKΨ consensus, while 30 (3.4%) fit both the direct and the inverted consensus motif. The remaining 229 sites (26.1%) do not conform to any of these motifs and are considered as non-consensus. Based on these observations we grouped the experimental SUMOylation sites in three clusters which encompass all the sequences of the DB (ALL), the sites following the direct ΨKxα (DIRECT) or the inverted αxKΨ consensus (INVERTED). We reasoned that the DB of JASSA being enriched in sequences that match the direct ΨKxα motif (>68%), while inverted motifs are underrepresented (<10%), could undermine the prediction of sites that fit the αxKΨ pattern.

Next, we analyzed the local target protein context in a 21-mer window centered on the SUMOylated K (position 0) for each cluster with WebLogo3 (Crooks *et al.*, 2004) (Fig. 1). The sequence logo for the cluster ALL agrees with the prevalence of the ΨKxα motif in the DB. At position −1 there is a higher occurrence of hydrophobic amino acids most of which are residues with aliphatic side chains (67.5% among which 27.7% are I, 12.7% are L and 23.7% are V), whereas aromatic amino acids are rare (5.4% of which 4.6% are F) (Supplementary Table S1A). At position +2 acidic residues are significantly enriched (70% E, 5.4% D). Finally, no particular amino acid is overrepresented at position +1. As expected, the sequence logo of the cluster DIRECT is similar except that position −1 is occupied exclusively by hydrophobic residues, whereas E (~94%) and D (~6%) are the only amino acids found at position +2. Consistent with the fact that about a third of the SUMOylation sites of the collection match with the NDSM or the PDSM motifs, acidic and/or phosphorylatable residues are enriched downstream of the core consensus motif (Fig. 1 and Table 1). The amino acids distribution around the SUMO-acceptor K is different for the cluster

Table 1. List and proportion of known SUMOylation sites in the DB of JASSA

	Name	Motif	Nb	%	References
Consensus direct	Stong consensus	[Ψ_1]-[K]-[x]-[α]	498	56.8	Melchior 2000; Rodriguez *et al.*, 2001
	Consensus	[Ψ_2]-[K]-[x]-[α]	591	67.4	
	Weak consensus	[Ψ_3]-[K]-[x]-[α]	598	68.2	
	PDSM	[Ψ_2]-[K]-[x]-[α]-[x]$_2$-[S]-[P]	32	3.6	Hietakangas *et al.*, 2006
	NDSM	[Ψ_2]-[K]-[x]-[α]-[x]-[α]$_{2/6}$	231	26.3	Yang *et al.*, 2006
	HCSM	[Ψ_4]$_3$-[K]-[x]-[E]	105	12.0	Matic *et al.*, 2010
	SC-SUMO	[P/G]-[x]$_{(0–3)}$-[I/V]-[K]-[x]-[E]-[x]$_{(0–3)}$-[P/G]	110	12.5	Benson *et al.*, 2007
	Minimal SC-SUMO	[I/V]-[K]-[x]-[E]-[x]$_{(0–3)}$-[P]	178	20.3	Subramanian *et al.*, 2003
	SUMO-acetyl switch	[Ψ_2]-[K]-[x]-[α]-[P]	130	14.8	Stankovic-Valentin *et al.*, 2007
	pSuM	[Ψ_2]-[K]-[x]-[$_p$S]-[P]	1	0.1	Picard *et al.*, 2012
Consensus inverted	Strong consensus	[α]-[x]-[K]-[Ψ_1]	30	3.4	Ivanov *et al.*, 2007 Matic *et al.*, 2010
	Consensus	[α]-[x]-[K]-[Ψ_2]	77	8.8	
	Weak consensus	[α]-[x]-[K]-[Ψ_3]	80	9.1	
	Non consensus		229	26.1	

Note: PDSM, phosphorylation-dependent SUMOylation motif; NSDM, negatively charged amino acid-dependent SUMOylation motif; HCSM, hydrophobic cluster SUMOylation motif; SC-SUMO, synergy control motif; pSuM, phosphorylated SUMOylation motif. The abundance [number (Nb) and percentage (%)] of each motif in the DB is indicated. Ψ_1 = I, L or V; Ψ_2 = A, F, I, L, M, P,V or W; Ψ_3 = A, F, G, I, L, M, P, V, W or Y; Ψ_4 = A, F, G, I, L, P or V; α = D or E; pS/T, phosphorylated serine/threonine.

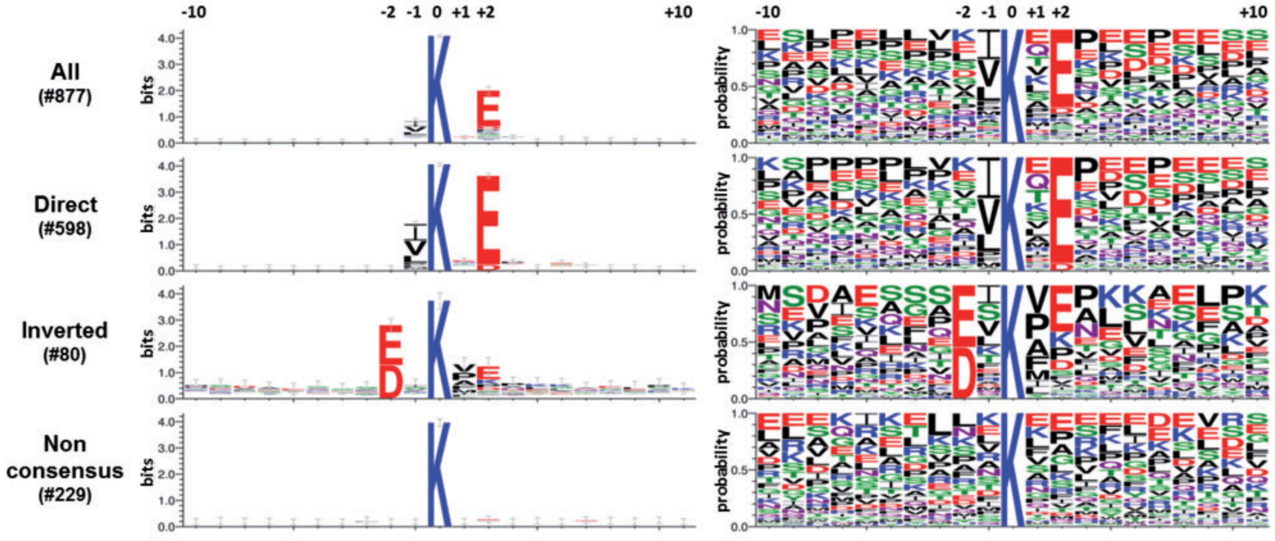

Fig. 1. Characterization of the collection of SUMOylation sites. The sequence logo representations of the 21-mer harboring the SUMO-acceptor K residue for motifs included in the cluster ALL, DIRECT or INVERTED were obtained with WebLogo3 (Crooks et al., 2004) using bits unit (left) or probability unit (right). The 5-mer (positions −2 to +2) used to define the frequency plots (Supplementary Table S1) is indicated

INVERTED (Fig. 1 and Supplementary Table S1A). A similar incidence of E (55%) and D (45%) is found at position −2. Non-polar residues are still enriched at position +1 (97.5 %). However, there is a greater variety of residues, with a higher rate of P (22.5%), A (15%), F (12.5%) and M (~9%), and a lower occurrence of I (~6.3%), L (~6.3%) and V (25%) compared with the sequences of the cluster ALL or DIRECT. Finally, the sequence logo for the non-consensus sites has an ambiguous profile, the SUMOylated K being the only conserved residue (Fig. 1). In agreement with the fact that the SUMOylation pathway is highly conserved among eukaryotes, similar sequence logos were obtained for the SUMOylation sites belonging to yeast proteins (Supplementary Fig. S2).

2.3 Characterization of the SIMs DB and motifs clustering strategy

The identification of SIMs, which act as SUMO recognition modules, is necessary to better understand the functional consequences of SUMOylation. To develop a protocol for computer-aided prediction, we generated a DB by collecting 102 experimentally validated SIMs, most of which are from human proteins (62.4%). The remaining sequences are from rodent (3.0%), yeast (12.9%) and viruses (16.8%) (Supplementary Fig. S1B).

Alignment of the hydrophobic core (positions −2 to +2) of these sequences showed that V and I represent more than 50% of the amino acids occurring at position −2 and −1, respectively, while having a rate ranging from 19 to 28% at other positions (Supplementary Table S1B). L is the most frequent residue at position +2 (45%). Although its occurrence is around 6% at other positions, L is nevertheless the more represented residue after V and I, the other amino acids having rates below 3%.

We manually curated the alignment of the 24-mer centered on the hydrophobic core for the 102 motifs of the DB and defined 5 types of SIMs (Fig. 2 and Table 2). The SIM consensus being degenerate, some experimental sites could fit with more than one pattern. Motifs

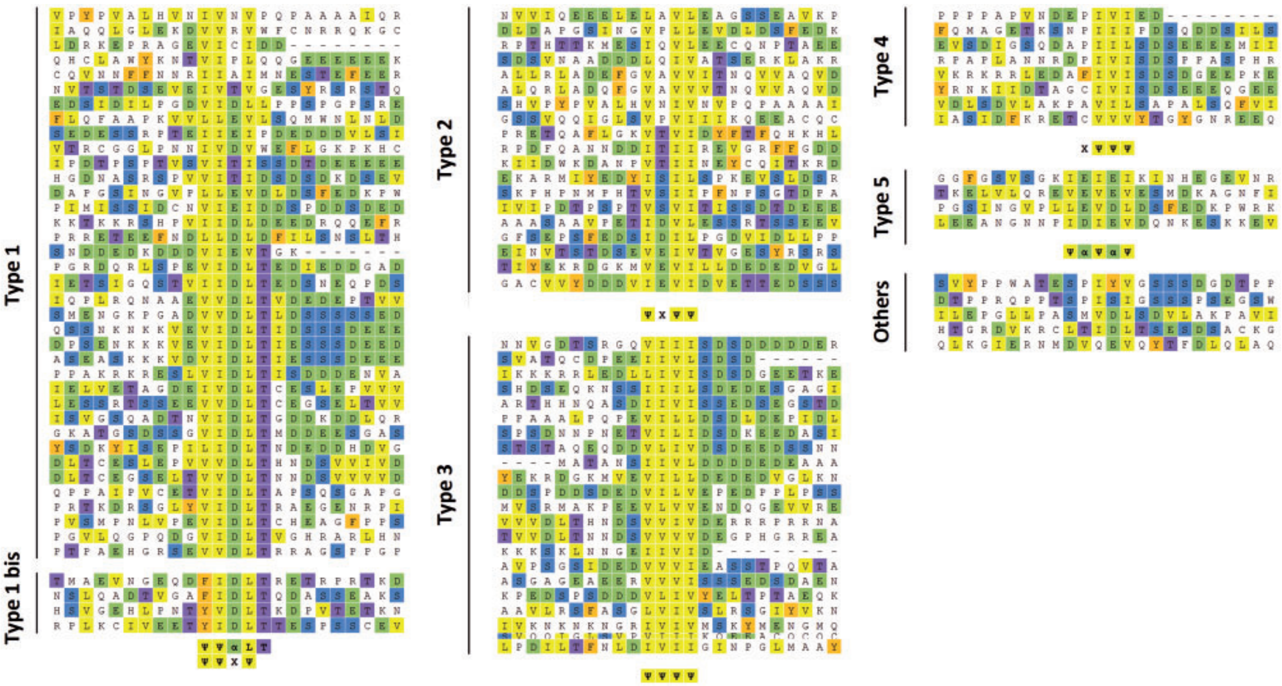

Fig. 2. Characterization of the collection of SIMs. Alignment of the 24 amino acid-long sequences harboring experimentally proved SIMs. The consensus motifs used to classify the sequences are shown below each cluster. Residues V, I, L (Ψ) are colored in yellow, D and E (α) in green, S in blue, T in purple and F and Y in orange

matching the consensus ΨΨxΨ (type 1), ΨxΨΨ (type 2), ΨΨΨΨ (type 3), xΨΨΨ (type 4) and ΨαΨαΨ (type 5) (where Ψ = V, I or L; α = D or E and x = any amino acids), represent ~60, 40, 20, 30 and 4% of the collection, respectively. A deeper analysis of the sequences of type 1 cluster showed that ~40% of the sites match with the V[I/V]DLT pattern. In this context, Sun and Hunter (2012) showed that aromatic residues can be accommodated at position −2 as in the SIMs of FLASH/CASP8AP2 and C5orf25. Thus, we named the group of sites which comprise an F or Y residue at position −2 'type 1 bis' cluster (Fig. 2). Only 5 out of 102 known SIMs could not be included in any cluster.

2.4 Determination of the prediction performances

To compute the area under the curve (AUC) of the Receiver Operating Characteristic (ROC) curves for the prediction of SUMOylation sites by JASSA using any of the three clusters (ALL, DIRECT, INVERTED) or GPS-SUMO, we used the R package pROC, notably the 'roc()' function with no smoothing method (R Core Team, 2015; Robin *et al.*, 2011).

The evaluation was performed using (i) the testing dataset of SUMOhydro which consists of 24 positive and 510 negative motifs (Chen *et al.*, 2012); (ii) the 4351 positive site identified by the Vertegaal's group in 1489 proteins (Hendriks *et al.*, 2014). As a matching negative dataset, we selected the 68325 K residues from the same proteins which SUMOylation was not detected. The results were plotted using the R package ggplot2 (Wickham, 2009).

3 Algorithm

3.1 Scoring strategy and definition of cut-off values for the prediction of SUMOylation sites

Sequence analysis of SUMOylation sites points to a hydrophobic and an acidic amino acids surrounding the target K (position 0) as major determinants of SUMO conjugation (Fig. 1). Thus, we established a scoring system where the occurrence of these residues, which are found at position −1 and +2 in direct consensus motifs or −2 and +1 in inverted consensus motifs, is used to quantify the eventuality of a K residue to lie within a SUMOylation site. The method is based on a PFM at four positions generated by aligning the sequences of the selected cluster (ALL, DIRECT or INVERTED) (Supplementary Table S1A). Each K residue of a query is given two predictive scores (PS) termed PSd and PSi. When either the cluster ALL or DIRECT is selected, the scores are defined as $\text{PSd} = f_{-1}(aa_{-1}) \times f_0(K_0) \times f_{+2}(aa_{+2}) \times 100$ and $\text{PSi} = f_{-1}(aa_{+1}) \times f_0(K_0) \times f_{+2}(aa_{-2}) \times 100$. When the cluster INVERTED is chosen, the scores are defined as $\text{PSd} = f_{+1}(aa_{-1}) \times f_0(K_0) \times f_{-2}(aa_{+2}) \times 100$ and $\text{PSi} = f_{+1}(aa_{+1}) \times f_0(K_0) \times f_{-2}(aa_{-2}) \times 100$. In these formulas, $f_p(aa_q)$ is the frequency at position p of the amino acid (amino acids) at position q for the selected cluster. Note that $f_0(K_0) = 1$ and the frequency of residues that are absent is set to 0.0001. Since the majority of SUMOylation sites of the DB conform to the canonical consensus pattern, hydrophobic residues are overrepresented at position −1 (Fig. 1 and Supplementary Table S1A). To avoid any bias when analyzing putative inverted SUMOylation sites, the frequencies of the amino acids at position +1 and −1 are not considered in the calculation of the PSd and the PSi, respectively.

Two cut-off values were defined by decision tree analysis for each cluster of SUMOylation sites (Fig. 3A–C). To this aim, a positive and a negative datasets composed of 358 bona fide SUMOylated motifs and 8071 non-SUMOylated sites, respectively (Chen *et al.*, 2012), were analyzed with JASSA choosing either the cluster ALL, DIRECT or INVERTED. The best score (either PSd or PSi) for each sequence tested was submitted to SIPINA Research (http://eric.univ-lyon2.fr/~ricco/sipina.html).

3.2 Scoring strategy and definition of cut-off values for the prediction of SIMs

For the detection of putative SIMs, the query protein is scanned against multiple consensus motifs (Table 2), and a PS is calculated

Table 2. List and proportion of known SIMs and abundance in the DB of JASSA

	Name	Motif	Nb	%	References
SIM	Type 1	$[\Psi_1]$-$[\Psi_1]$-[x]-$[\Psi_1]$	63	61.8	
	Type 2	$[\Psi_1]$-[x]-$[\Psi_1]$-$[\Psi_1]$	41	40.2	
	Type 3	$[\Psi_1]$-$[\Psi_1]$-$[\Psi_1]$-$[\Psi_1]$	23	22.5	This study
	Type 4	[x]-$[\Psi_1]$-$[\Psi_1]$-$[\Psi_1]$	31	30.4	
	Type 5	$[\Psi_1]$-$[\alpha]$-$[\Psi_1]$-$[\alpha]$-$[\Psi_1]$	4	3.9	This study; Ouyang *et al.*, 2009
	Type α	[V/I]-[x]-[V/I]-[V/I]	24	23.5	Song *et al.*, 2005
	Type β	[V/I]-[V/I]-[x]-[V/I/L]	54	52.9	
	Type a	[P/I/L/V/M]-[I/L/V/M]-[x]-[I/L/V/M]-$[\alpha/S]_3$	17	16.7	Miteva *et al.*, 2010; Sun and Hunter., 2012
	Type b	[P/I/L/V/M/F/Y]-[I/L/V/M]-[D]-[L]-[T]	24	23.5	
	Type r	$[\alpha/S]_3$-[I/L/V/M]-[x]-$[I/L/V/M/F]_2$	6	5.9	
	Type H	[K]-$[x]_{3\text{–}5}$-[I/V]-[I/L]-[I/L]-$[x]_3$-$[\alpha/Q/N]$-$[\alpha]_2$	7	6.9	Hannich *et al.*, 2005
	Type M	$[\Psi_2]_2$-[x]-[S]-[x]-[S/T]-$[\alpha]_3$	3	2.9	Minty *et al.*, 2000
	Non consensus		5	4.9	

Note: The abundance [number (Nb) and percentage (%)] of each motif in the DB is indicated. Ψ_1 = I, L or V; Ψ_2 = A, F, I, L, M, P, V or W; α = D or E.

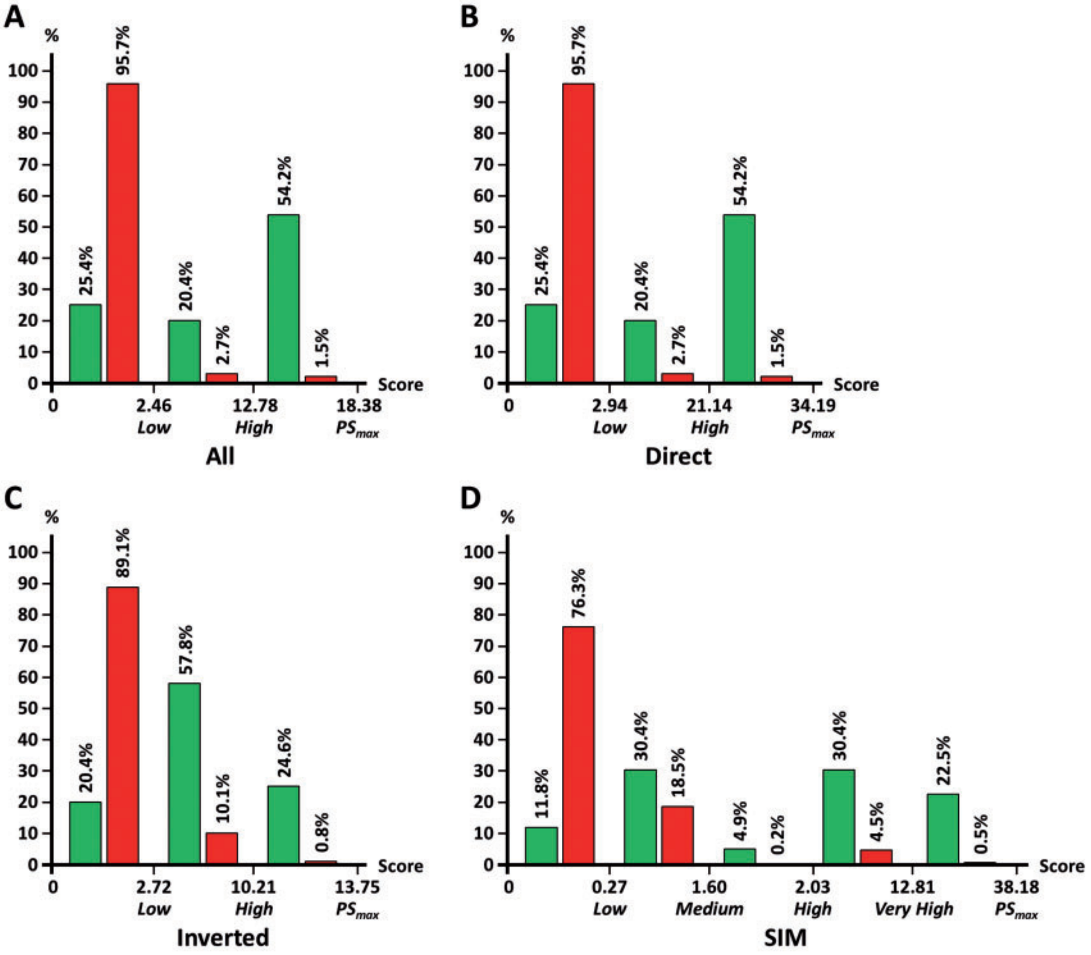

Fig. 3. Establishment of cut-off values for the prediction of SUMOylation sites and for SIM. Decision trees were generated with SIPINA Research v3.9, induction method C4.5, following analysis with JASSA of the SUMOylated and non-SUMOylated sequences from the control datasets of Chen *et al.* (2012) with cluster ALL **(A)**, DIRECT **(B)**, INVERTED **(C)**. The distribution (in %, *y*-axis) of the positive (green) and the negative (red) motifs of the control clusters after separation by the two cut-off values is shown. The cut-off values for the prediction of SIM motifs **(D)** were defined with the same methodology as for the SUMOylation sites. The distribution (in %, *y*-axis) of the positive (green) and the negative (red) motifs of the control datasets after separation by the four cut-off values is shown

for each tetrapeptide that fits at least one of them. The algorithm for SIMs prediction is based on a PFM at four positions derived from the alignment of the sequences of the training DB. The score is calculated as $\mathrm{PS} = 100 \times \prod_{i=1}^{4} f_i(aa_i)$, where f_i is the frequency of a given amino acid at position i in the frequency table (Supplementary Table S1B).

We defined four different thresholds for the SIM predictor with the same decision tree methodology used for the SUMOylation sites

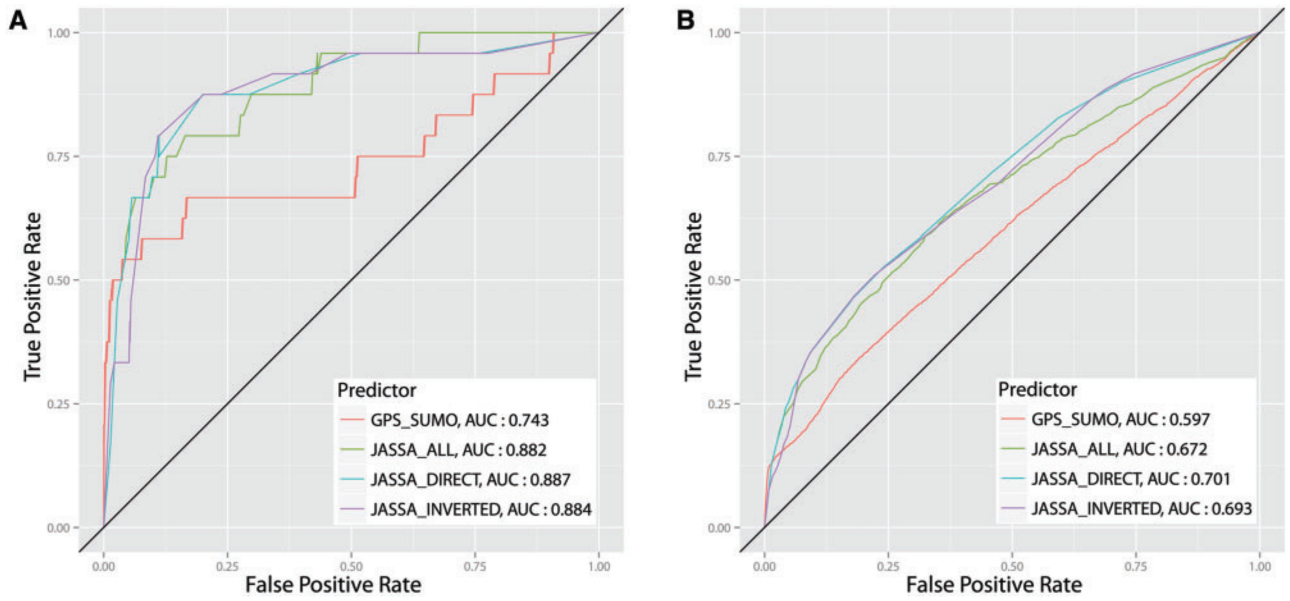

Fig. 4. Evaluation of the performances of JASSA and GPS-SUMO for the prediction of SUMOylation sites. The ROC curves were plotted and the AUC calculated. **(A)** AUC obtained using the SUMOhydro dataset. **(B)** AUC obtained using the Vertegaal's group dataset

(Fig. 3D). The positive test dataset consists of the 102 known SIMs. The negative test dataset is composed of the remaining 868 motifs that match at least one of the previously defined consensus patterns (Table 2) within the same proteins, and not included in the positive dataset.

3.3 Prediction of a negatively charged-residue stretch flanking a putative SIM

We manually curated the 10-amino acids long sequences upstream and downstream the 102 SIMs of the DB and arbitrarily decided on the presence or the absence of a [DES] stretch. Each sequence was given a score calculated as follows: $\mathrm{PS}_{\mathrm{stretch}} = v_i \times \sum_{j=i-2}^{i+2} v_j$, where $v_i = 1$ when E, D or S is found at position i; or $v_i = 0$ if any other residue is found position i. Next, the positive and negative clusters were analyzed with SIPINA Research and a cut-off was set at $\mathrm{PS}_{\mathrm{stretch}} = 5$ (Supplementary Fig. S3). The letter 'N' ($\mathrm{PS}_{\mathrm{stretch}} \leq 5$) or 'Y' ($\mathrm{PS}_{\mathrm{stretch}} > 5$) in the result output indicates the absence or the presence of a [DES] cluster, respectively.

3.4 Predictive performance assessment and comparison with other predictors

The prediction performances of JASSA were evaluated using the independent test datasets of Chen *et al.* (2012). Accuracy (Ac), sensitivity (Sn), specificity (Sp) and strength (St) were calculated as follows:

$$\mathrm{Ac} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}};\ \mathrm{Sn} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}};\ \mathrm{Sp} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}};$$
$$\mathrm{St} = \frac{\mathrm{Sn}+\mathrm{Sp}}{2};$$

TP (true positives) and TN (true negative) are the number of SUMOylated and non-SUMOylated K residues that were correctly predicted or rejected, respectively; FP (false positive) and FN (false negative) are the number of over-predicted or under-predicted K residues, respectively. The Matthews' correlation coefficient (MCC) is calculated as follows:

Table 3. Comparison of the performance of the SIM prediction tools of JASSA and GPS-SUMO

Method	Threshold	Sn (%)	Sp (%)	Ac (%)	St (%)	MCC
JASSA	**Low**	**73.9**	76.3	76.0	75.1	0.343
	Medium	50.0	94.0	89.3	72.0	0.442
	High	45.3	95.4	90.0	70.3	0.442
	Very High	20.9	**99.5**	**91.0**	60.2	0.388
GPS-SUMO	**Low**	72.8	76.3	75.9	74.6	0.336
	Medium	68.9	89.7	87.5	**79.3**	**0.489**
	High	44.4	95.4	89.9	69.9	0.435

Note: The highest value obtained for each parameter is in bold. Ac, accuracy; Sp, specificity; Sn, sensitivity; St, strength; MCC, Matthews' correlation coefficient.

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

The results of these assessments for JASSA are shown in Supplementary Table S2. The performances of JASSA were also benchmarked against GPS-SUMO (Zhao *et al.*, 2014). Indeed, the authors of GPS-SUMO proved the superiority of their algorithm over other available methods and their web-based service allows for batched prediction.

We used the test dataset from SUMOhydro (Chen *et al.*, 2012) and the dataset created from the extensive results of Vertegaal's lab (Hendriks *et al.*, 2014). We plotted the ROC curves and calculated the AUC for JASSA (with any of the three clusters ALL, DIRECT and INVERTED) and for GPS-SUMO.

In all the cases, JASSA displayed superior performances (Fig. 4). The performances of the SIM prediction tool of JASSA were evaluated by a bootstrap approach and compared with the SIM predictor of GPS-SUMO (Zhao *et al.*, 2014). The whole DB was randomly distributed in five equal-sized portions. Four portions were used to create a frequency table. The remaining 20% of the positive dataset was tested in parallel with 20% of the negative dataset to establish the number of TP, TN, FP and FN. This analysis was reiterated a thousand times. Since GPS-SUMO does not analyze some of the motifs contained in the DB of JASSA (19/102 positive and 140/826 negative), their score was set to 0, not to introduce a strong bias in favor of JASSA. Overall, the performances of JASSA are better or on par with those of GPS-SUMO (Table 3). Notably, JASSA features a 'Very High' threshold, which allows the identification of putative SIMs with Sp reaching 99.5% and an Ac of 91% (Table 3).

Altogether these results indicate that JASSA is a robust and accurate tool for the prediction of both SUMOylation sites and SIMs.

4 Implementation

4.1 Additional features of JASSA complement the scoring system and improve prediction evaluation

Identification of potential SUMOylation sites and SIMs faces major challenges represented by sequences matching the consensus but that are not functional (false positive) or functional sequences diverging from the consensus (false negative). We implemented several options that complement the scoring system towards a better evaluation of the predication.

- **Pattern search against extended motifs.** Although the sequence of the core consensus motifs is a key determinant, other factors like the local protein composition influence the propensity of a substrate to interact with SUMO in either a covalent or a non-covalent manner. For instance, more than a half of the experimental sites fitting the ΨKxα pattern are included within extended motifs characterized by a high occurrence of phosphorylatable and/or negatively charged residues (Table 1). JASSA systematically scans the sequence flanking each candidate site against known extended SUMOylation motifs or SIM variants (Tables 1 and 2). A positive match, which is indicated in the output with the corresponding sequence highlighted (Supplementary Fig. S4), might support the prediction of a functionally relevant site.
- **Analysis of the physico-chemical properties of adjacent amino acids.** Since extended variants of inverted SUMOylation motifs have not been characterized yet, we designed an option to highlight negatively charged, phosphorylatable or hydrophobic amino acids within the 21-mer surrounding each candidate site. By analyzing the sequence surrounding experimental SIMs we noted a high prevalence of negatively-charged and/or phosphorylatable residues. In particular, more than 50% of the known SIMs position +3 is occupied by E, S, D and T (Fig. 2). We therefore developed an algorithm to predict whether negatively charged amino acids and/or serine residues are enriched near the candidate site. Given that the consensus for SIM is degenerate, this information will be helpful to select or reject a candidate site.
- **Graphical representation of the prediction with secondary protein elements and within the 3D fold of the protein.** An important parameter which impacts on both SUMO conjugation and the SUMO-SIM interaction is the topology of the candidate site. Most validated SUMOylation events occur in extended loops or intrinsically disordered regions of the substrate outside its globular fold, and all known SIMs display a β-strand conformation (Sekiyama *et al.*, 2008; Song *et al.*, 2005). To have an insight on the local protein conformation at a candidate site, an option is provided to represent secondary structural elements extracted from a PDB file below the protein sequence where candidate sites are annotated allowing an easy comparison (Supplementary Fig. S4B and C). This feature might ease for instance the identification of tandem arrays of SIMs, a feature of SUMO-dependent Ubiquitin ligases (STUbL) (Sun and Hunter, 2012).
- When a PDB reference is provided, a PyMol script file is also generated, where candidate sites are annotated within the 3D fold of the protein of interest (Supplementary Fig. S4C). Altogether, these informations will help reject false positive, i.e. sites having a high PS because their sequence matches the consensus motif but which do not adapt a favorable conformation or are buried within the globular fold of the protein.
- **'DB hit' finder:** This feature indicates whether a candidate site matches a previously validated SUMOylation site or SIM. Each 5-mer centered on a K and each putative SIM of the query is systematically scanned against the proper DB to retrieve matching hits. The number of experimentally validated SUMOylation site or SIM corresponding to the candidate site is returned in the output. For each hit JASSA provides the link to the protein deposited on NCBI and UniProt and the PubMed reference of the publication. This option allows the detection of sites that diverge from the consensus and would otherwise be rejected because of their low PS. Moreover, this feature has been updated with a list of motifs known until 2014 (Hendriks *et al.*, 2014; Tammsalu *et al.*, 2014; Zhao *et al.*, 2014).

4.2 Use of JASSA and output description

JASSA is a freely accessible web-based tool which offers a user-friendly interface. The default parameters are designed to give the best results in most cases, but they can be adjusted for particular needs. The user can submit a sequence (protein or nucleotide sequence) in either Text or FASTA format, upload a FASTA file or input a UniProt protein ID. The results are returned in the form of a table-list reporting the position of the candidate K residue, the sequence of the 21-mer centered on it, the PS, the consensus type and the number of DB hits. Under default settings ('Analyze with: Best Predictions'), every 5-mer centered on a K is scored using independently the cluster ALL, DIRECT and INVERTED. The default parameter for the result output (only 'interesting' results) displays the sites that either have a PS at least above the low threshold or fit with a consensus motif or match a motif in the DB. A second table resumes the results for putative SIMs: position, PS, consensus type, presence of a potential acidic/serine stretch and the DB hit count. A graphical representation of the protein sequence of the query, where candidate sites are annotated is also provided. The results of a query are exportable under the GeneProt format. When a .pdb file is given, corresponding secondary structural elements can be visualized in the output and an annotated 3D structure can be downloaded under the form of a PyMol script.

5 Discussion

We designed JASSA to provide a comprehensive overview of potential SUMOylation sites and SIMs of a protein of interest, assisting in the selection of candidate sites for further experimentation.

Prediction of SUMO-modified K is based on a large DB encompassing 877 experimentally characterized SUMOylation sites, among which ~68% match the direct ΨKxα motif and ~9% the inverted αxKΨ motif. We also observed that either consensus pattern displays specific features in terms of amino acid composition. As an example, the acidic residue (α) is almost exclusively an E in direct sites, whereas a similar incidence of E and D is found in inverted motifs. Regarding the hydrophobic residue (Ψ), this position is occupied by P in more than 22% of the inverted sites, whereas this amino acid is found in only 4.2% of the direct sites. Based on these findings, we established two clusters including sites matching either the direct or the inverted SUMOylation consensus motif, to favor the prediction of SUMO-conjugation sites which are underrepresented in the full DB, such as inverted SUMOylation sites. As an example the peptide DVKA obtains a PSi of 0.174 or 0.114 using the clusters ALL or DIRECT, respectively. Both scores being below the low threshold (Fig. 3), this site would be rejected. The same query would receive a score above the threshold (PSi = 6.750), if the prediction is performed choosing the cluster INVERTED, and would therefore be considered further.

JASSA also allows identifying putative SIMs, a feature currently performed only by GPS-SUMO (Zhao *et al.*, 2014). SIMs are found in a wide variety of proteins such as downstream effectors of the SUMO pathway and SUMO enzymes. Some SUMO substrates [i.e. Daxx (Chang *et al.*, 2011) and USP25 (Meulmeester *et al.*, 2008)] also harbor SIMs, which were proposed to recruit SUMO-loaded Ubc9 leading to the conjugation of SUMO to proximal sites including K within non-consensus sequences.

The major shortcoming of SUMOylation sites and SIMs prediction is the difficulty to correctly classify the false positive and the false negative. To overcome this problem, we developed features that can help the user to refine the prediction. Indeed, JASSA's output can give informations about the structure and the physico-chemical environment of the candidate site and whether or not it has already been experimentally validated.

We benchmarked JASSA against existing predictors and found that it reaches better or equivalent levels of performance. Thus, we believe that JASSA will be a valuable tool to help biologists identifying the best candidate sites for experimentations aiming to better understand SUMOylation and its contribution to cell biology.

Acknowledgements

We are grateful to Dr JP Feugeas (UP7), M. Montes (CNAM) and all lab members for discussion.

Funding

EU FP7 [HEALTH-2012-INNOVATION-1 'HIVINNOV'] (Grant no. 305137). G.B. is supported by the French Ministère de la Recherche.

Conflict of Interest: none declared.

References

Apweiler,R. *et al.* (2004) UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res.*, **32**, D115–D119.

Benson,M.D. *et al.* (2007) SUMO modification regulates inactivation of the voltage-gated potassium channel Kv1.5. *Proc. Natl. Acad. Sci. U. S. A.*, **104**, 1805–1810.

Blomster,H.A. *et al.* (2009) Novel proteomics strategy brings insight into the prevalence of SUMO-2 target sites. *Mol. Cell. Proteomics*, **8**, 1382–1390.

Chang,C.-C. *et al.* (2011) Structural and functional roles of Daxx SIM phosphorylation in SUMO paralog-selective binding and apoptosis modulation. *Mol. Cell*, **42**, 62–74.

Chen,Y.-Z. *et al.* (2012) SUMOhydro: a novel method for the prediction of sumoylation sites based on hydrophobic properties. *PLoS One*, **7**, e39195.

Crooks,G.E. *et al.* (2004) WebLogo: a sequence logo generator. *Genome Res.*, **14**, 1188–1190.

Droescher,M. *et al.* (2013) SUMO rules: regulatory concepts and their implication in neurologic functions. *Neuromolecular Med.*, **15**, 639–660.

Flotho,A. and Melchior,F. (2013) Sumoylation: a regulatory protein modification in health and disease. *Annu. Rev. Biochem.*, **82**, 357–385.

Gareau,J.R. and Lima,C.D. (2011) The SUMO pathway: emerging mechanisms that shape specificity, conjugation and recognition. *Nat. Rev. Mol. Cell Biol.*, **11**, 861–871.

Hannich, J.T. *et al.* (2005) Defining the SUMO-modified proteome by multiple approaches in *Saccharomyces cerevisiae*. *J. Biol. Chem.*, **280**, 4102–4110.

Hecker,C.-M. *et al.* (2006) Specification of SUMO1- and SUMO2-interacting motifs. *J. Biol. Chem.*, **281**, 16117–16127.

Hendriks,I.A. *et al.* (2014) Uncovering global SUMOylation signaling networks in a site-specific manner. *Nat Struct. Mol. Biol.*, **21**, 927–936.

Hietakangas,V. *et al.* (2006) PDSM, a motif for phosphorylation-dependent SUMO modification. *Proc. Natl. Acad. Sci. U. S. A.*, **103**, 45–50.

Hornbeck,P.V. *et al.* (2004) PhosphoSite: a bioinformatics resource dedicated to physiological protein phosphorylation. *Proteomics*, **4**, 1551–1561.

Impens,F. *et al.* (2014) Mapping of SUMO sites and analysis of SUMOylation changes induced by external stimuli. *Proc. Natl Acad. Sci. U.S.A.*, **111**, 12432–12437.

Ivanov,A.V. *et al.* (2007) PHD domain-mediated E3 ligase activity directs intramolecular sumoylation of an adjacent bromodomain required for gene silencing. *Mol. Cell*, **28**, 823–837.

Kerscher,O. (2007) SUMO junction-what's your function? New insights through SUMO-interacting motifs. *EMBO Rep.*, **8**, 550–555.

Matic,I. *et al.* (2010) Site-specific identification of SUMO-2 targets in cells reveals an inverted SUMOylation motif and a hydrophobic cluster SUMOylation motif. *Mol. Cell*, **39**, 1–19.

Melchior,F. (2000) SUMO—nonclassical ubiquitin. *Annu. Rev. Cell Dev. Biol.*, **16**, 591–626.

Meulmeester,E. *et al.* (2008) Mechanism and consequences for paralog-specific sumoylation of ubiquitin-specific protease 25. *Mol. Cell*, **30**, 610–619.

Minty,A. *et al.* (2000) Covalent modification of p73alpha by SUMO-1. Two-hybrid screening with p73 identifies novel SUMO-1-interacting proteins and a SUMO-1 interaction motif. *J. Biol. Chem.*, **275**, 36316–36323.

Miteva,M. *et al.* (2010) Sumoylation as a signal for polyubiquitylation and proteasomal degradation. *Subcell. Biochem.*, **54**, 195–214.

Mohideen,F. *et al.* (2009) A molecular basis for phosphorylation-dependent SUMO conjugation by the E2 UBC9. *Nat. Struct. Mol. Biol.*, **16**, 945–952.

Ouyang,J. *et al.* (2009) Direct binding of CoREST1 to SUMO-2/3 contributes to gene-specific repression by the LSD1/CoREST1/HDAC complex. *Mol. Cell*, **34**, 145–154.

Picard,N. *et al.* (2012) Identification of estrogen receptor *β* as a SUMO-1 target reveals a novel phosphorylated sumoylation motif and regulation by glycogen synthase kinase 3*β*. *Mol. Cell. Biol.*, **32**, 2709–2721.

Pichler,A. *et al.* (2005) SUMO modification of the ubiquitin-conjugating enzyme E2-25K. *Nat. Struct. Mol. Biol.*, **12**, 264–269.

R Core Team (2015) *R: a Language and Environment for Statistical Computing*. R Foundation for Statistical Computing Vienna, Austria.

Ren,J. *et al.* (2009) Systematic study of protein sumoylation: Development of a site-specific predictor of SUMOsp 2.0. *Proteomics*, **9**, 3409–3412.

Robin,X. *et al.* (2011) pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, **12**, 77.

Rodriguez,M.S. *et al.* (2001) SUMO-1 conjugation in vivo requires both a consensus modification motif and nuclear targeting. *J. Biol. Chem.*, **276**, 12654–12659.

Stankovic-Valentin,N. *et al.* (2007) An acetylation/deacetylation-SUMOylation switch through a phylogenetically conserved psiKXEP motif in the tumor suppressor HIC1 regulates transcriptional repression activity. *Mol. Cell. Biol.*, **27**, 2661–2675.

Sampson,D.A. *et al.* (2001) The small ubiquitin-like modifier-1 (SUMO-1) consensus sequence mediates Ubc9 binding and is essential for SUMO-1 modification. *J. Biol. Chem.*, **276**, 21664–21669.

Sarge,K.D. and Park-Sarge,O.-K. (2009) Sumoylation and human disease pathogenesis. *Trends Biochem. Sci.*, **34**, 200–205.

Sekiyama,N. *et al*. (2008) Structure of the small ubiquitin-like modifier (SUMO)-interacting motif of MBD1-containing chromatin-associated factor 1 bound to SUMO-3. *J. Biol. Chem*., **283**, 35966–35975.

Song,J. *et al*. (2005) Small ubiquitin-like modifier (SUMO) recognition of a SUMO binding motif: a reversal of the bound orientation. *J. Biol. Chem*., **280**, 40122–40129.

Subramanian,L. *et al*. (2003) A synergy control motif within the attenuator domain of CCAAT/enhancer-binding protein alpha inhibits transcriptional synergy through its PIASy-enhanced modification by SUMO-1 or SUMO-3. *J. Biol. Chem*., **278**, 9134–9141.

Sun,H. and Hunter,T. (2012) Poly-small ubiquitin-like modifier (PolySUMO)-binding proteins identified through a string search. *J. Biol. Chem*., **287**, 42071–42083.

Tammsalu,T. *et al*. (2014) Proteome-wide identification of SUMO2 modification sites. *Sci. Signal*., **7**, rs2.

Tatham,M.H. *et al*. (2011) Comparative proteomic analysis identifies a role for SUMO in protein quality control. *Sci. Signal*., **4**, rs4.

Teng,S. *et al*. (2012) Predicting protein sumoylation sites from sequence features. *Amino Acids*, **43**, 447–455.

Wickham,H. (2009) ggplot2: elegant graphics for data analysis Springer New York.

Wilson,V.G. (2012) Sumoylation at the Host-Pathogen Interface. *Biomolecules*, **2**, 203–227.

Wimmer,P. *et al*. (2012) Human pathogens and the host cell SUMOylation system. *J. Virol*., **86**, 642–654.

Yang,S.-H. *et al*. (2006) An extended consensus motif enhances the specificity of substrate modification by SUMO. *EMBO J*., **25**, 5083–5093.

Zhao,Q. *et al*. (2014) GPS-SUMO: a tool for the prediction of sumoylation sites and SUMO-interaction motifs. *Nucleic Acids Res*., **42**, W325–W3230.