Conjugation of small ubiquitin-like modifier (SUMO) to substrates is involved in a large number of cellular processes. Typically, SUMO is conjugated to lysine residues within a SUMO consensus site; however, an increasing number of proteins are sumoylated on non-consensus sites. To appreciate the functional consequences of sumoylation, the identification of SUMO attachment sites is of critical importance. Discovery of SUMO acceptor sites is usually performed by a laborious mutagenesis approach or using MS. In MS, identification of SUMO acceptor sites in higher eukaryotes is hampered by the large tryptic fragments of SUMO1 and SUMO2/3. MS search engines in combination with known databases lack the possibility to search MSMS spectra for larger modifications, such as sumoylation. Therefore, we developed a simple and straightforward database search tool ("ChopNSpice") that successfully allows identification of SUMO acceptor sites from proteins sumoylated in vivo and in vitro. By applying this approach we identified SUMO acceptor sites in, among others, endogenous SUMO1, SUMO2, RanBP2, and Ubc9. Molecular & Cellular Proteomics 8:2664–2675, 2009.

Post-translational modification with ubiquitin and ubiquitin-like modifiers (Ubls) such as SUMO plays an important role in most, if not all, cellular processes (1–6). Conjugation of Ubls to their targets involves an isopeptide bond between the carboxyl group of the modifier and the ε-amino group of a lysine residue within the targets. Attachment of Ubls to specific targets involves an enzymatic cascade. First the Ubls are processed to expose their C-terminal diglycine motif. The mature Ubl is then transferred to its target via a cascade of E1 (activating), E2 (conjugating), and E3 (ligase) enzymes. The conjugation system for SUMO consists of a heterodimeric activating enzyme, Aos1/Uba2; a conjugating enzyme, Ubc9; and E3 ligases, such as RanBP2 or members of the PIAS family. The conjugation status undergoes perpetual change and is governed by a small family of SUMO proteases that hydrolyze the isopeptide bond between SUMO and its target (7, 8). Although in lower eukaryotes only one SUMO is present, vertebrates express at least three different SUMO paralogs: SUMO1, SUMO2, and SUMO3. Mature SUMO2 and SUMO3 (referred to as SUMO2/3) are 97% identical but differ substantially from SUMO1 (~50% identity).

Although the list of known SUMO substrates is growing rapidly, our understanding of the functional consequences for many of these targets is lagging behind. At a molecular level, the functional consequences of SUMO conjugation can be explained by a gain or loss of interaction with other macromolecules (3, 4). SUMO-dependent intramolecular conformational changes have also been described (9, 10). Thus, to appreciate the role that SUMO plays in the regulation of specific substrates, identification of the acceptor site(s) for SUMO conjugation is of key importance.

So far, identification of SUMO acceptor sites has relied largely on mutation of the SUMO consensus site, which consists of a short motif with the sequence ψKXE (ψ represents a bulky hydrophobic residue, and X represents any amino acid). This motif is recognized by Ubc9 if presented in an extended conformation (11–13). However, an increasing number of proteins, such as PCNA, E2-25K, Daxx, and USP25, turned out to be sumoylated on lysine residues that do not conform to the SUMO consensus site (14–17). For this category of proteins, as well as for proteins that contain a large number of SUMO consensus sites, the identification of acceptor lysines is a burdensome task that often involves mutagenesis of each lysine residue within the substrate in turn.

From the ‡Bioanalytical Mass Spectrometry Group, §§Department of NMR-based Structural Biology, Max Planck Institute for Biophysical Chemistry, Am Fassberg 11, 37077 Göttingen, Germany, †Department of Biochemistry I, Faculty of Medicine, Georg August University of Göttingen, Humboldtallee 23, 37073 Göttingen, Germany, ||Department of Molecular Cell Biology, Leiden University Medical Center, 2300 RC Leiden, The Netherlands, and ¶¶Center for Molecular Biology (ZMBH), University of Heidelberg, Im Neuenheimer Feld 282, 69120 Heidelberg, Germany

Received, February 17, 2009, and in revised form, July 10, 2009
Published, MCP Papers in Press, August 31, 2009, DOI 10.1074/mcp.M900087-MCP200

© 2009 by The American Society for Biochemistry and Molecular Biology, Inc.
This paper is available on line at http://www.mcponline.org
Identification of SUMO Conjugation Sites in Vivo

MS is currently one of the state-of-the-art technologies to identify protein factors and their post-translational modifications in an unbiased and sensitive manner. Several groups have shown that, using overexpressed tagged SUMO, MS can be efficiently exploited to identify endogenous substrates for SUMO conjugation (18–20). However, the identification of SUMO acceptor lysines using MS has remained a more challenging task (18, 21, 23, 24). So far, using tagged SUMO, unbiased identification of acceptor lysines for endogenous substrates has only been observed in Saccharomyces cerevisiae (18). The identification of substrates in higher eukaryotes has been hampered by the large conjugated SUMO peptide that arises upon tryptic digestion (>2154 Da with human SUMO1 and >3568 Da with human SUMO2/3 compared with 484 Da for Smt3 in S. cerevisiae). Such large fragments, in addition to the mass of the conjugated peptide, can impede their in-gel digestion, extraction, detection, and sequencing in MS. To overcome some of these limitations, several different strategies have been developed: 1) mutation of the tryptic fragment of SUMO, yielding a smaller tryptic fragment (23), 2) development of an automated recognition pattern tool (SUMmOn) (24), and 3) identification of targets using an in vitro approach (21). Although these approaches have been applied successfully for the identification of SUMO conjugates in vitro and in vivo, unbiased identification of SUMO conjugates in vivo has not been achieved in higher eukaryotes. Another hurdle to such identification of SUMO conjugates is the variety of masses that can theoretically arise for just one SUMO-conjugated lysine in a given protein because of tryptic miscleavages. Thus, the unambiguous identification of SUMO acceptor sites requires the mass of the modified peptide carrying the conjugated SUMO (fragment) to be measured with high accuracy, and most importantly, it requires sequence analysis of the modified peptides. Because available proteomics search engines lack the possibility to search MSMS spectra for larger modifications, e.g. those that occur upon sumoylation, we developed a novel, simple, and straightforward database search tool (“ChopNSpice”) that, in combination with current proteomics search engines (such as MASCOT (25) or SEQUEST (26)), allows one to identify SUMO1 and SUMO2/3 acceptor sites unambiguously. We confirmed this strategy in vitro on various substrates and demonstrate the power of this technique by the identification of acceptor lysines within several endogenous targets from HeLa cells.

EXPERIMENTAL PROCEDURES

Software—ChopNSpice is written in PHP. The software tools that we have developed and presented in this study, along with further documentation, are freely available on line and also released as open source under the terms of the General Public License v3 (GPLv3).

In Vitro Sumoylation Assays—SUMO conjugation reactions were performed at 30 °C for 1 h in the presence or absence of 5 mM ATP in 20 μl of TB (20 mM Heps/KOH, pH 7.3, 110 mM potassium acetate, 2 mM magnesium acetate, 0.5 mM EGTA, 1 mM DTT supplemented with protease inhibitors). Reactions contained 100 ng of Aos1/Uba2, 200 ng of Ubc9, 2.5 μg of SUMO1 or SUMO2, and 1 μg of target protein (GST-p53, mouse RanGAP1, GST-Sp100, or Aos1/Uba2) in a volume of 20 μl.

Cell Culture, Immunoprecipitation, and Immunoblotting—HeLa-S3 cells were maintained in Joklik’s medium supplemented with 10% fetal bovine serum and antibiotics. To immunoprecipitate SUMO1 conjugates, 1 × 10⁶ HeLa cells were washed twice with PBS containing 10 mM NEM and lysed in 2 pellet volumes of radioimmunoprecipitation assay buffer (20 mM NaP, pH 7.4, 150 mM NaCl, 1% Triton, 0.5% sodium deoxycholate, 0.1% SDS) supplemented with protease inhibitors and 10 mM NEM. Lysates were centrifuged (16,000 × g for 15 min at 4 °C) and filtered (0.45 μm) prior to addition of 25 μg of monoclonal α-SUMO1 antibodies. After 2-h incubation at 4 °C, the lysates were centrifuged (16,000 × g for 15 min at 4 °C), and the supernatant was incubated for another 2 h at 4 °C with protein G-agarose. After collection and extensive washing of bound proteins, samples were eluted with 2× sample buffer and separated by SDS-PAGE followed by Coomassie staining or Western blotting. In a second larger experiment, 1 × 10⁹ cells were lysed in TB (with 0.1% Triton and 10 mM ATP) and treated with 10 mM NEM after lysis. Immunoprecipitation using 100 μg of GMP1 antibodies was similar to that described above. The SUMO acceptor site in RanGAP1 was observed in both purification methods, whereas the other targets were identified in the second scaled up experiment.

Antibodies—Mouse monoclonal α-SUMO1 antibodies were kindly provided by M. Matunis, and goat anti-SUMO1 antibodies have been described previously (27, 28). Secondary antibodies were obtained from Jackson ImmunoResearch Laboratories.

Plasmids—Plasmids for bacterial expression of Aos1/Uba2, Ubc9, SUMO1, SUMO2 (GenBankTM accession number NM_006937), GST-Sp100, and RanGAP1 have been described previously (17, 29). A plasmid for GST-p53 was kindly provided by Dr. Moshe Oren.

Recombinant Proteins—Protein purification for SUMO1, SUMO2, Aos1/Uba2, Ubc9, RanGAP1, GST-Sp100, GST-p53, and PIAS1 has been described previously (17, 29–31).

Mass Spectrometry and Data Analysis—SUMO-conjugated proteins were excised from the gel, reduced with 50 mM DTT for 1 h, alkylated for 1 h with 100 mM iodoacetamide, and in-gel digested with modified trypsin (Promega) overnight, all at 37 °C. SUMO-conjugated proteins from solution were reduced with 50 mM DTT for 1 h, alkylated for 1 h with 100 mM iodoacetamide, and subsequently digested with modified trypsin overnight, all at 37 °C. Tryptic peptides were dissolved in 2 μl of 50% acetonitrile with 0.1% formic acid and added to 18 μl of 0.1% formic acid for further MS analysis. MS analysis was performed by nanoscale LC-MSMS using an LTQ-Orbitrap mass spectrometer (Thermo Fisher Scientific) equipped with a nanoelectrospray ion source and coupled to an Agilent 1100 HPLC system (Agilent Technologies) fitted with a self-made C18 trap column. Tryptic peptides were first loaded at a flow rate of 10 μl/min onto a C18 trap column (1.5 cm, 360-μm outer diameter, 150–μm inner diameter, Reprosil-Pur 120 Å, 5 μm, C18-AQ, Dr. Maisch GmbH, Ammerbuch-Entringen, Germany). Retained peptides were eluted and separated on a analytical C18 capillary column (15 cm, 360-μm outer diameter, 75-μm inner diameter, Reprosil-Pur 120 Å, 5 μm, C18-AQ, Dr. Maisch GmbH) at a flow rate of 300 nl/min with a gradient from 7.5 to 37.5% ACN in 0.1% formic acid for 60 min. Typical MS conditions were as follows: spray voltage of 1.8 kV, heated capillary temperature of 150 °C, and normalized CID collision energy of 37.5% for MSMS in the LTQ. An activation q = 0.25 and activation time of 30 ms were used. The mass spectrometer was operated in the data-dependent mode to automatically switch between MS and MSMS acquisition. Survey full-scan MS spectra (from m/z 250 to 2000) were acquired in

Molecular & Cellular Proteomics 8.12 2665
Identification of SUMO Conjugation Sites in Vivo

the orbitrap with resolution $R = 30,000$ at $m/z$ 400 (after accumulation to a “target value” of 1,000,000 in the orbitrap). The five most intense ions were isolated sequentially and fragmented in the linear ion trap using CID at a target value of 100,000. For all measurements with the orbitrap detector a lock mass ion from ambient air ($m/z$ 445.120025) was used for internal calibration. For high mass data-dependent mode, the mass range for selecting MS data-dependent masses was 2154–1,000,000 and 3568–1,000,000 for SUMO1 and SUMO2/3, respectively, using $m/z$ values as masses. For protein identification, all MSMS spectra were searched against a Swiss-Prot database using MASCOT with the following parameters: mass tolerance of 10 ppm in MS mode and 0.8 Da in MSMS mode; allow up to two missed cleavages; consider methionine oxidation and cysteine carboxamidomethylation as variable modifications. The sequence of the protein of interest was manually saved to a FASTA file, and ChopNSpice was used to create a new FASTA file with the following parameters: spice species was $H. sapiens$; spice sequences were SUMO1 and SUMO2, respectively; spice site was KX; spice mode was once per fragment; include unmodified fragments in output; enzyme was trypsin (Lys/Arg, do not cleave at Pro); allow up to three protein miscleavages; allow up to one miscleavages in the “spice sequence”; output formatting was FASTA (single protein sequence); mark all cleaved sites (“J”); retain comments in FASTA format without line breaks in FASTA output. For sumoylated site identification with MASCOT or SEQUEST, all MSMS spectra were searched against a new FASTA file that was created by ChopNSpice with the following parameters: mass tolerance of 10 ppm in MS mode and 0.8 Da in MSMS mode; allow zero missed cleavages; consider methionine oxidation and cysteine carboxamidomethylation as variable modifications; enzyme cleaved at J at N and C termini for MASCOT or no enzyme must be used for SEQUEST. If the search was performed with the in-house MASCOT server, the file “quant_subs.pl” must be changed from J to N and C termini for MASCOT or no enzyme must be used for SEQUEST. The MSMS spectra were confirmed manually to identify the SUMO acceptor site. The symbol of the amino acid that was before and after the identified SUMO conjugated peptide must be J. All high abundance peaks had to be assigned to y- or b-ion series.

RESULTS

ChopNSpice—A typical work flow in MS-based proteomics comprises digestion of proteins with endoproteinases, separation of the generated peptides by LC, and ionization and subsequent fragmentation of the peptides. Finally, automated searching of the fragment spectra against a database allows identification of the corresponding protein (for a review, see Aebersold and Mann (32)). Identification of post-translational modifications by MS requires, in addition to a highly accurate mass determination of the precursor, sequencing of the peptide that contains the modification.

Accordingly, our approach to identify SUMO acceptor sites is based on the fragmentation pattern of conjugated sumoylated peptides after digestion with trypsin. Such digestion results in peptides in which a missed (i.e. non-cleaved because of SUMO modification) lysine residue is branched with a SUMO tryptic peptide (Fig. 1A). In practice, we and others observed that the MSMS fragmentation of such a branched peptide is similar to the fragmentation of a linear tryptic peptide that has a non-cleaved lysine residue and the SUMO peptide at its N terminus (Fig. 1, A and B) (21, 33). Identification of SUMO acceptor lysines using such MSMS spectra in a database search is only possible when the peptide sequences within the database are also modified by SUMO. However, available search engines for experimental fragment spectra do not include SUMO as a putative modification at lysine residues. Simple addition of the molecular weight of the tryptic SUMO fragment to that of a lysine residue within the target protein, without obtaining sequence information, would generate a large number of false positive hits in database searches. In addition, because sumoylation can theoretically occur at every lysine residue within a protein, manual construction of such artificial peptides is a time-consuming process. Accordingly, we generated an algorithm to automate the generation of such SUMO-modified FASTA sequences of proteins in silico (Fig. 2A). Subsequently, the novel FASTA sequences are implemented in a database search with commonly used search engines to identify acceptor sites for SUMO conjugation (Fig. 2B).

More specifically, the FASTA sequence of a putatively sumoylated protein is “chopped” into tryptic fragments (allowing 0, 1, 2, or n missed cleavages). The tryptic “spice” sequence (e.g. tryptic peptides from SUMO1 or any other ubiquitin-like protein) is attached to the N terminus of each tryptic peptide that contains a Lys as a missed cleavage site. It is of note that also the ubiquitin-like proteins are allowed to contain 0, 1, 2, or n miscleavage(s). To prevent the appearance of non-natural peptides, a virtual amino acid, J, is attached to the C terminus of each tryptic fragment before ligation of the generated tryptic fragments into one large FASTA sequence. This large artificial protein sequence is submitted into the database search in which the virtual cleavage site J is recognized by an artificial endoproteinase that directly cleaves N- and C-terminally to J to generate the tryptic fragments for the selected missed cleavages. Subsequently, the SUMO acceptor site can be identified by using the applied search engine (e.g. MASCOT, XITandem, or SEQUEST). A work flow to set up a modified FASTA sequence in which certain proteins (or entire databases) can be generated by a user-defined modifier is implemented in the program ChopNSpice.

In practical terms, after enrichment of endogenous SUMO-conjugated proteins or proteins sumoylated in vitro, putative SUMO substrates are identified by a standard MS-based protein identification; i.e. samples are digested with trypsin, and the tryptic fragments are separated by LC, detected, and sequenced by MS. Corresponding proteins in the sample are identified by (i) the highly accurate mass of the peptide and (ii) searching the fragment spectra against a database using e.g. MASCOT, XITandem, or SEQUEST as search engine. A second MS and MSMS analysis under “high mass” conditions is performed where only those precursors are selected for sequencing that exceed a certain size, i.e. ≥2154 Da for SUMO-1 and ≥3568 Da for SUMO-2/3 (see also below).

Once one or several putatively sumoylated proteins have been identified in both the analyses after merging the data/
results, MS and MSMS data are resubmitted for search against the database containing the virtual sumoylated protein sequence generated by ChopNSpice (Fig. 2B). In a subsequent experiment, the same sample can be reinvestigated by extended/modified LC-MSMS analysis to identify the SUMO acceptor site(s).

Note that both of the search engines used in this study (MASCOT and SEQUEST) have some shortcomings. MASCOT
Identification of SUMO Conjugation Sites \textit{in Vivo}

\section*{A}

\begin{center}
\begin{figure}
\centering
\includegraphics[width=\textwidth]{Fig2A}
\caption{Concept of ChopNSpice software. \textbf{A}, basic work flow of ChopNSpice to generate a “spiced” FASTA sequence from an initial protein sequence in which all lysine residues are putatively modified by SUMO1 or SUMO2/3. The spiced FASTA sequence is subsequently used in database searches (see text for details). \textbf{B}, general work flow for identification of SUMO acceptor sites. Sumoylated proteins are
\end{figure}
\end{center}
for instance does not efficiently search fragment spectra that contain fragment ions with a charge state higher than 2; as a consequence, larger sumoylated peptides with charge state of 4+ show a very low score in MASCOT searches or are not identified at all (data not shown). This problem can be circumvented by using either SEQUEST or other search engines (e.g. X!Tandem) or, alternatively, by using the software tool Raw2msn to deconvolute the higher charge stages of the fragment ions in the raw data to singly charged fragment ions for MASCOT search (34). However, a prerequisite for deconvolution is that MSMS spectra (generated either by CID or by high energy collision-induced dissociation) are recorded in the FT analyzer/detector of the orbitrap with sufficient resolution for charge state recognition, and this in turn decreases sensitivity (35). A comparison between the different systems for processing raw data and the different detection modes of the orbitrap mass spectrometer are shown in supplemental Fig. S1. SEQUEST on the other hand does not allow for cleavage with endoproteinase both N- and C-terminally to J but rather either N-terminally or C-terminally. Therefore, cleavage of the FASTA sequence is performed unspecifically; i.e. no enzyme is used in silico, and matched spectra are validated manually. Confidence in the results from the search engine is achieved by the high mass accuracy of the orbitrap instrument (<10 ppm) and by the fact that the validated sequence must be preceded or followed by the virtual amino acid J. Furthermore, all the abundant fragment ions must be assigned to y- and/or b-ion series. However, as a very simple alternative, the single concatenated peptide sequences can be submitted to the database without merging them into a single new FASTA sequence.

Identification of SUMO Conjugation Sites in Vivo—To validate our approach, we applied RanGAP1, Sp100, and p53 to an in vitro sumoylation reaction with SUMO1 (Fig. 3, A–C) and SUMO2 (data not shown). Proteins migrating on SDS-PAGE with a higher apparent molecular weight than the original proteins were considered to be sumoylated and were processed by LC-MSMS as described above. For identification of sumoylated peptides, we first tested SUMO as a variable modification of lysines (2154 Da for SUMO1 and 3568 Da for SUMO2) using two commonly used peptide identification tools, MASCOT and SEQUEST. However, like other groups (24), we were unable to identify any sumoylated peptides by the standard LC-MS proteomics and subsequent database search (see supplemental Data S1). Although manual identification of SUMO conjugation sites was possible, it required extensive searching in the MS spectra for modified peptides (17). In contrast, by using the ChopNSpice software on the identified protein sequences and subsequent database search with MASCOT and SEQUEST, we readily identified SUMO modification of RanGAP1 on lysine 526, of p53 on lysine 386, and of Sp100 on lysine 297 (Fig. 3, A–C). In addition, we observed several minor acceptor sites, also observed by others (Fig. 3, A–C, supplemental Table S1, and corresponding MASCOT search results and annotated spectra for RanGAP1, Sp100, and p53 are listed in supplemental Data S2) (36–38). Furthermore, we discovered that numerous, so far unidentified, lysine residues within the SUMO E1 activating enzyme Uba2 are conjugated with SUMO1 and SUMO2 (Fig. 3D and supplemental Data S2 and Table S1). Consistent with the identification of multiple acceptor sites, mutations of single lysine residues within
Uba2 did not significantly impair its sumoylation (data not shown).

Increasing Sensitivity Using High Mass Acquisition—In earlier work, we mapped two SUMO conjugation sites within USP25 by which we identified one site (lysine 141) using a mutagenesis approach, whereas the other (lysine 99) was identified using an MS approach. It is of note that in our previous study we used a small fragment of USP25 that was conjugated with SUMO2 in bacteria followed by purification by gel filtration and anion-exchange chromatography (17). However, manual examination of full-length USP25 sumoylated in vitro did not reveal any SUMO acceptor site. To test whether our ChopNSpice method has an increased sensitivity to identify the acceptor sites of this more complex sample, we conjugated full-length USP25 with SUMO2 in vitro, using the E3 ligase PIASXα, as described previously (17). Next, the mixture was digested with trypsin in solution. Subsequently, to increase sensitivity for the identification of SUMO acceptor sites, we also used high mass MSMS acquisition conditions (Fig. 4A, compare the standard (upper panel) with the high mass (lower panel)). Under these conditions, only peptides with a mass exceeding 2154 Da (for SUMO1) or 3568 Da (for SUMO2/3) are selected (see above).

This approach is highly suitable for the accurate detection and sequencing of larger peptides and additionally facilitates detection of lower abundance SUMO conjugates (see also Fig. 4A and below). A database search against modified sequences (achieved by the program ChopNSpice in combination with MASCOT) demonstrated that sumoylated peptides were enriched by high mass MSMS acquisition (supplemental Data S3 and Table S2). Using this strategy, we went on to identify several additional SUMO acceptor sites within full-length USP25 (supplemental Data S3 and Table S2), including lysine 5, which had previously been identified only by a mutational approach. In addition, we observed lysine 141 in SUMO2 as an acceptor site for chain formation, consistent with a previous report (21).

Identification of SUMO-conjugated Sites in Vivo—Although the identification of SUMO conjugation sites in endogenous proteins from yeast has been performed before (18), unbiased identification of SUMO acceptor sites in higher eukaryotes has remained a technical challenge. This can partly be accounted for by the high mass of SUMO after hydrolysis with trypsin in higher eukaryotes combined with the low abundance of post-translational modifications per se as compared with the amount of non-modified protein. Additionally, chemical enrichment for modifications with SUMO prior to MS has not been described as is the case for instance with phosphorylation (39–41).

To examine the power of our strategy for identification of SUMO conjugation sites, we purified endogenous SUMO1 conjugates from HeLa cells (Fig. 5A). Although the overall protein composition in the immunoprecipitation of SUMO1 conjugates seems indistinguishable from the control immuno-
Identification of SUMO acceptor sites in endogenous proteins. A, SUMO1-conjugated proteins were isolated from HeLa cells using SUMO1 antibodies (Ab) coupled to protein G-agarose or control (Ctr) protein G-agarose. Immunoprecipitates were extensively washed and eluted with sample buffer. Five percent of the sample was loaded to detect SUMO1-conjugated species by Western blot; the rest of the

---

MS Spectrum

4+

M.W (calc): 3644.9115
M.W (expt): 3644.9167
Error: 0.0052Da (1.35 ppm)
XCorr: 2.90

---

Fig. 5. Identification of SUMO acceptor sites in endogenous proteins. A, SUMO1-conjugated proteins were isolated from HeLa cells using SUMO1 antibodies (Ab) coupled to protein G-agarose or control (Ctr) protein G-agarose. Immunoprecipitates were extensively washed and eluted with sample buffer. Five percent of the sample was loaded to detect SUMO1-conjugated species by Western blot; the rest of the
Identification of SUMO Conjugation Sites in Vivo

In vivo sumoylated proteins derived after immunoprecipitation from HeLa cells using anti-SUMO1 antibody (see “Experimental Procedures”) Table I

| Modification and conjugated protein | Swiss-Prot accession number | Sequence | Conjugated position | MASCOT score | SEQUEST XCorr |
|------------------------------------|-----------------------------|----------|---------------------|--------------|---------------|
| SUMO1                              |                             |          |                     |              |               |
| Ran GTPase-activating protein 1     | P46060                      | 510LLVHMGGLKSEDVKV530 | Lys524       | 70.29        | 6.10          |
| Small ubiquitin-related modifier 1  | P63165                      | 17KGEYK23 | Lys17              | 92.87        | 5.60          |
| Small ubiquitin-related modifier 2  | P61956                      | 8EGVKTNENHDHNL21 | Lys11          | 79.50        | 5.57          |
| SUMO-conjugating enzyme UBC9        | P63279                      | 142VEYERK147 | Lys146         | 59.84        | 5.18          |
| E3 SUMO-protein ligase RanBP2       | P49792                      | 1090GICKAEDVDTGRAKYVL120 | —             | 14.31        | 4.34          |
|                                    |                             | 1408RALVTPK1415 | Lys1414       | 36.53        | 4.22          |
|                                    |                             | 1775SGEGMTK1724 | Lys1723       | 106.23       | 5.46          |
|                                    |                             | 2255KNLFR2259 | Lys2255       | 35.55        | 2.86          |
|                                    |                             | 2424FLGQDVADSFK2434 | Lys2434     | 94.41        | 5.12          |
|                                    |                             | 2507AVSPPPKFVGESV2522 | Lys2522     | 29.65        | 2.59          |
|                                    |                             | 2581NSDIEQSSDSKV2594 | Lys2592     | 97.71        | 5.67          |
|                                    |                             | 2617AKEK2620 | Lys2618       | 44.60        | 6.22          |
| RanBP2-like and GRIP domain-containing protein 4 | Q7Z3J3 | 891KAEDIANDSPEEQECK709 | Lys691      | 7.29         | —             |
| Chromodomain-helicase-DNA-binding protein 8 | Q9HCK8 | 581YTEDLDIKKITDEEEEEVDVTGPK609 | Lys592 | 80.62 | —           |
| Cytoplasmic dynein 1 heavy chain 1 | Q14204                      | 3207KIKEVTVGVEELR3219 | Lys3207     | 20.56        | 4.21          |
| Very long-chain-specific acyl-CoA dehydrogenase | P49748 | 527NFKSIK530 | Lys535        | 46.13        | —             |
| Bifunctional aminoacyl-tRNA synthetase | P07814 | 31NPDKELQLMOMEIK327 | Lys318        | 51.70        | —             |
| SUMO2                              |                             |          |                     |              |               |
| Small ubiquitin-related modifier 2  | P61956                      | 8EGVKTNENHDHNL21 | Lys11         | 37.99        | 7.31          |

In this study, we present a freely available computational approach to identify post-translational modifications by mass spectrometry that cannot easily be explored by using common search engines such as SEQUEST and/or MASCOT. We demonstrate that our approach is of value in MS-based analysis and subsequent database search for the identification of SUMO conjugation sites within proteins that have been sumoylated either in vitro or in vivo.

In particular, mammalian sumoylated proteins and peptides present a challenge in MS-based detection. In contrast to yeast (S. cerevisiae), where after digestion of sumoylated proteins only an EQIGG peptide (484 Da) is conjugated to its respective SUMO acceptor, the large tryptic fragments of mammalian SUMO1 (2154 Da) and SUMO2/3 (3568 Da) are not easily identified in MS. These difficulties are in part due to the presence of long peptide conjugates, which resemble cross-linked peptides (but without cross-linker). Consequently, MS and MSMS result in fragment ion spectra that are too complex to interpret manually. To circumvent these problems, a mutational approach has been proposed to yield a smaller tryptic fragment of SUMO that simplifies the identification...
Identification of SUMO Conjugation Sites in Vivo

Although this method has proved efficient for the identification of SUMO acceptor sites from proteins sumoylated in vitro, the tailored SUMO proteins may be conjugated/deconjugated less efficiently in vivo. Another MS-based method that has been utilized to identify SUMO acceptor sites is a software tool (SUMmOn) designed to interpret the complex fragment ion pattern that allows one to work with low accuracy mass spectrometers (24). Also in this study, relatively simple in vitro conjugation mixtures were examined, whereas more complex samples from in vivo experiments are expected to cause problems in the unambiguous identification of SUMO acceptor sites. We also have used the SUMmOn pattern recognition software to identify SUMO acceptor sites in proteins sumoylated in vitro and in vivo. In fact, the analysis of our raw data with SUMmOn delivered a similar, but smaller set of sites compared with ChopNSpice in conjunction with a MASCOT-based database search (see supplemental Table S5). In addition, another software tool (Ubl finder) is available, but it suffers from the weakness that only ubiquitin and SUMO (T95R) mutants can be searched (23). By making use of highly accurate and resolving MS techniques, Matic et al. used an in vitro to in vivo approach (21). In vitro sumoylated proteins were analyzed for SUMO acceptor sites in an Orbitrap mass spectrometer and were subsequently confirmed in vivo.

We followed a different approach and combined high end MS with a commonly used database search that was slightly modified. The prerequisite for the detection of post-translational modifications per se by MS is the unambiguous identification of the site of modification within the peptide. This in turn requires MSMS sequence analyses and subsequent database searches using search engines that compare the m/z values of experimental data (i.e., the MSMS fragment spectra) with the m/z values generated in silico. In this manner, (post-translational) modifications that are attached to any amino acid can also be identified through the extra mass of the modification that is added to all the respective amino acids in the database. In a similar manner, putative ubiquitylation sites after tryptic digestion (GG diaminoo acid conjugated to its acceptor site) can be identified with available search engines. Nonetheless, even highly accurate MS analysis can lead to false positive identification when only the exact mass of the modification is taken into account. For example, it has recently been reported that iodoacetamide-induced artifacts mimic ubiquitylation in mass spectrometry (22). Thus, it is of the utmost importance, in particular when one is dealing with longer conjugates, to obtain sequence information not only from the substrate peptide but also from the modifier. However, although search engines are capable of taking experimental parameters (e.g., proteases used and modifications) into account, they rely solely on databases that contain putative protein sequences for identification and, in the case of modifications, the extra mass added to a particular amino acid. Search engines such as MASCOT and/or SEQUEST are commonly used by the proteomics researchers who use MS, and the output format of these search engines (including their scoring systems) are widely accepted in the community. To that end, we developed a software tool that makes use of these search engines and adds new modified protein sequences (sumoylated sequences) to the standard databases against which standard MS search engines can then compare and have made the new tool freely available.

The program ChopNSpice for the identification of SUMO acceptor sites is unique in its ability to allow the user (i) to combine two protein sequences in a linear manner, (ii) to generate any modified linear protein sequence that contains any modifications at the N terminus of the novel fused sequence, (iii) to introduce defined extra masses in either of the two protein sequences so that also peptide-peptide cross-links (using a cross-linking reagent) after tryptic digestion of cross-linked proteins can be searched and identified, and (iv) to generate an m/z list of all linearly fused peptides. The latter is particularly useful when users do not have access to e.g., an Orbitrap mass spectrometer but instead would like to use a simple peptide mass fingerprint analysis by MALDI MS of putatively sumoylated proteins. In addition, the list serves as an inclusion list in LC-MSMS analysis such that predicted modified (e.g., sumoylated) peptides are chosen for fragmentation within the mass spectrometer.

A similar strategy for the generation of concatenated peptides (proteins) has been discussed in conjunction with the analysis of protein-protein cross-linking MS data (33), but to date, no software is publicly available to facilitate the generation of the required FASTA library files, and Maiolica et al. (33) did not describe how the user should generate a dedicated database containing concatenated peptides. Against this background and for the first time, our approach provides a broad community with the possibility to generate every type of FASTA sequences, including various modifications that can then be used for a database search using common search engines, if required, in a high throughput approach. For the latter, entire databases (e.g., Swiss-Prot human) can be modified with ChopNSpice to generate e.g., sumoylated proteins from each entry. In addition to this feature, a number of modified databases (e.g., sumoylated Swiss-Prot human) are available via the ChopNSpice web site for added convenience.

We further show that the database search of the MSMS fragment spectra (values), including the modified linear sequence(s), is highly specific. Importantly, no hits with MASCOT or SEQUEST were obtained when the modifier sequence was reversed and attached to the C terminus of the tryptic peptides (data not shown). Moreover, a search against the human Swiss-Prot database in which all proteins were modified with SUMO1 and SUMO2/3 by ChopNSpice gave the same hits for a distinct sumoylated protein as in a search where only the protein sequence of interest was modified with ChopNSpice and submitted to the Swiss-Prot database (data...
not shown). As we aim to reach a broad proteomics community by this approach, we determined the rate of false positives in a decoy database search, finding it to be \( \leq 0.33\% \) (see supplemental Table S6), and thus demonstrate that our approach can be applied to shotgun proteomics projects. Importantly, the false positive rate remains low because of the applied proteomics work flow (see “Results”), although in some cases, we observed mainly product ions of the SUMO peptide and less of the product ions derived from the acceptor peptide (see supplemental Data S3).

In summary, here we present an approach to identify SUMO acceptor sites in endogenous proteins by mass spectrometry in a rapid and sensitive manner, and we describe examples of its successful application. We believe that this approach has the potential to be widely used mainly because (i) the necessary software for the generation of modified protein sequence (ChopNSpice) is provided, (ii) it uses established search engines for protein identification, and (iii) it facilitates the identification of sites of modification in large immunoprecipitation studies and shotgun approaches. Importantly, the idea of the generation of novel modified sequences is not restricted to ubiquitin modifiers or Ubls but can be applied to any type of (user-defined) modification.

Acknowledgments—We are grateful to Monika Raabe and Johanna Lehne for technical assistance in MS and Nicolas Stankovic for critical reading of the manuscript, and we thank all the other members of our laboratory for discussions. We are indebted to M. Matunis for the kind reading of the manuscript, and we thank all the other members of our framework program. To whom correspondence may be addressed. NET) young investigator program grant from the 6th European Union (EURAS) and Deutsche Forschungsgemeinschaft.** Supported by grants from the European Union (Rubicon Network of Excellence) and Deutsche Forschungsgemeinschaft.‡‡ To whom correspondence may be addressed: Dept. of Molecular Cell Biology, Leiden University Medical Center, P. O. Box 9600, 2300 RC Leiden, The Netherlands. Tel.: 31-71-526-9273; Fax: 31-71-526-8270; E-mail: E.Meulmeester@LUMC.nl.

Supported by a European Alternative Splicing Network (EURAS) young investigator program grant from the 6th European Union framework program. To whom correspondence may be addressed. Tel.: 49-551-201-1060/1470; Fax: 49-551-201-1197; E-mail: henningurlaub@mplbcp.mpg.de.

REFERENCES

1. Kerscher, O., Felberbaum, R., and Hochstrasser, M. (2006) Modification of proteins by ubiquitin and ubiquitin-like proteins. Annu. Rev. Cell Dev. Biol. 22, 159–180

2. Hay, R. T. (2005) SUMO: a history of modification. Mol. Cell 18, 1–12

3. Meulmeester, E., and Melchior, F. (2008) Cell biology: SUMO. Nature 452, 709–711

4. Geiss-Friedlander, R., and Melchior, F. (2007) Concepts in sumoylation: a decade on. Nat. Rev. Mol. Cell Biol. 8, 947–956

5. Hershko, A., and Ciechanover, A. (1998) The ubiquitin system. Annu. Rev. Biochem. 67, 425–479

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

7. Hay, R. T. (2007) SUMO-specific proteases: a twist in the tail. Trends Cell Biol. 17, 370–376

8. Mukhopadhyay, D., and Dasso, M. (2007) Modification in reverse: the SUMO proteases. Trends Biochem. Sci. 32, 286–295

9. Steinacher, R., and Schär, P. (2005) Functionality of human thymine DNA glycosylase requires SUMO-regulated changes in protein conformation. Eur. Biol. 15, 616–623

10. Baba, D., Maia, N., Jee, J. G., Uchimura, Y., Saitho, H., Sugasawa, K., Hanaoka, F., Tachio, H., Hiroaki, H., and Shirakawa, M. (2005) Crystal structure of thymine DNA glycosylase conjugated to SUMO-1. Nature 435, 979–982

11. Sampson, D. A., Wang, M., and Matunis, M. J. (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

12. Lin, D., Tatham, M. H., Yu, B., Kim, S., Hay, R. T., and Chen, Y. (2002) Identification of a substrate recognition site on Ubc9. J. Biol. Chem. 277, 21740–21748

13. Bernier-Villamor, V., Sampson, D. A., Matunis, M. J., and Lima, C. D. (2002) Structural basis for E2-mediated SUMO conjugation revealed by a comparison between ubiquitin-conjugating enzyme Ubc9 and RanGAP1. Cell 108, 345–356

14. Hoege, C., Pfander, B., Moldovan, G. L., Pyrowolakis, G., and Jentsch, S. (2002) RAD6-dependent DNA repair is linked to modification of PCNA by ubiquitin and SUMO. Nature 419, 135–141

15. Pichler, A., Knipscheer, P., Oberhofer, E., van Dijk, W. J., Körner, R., Olsen, J. V., Jentsch, S., Melchior, F., and Sixma, T. K. (2005) SUMO modification of the ubiquitin-conjugating enzyme E2–25K. Nat. Struct. Mol. Biol. 12, 264–269

16. Lin, D. Y., Huang, Y. S., Jeng, J. C., Kuo, H. Y., Chang, C. C., Chao, T. T., Ho, C. C., Chen, Y. C., Lin, T. P., Fang, H. I., Hung, C. C., Suen, C. S., Hwang, M. J., Chang, K. S., Mau, G. G., and Shih, H. M. (2006) Role of SUMO-interacting motif in Daxx SUMO modification, subnuclear localization, and repression of sumoylated transcription factors. Mol. Cell 24, 341–354

17. Meulmeester, E., Kunze, M., Hsiao, H. H., Urlaub, H., and Melchior, F. (2008) Mechanism and consequences for paralog-specific sumoylation of ubiquitin-specific protease 25. Mol. Cell 30, 610–619

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. Vertegaal, A. C., Andersen, J. S., Ogg, S. C., Hay, R. T., Mann, M., and Lamond, A. I. (2006) Distinct and overlapping sets of SUMO-1 and SUMO-2 target proteins revealed by quantitative proteomics. Mol. Cell Proteomics 5, 2298–2310

20. Hannich, J. T., Lewis, A., Krotetz, 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

21. Katic, I., van Hagen, M., Schimmel, J., Macek, B., Ogg, S. C., Tatham, M. H., Hay, R. T., Lamond, A. I., Mann, M., and Vertegaal, A. C. (2008) In vivo identification of human small ubiquitin-like modifier polymerization sites by high accuracy mass spectrometry and an in vitro to in vivo strategy. Mol. Cell Proteomics 7, 132–144

22. Nielsen, M. L., Vermeulen, M., Bonaldi, T., Cox, J., Moroder, L., and Mann, M. (2008) Iodoacetamide-induced artifact mimics ubiquitination in mass spectrometry. Nat. Methods 5, 459–460

23. Knesel, M., Cheung, H. T., Hamady, M., Barthel, K. K., and Liu, X. (2005) A method of mapping protein sumoylation sites by mass spectrometry using a modified small ubiquitin-like modifier 1 (SUMO-1) and a computational program. Mol. Cell Proteomics 4, 1628–1636

24. Pedrioli, P. G., Raught, B., Zhang, X. D., Rogers, R., Aitchison, J., Matunis, M., and Aebersold, R. (2006) Automated identification of SUMOylation sites using mass spectrometry and SUMOn pattern recognition software. Nat. Methods 3, 533–539

25. Perkins, D. N., Pappin, D. J., Creasy, D. M., and Cottrell, J. S. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis 20, 3551–3567

26. Jimmy K. Eng, Ashley L. McCormack, and John R. Yates, I. (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. J. Am. Soc. Mass Spectrom. 5,
27. Matunis, M. J., Coutavas, E., and Blobel, G. (1996) A novel ubiquitin-like modification modulates the partitioning of the Ran-GTPase-activating protein RanGAP1 between the cytosol and the nuclear pore complex. J. Cell Biol. 135, 1457–1470

28. Bossis, G., and Melchior, F. (2006) Regulation of SUMOylation by reversible oxidation of SUMO conjugating enzymes. Mol. Cell 21, 349–357

29. Pichler, A., Gast, A., Seeler, J. S., Dejean, A., and Melchior, F. (2002) The nucleoporin RanBP2 has SUMO1 E3 ligase activity. Cell 108, 109–120

30. Mahajan, R., Delphin, C., Guan, T., Gerace, L., and Melchior, F. (1997) A small ubiquitin-related polypeptide involved in targeting RanGAP1 to nuclear pore complex protein RanBP2. Cell 88, 97–107

31. Werner, A., Moutty, M. C., Möller, U., and Melchior, F. (2009) Performing in vitro sumoylation reactions using recombinant enzymes. Methods Mol. Biol. 497, 187–199

32. Aebersold, R., and Mann, M. (2003) Mass spectrometry-based proteomics. Nature 422, 198–207

33. Maiolica, A., Cittaro, D., Borsotti, D., Sennels, L., Ciferri, C., Tarricone, C., Musacchio, A., and Rappsilber, J. (2007) Structural analysis of multiprotein complexes by cross-linking, mass spectrometry, and database searching. Mol. Cell. Proteomics 6, 2200–2211

34. Olsen, J. V., de Godoy, L. M., Li, G., Macek, B., Mortensen, P., Pesch, R., Makarov, A., Lange, O., Horning, S., and Mann, M. (2005) Parts per million mass accuracy on an Orbitrap mass spectrometer via lock mass injection into a C-trap. Mol. Cell. Proteomics 4, 2010–2021

35. Olsen, J. V., Macek, B., Lange, O., Makarov, A., Horning, S., and Mann, M. (2007) Higher-energy C-trap dissociation for peptide modification analysis. Nat. Methods 4, 709–712

36. Mahajan, R., Gerace, L., and Melchior, F. (1998) Molecular characterization of the SUMO-1 modification of RanGAP1 and its role in nuclear envelope association. J. Cell Biol. 140, 259–270

37. Rodriguez, M. S., Desterro, J. M., Lain, S., Midgley, C. A., Lane, D. P., and Hay, R. T. (1999) SUMO-1 modification activates the transcriptional response of p53. EMBO J. 18, 6455–6461

38. Stensdorpf, T., Jensen, K., Reich, B., and Will, H. (1999) The nuclear dot protein sp100, characterization of domains necessary for dimerization, subcellular localization, and modification by small ubiquitin-like modifiers. J. Biol. Chem. 274, 12555–12566

39. Viillén, J., and Gygi, S. P. (2006) The SCX/IMAC enrichment approach for global phosphorylation analysis by mass spectrometry. Nat. Protoc. 3, 1630–1638

40. Larsen, M. R., Thingholm, T. E., Jensen, O. N., Røpstorff, P., and Jørgensen, T. J. (2005) Highly selective enrichment of phosphorylated peptides from peptide mixtures using titanium dioxide microcolumns. Mol. Cell. Proteomics 4, 873–886

41. Thingholm, T. E., Jørgensen, T. J., Jensen, O. N., and Larsen, M. R. (2006) Highly selective enrichment of phosphorylated peptides using titanium dioxide. Nat. Protoc. 1, 1929–1935

42. Wohlschlegel, J. A., Johnson, E. S., Reed, S. I., and Yates, J. R., 3rd (2006) Improved identification of SUMO attachment sites using C-terminal SUMO mutants and tailored protease digestion strategies. J. Proteome Res. 5, 761–770