

# Sequence determinants directing conversion of cysteine to formylglycine in eukaryotic sulfatases

Thomas Dierks<sup>1</sup>, M.Rita Lecca,  
Petra Schlotterhose, Bernhard Schmidt and  
Kurt von Figura

Institut für Biochemie und Molekulare Zellbiologie, Abteilung  
Biochemie II, Universität Göttingen, Gosslerstrasse 12d,  
D-37073 Göttingen, Germany

<sup>1</sup>Corresponding author  
e-mail: dierks@ukb2-00.uni-bc.gwdg.de

**Sulfatases carry at their catalytic site a unique post-translational modification, an  $\alpha$ -formylglycine residue that is essential for enzyme activity. Formylglycine is generated by oxidation of a conserved cysteine or, in some prokaryotic sulfatases, serine residue. In eukaryotes, this oxidation occurs in the endoplasmic reticulum during or shortly after import of the nascent sulfatase polypeptide. The modification of arylsulfatase A was studied *in vitro* and was found to be directed by a short linear sequence, CTPSR, starting with the cysteine to be modified. Mutational analyses showed that the cysteine, proline and arginine are the key residues within this motif, whereas formylglycine formation tolerated the individual, but not the simultaneous substitution of the threonine or serine. The CTPSR motif was transferred to a heterologous protein leading to low-efficient formylglycine formation. The efficiency reached control values when seven additional residues (AALLTGR) directly following the CTPSR motif in arylsulfatase A were present. Mutating up to four residues simultaneously within this heptamer sequence inhibited the modification only moderately. AALLTGR may, therefore, have an auxiliary function in presenting the core motif to the modifying enzyme. Within the two motifs, the key residues are fully, and other residues are highly conserved among all known members of the sulfatase family.**

**Keywords:** cysteine/endoplasmic reticulum/multiple sulfatase deficiency/protein modification/sulfatase

## Introduction

Members of the sulfatase protein family (for review, see von Figura *et al.*, 1998) are highly conserved from bacteria to man, showing strong homology at the level of their amino acid sequence (Franco *et al.*, 1995; Parenti *et al.*, 1997; Knaust *et al.*, 1998) and also at the level of their three-dimensional structure (Bond *et al.*, 1997; Lukatela *et al.*, 1998; Waldow *et al.*, 1999). The highest sequence homology is found in the N-terminal third including the unusual amino acid derivative C $\alpha$ -formylglycine (FGly) that has been found in sulfatases of prokaryotic (Dierks *et al.*, 1998a; Miech *et al.*, 1998), lower eukaryotic (Selmer *et al.*, 1996) and human origin (Schmidt *et al.*, 1995). The FGly residue

is part of the catalytic site of the enzyme, as has been shown by crystallographic analysis of two lysosomal sulfatases (Bond *et al.*, 1997; Lukatela *et al.*, 1998). The formyl group most probably is hydrated and acts as a geminal diol (Lukatela *et al.*, 1998). It directly participates in sulfate ester cleavage by a substitution/elimination mechanism. One of the hydroxyls accepts the sulfate from the substrate, leading to a covalently sulfated enzyme intermediate, and the other hydroxyl is required for subsequent sulfate elimination and regeneration of the aldehyde (Lukatela *et al.*, 1998; Recksiek *et al.*, 1998). Deficiency of FGly formation is observed in multiple sulfatase deficiency (Schmidt *et al.*, 1995), a rare lysosomal storage disorder in which all known human sulfatases are inactive (Kolodny and Fluharty, 1995).

The FGly residue is generated post-translationally by oxidation of a cysteine residue (Schmidt *et al.*, 1995; Dierks *et al.*, 1997) or, as found in one bacterial sulfatase, a serine residue (Miech *et al.*, 1998). The position of the cysteine or serine residue, which can be found in the unmodified translation product (Schmidt *et al.*, 1995; Dierks *et al.*, 1997; Miech *et al.*, 1998), is highly conserved. In eukaryotes, FGly formation occurs in the endoplasmic reticulum (ER) at a stage when the newly synthesized protein is not yet folded into its native structure (Dierks *et al.*, 1997, 1998b). Residues 58–80 of arylsulfatase A (ASA), which comprise the critical Cys69 and several other highly conserved residues (von Figura *et al.*, 1998), are obviously sufficient to direct generation of FGly69. This has been shown by analysis of truncated ASA polypeptides synthesized *in vitro* in the presence of translocation-competent microsomes (Dierks *et al.*, 1997).

In another approach to establish a sequence motif directing FGly formation, alanine/glycine mutants of ASA covering residues 68–86 were expressed in cell culture. The mutant sulfatases were purified and analyzed for the presence of FGly. However, a sequence motif could not be determined from these investigations since, apart from substitution of Cys69, other substitutions only partially impaired FGly formation (Knaust *et al.*, 1998). This led to the conclusion that the structural motif is of an extended nature, tolerating the substitution of single amino acids except for Cys69. To define this motif more closely, we took advantage of the *in vitro* translation–translocation system allowing the testing of a large number of ASA mutants for FGly formation within a reasonable time. These mutants comprised single to quadruple amino acid substitutions, including also those mutations that rendered ASA unstable when expressed in cell culture. The data obtained identified a pentameric linear sequence motif, starting with the cysteine to be modified, and a downstream element that plays a more indirect role.

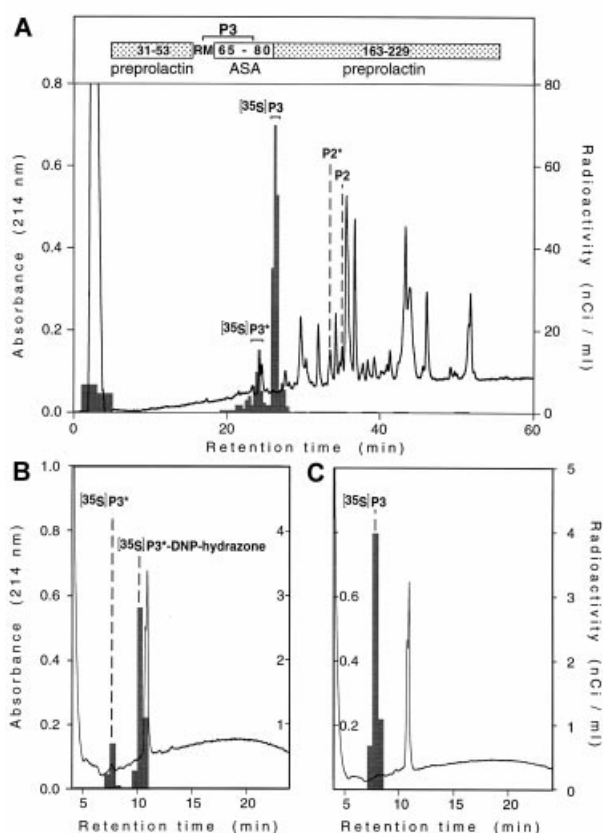
## Results

Conversion of cysteine to FGly can be studied *in vitro* using sulfatase polypeptides synthesized *de novo* in a

reticulocyte lysate in the presence of transport-competent microsomes (Dierks *et al.*, 1997, 1998b). *In vitro* expression is programmed with a cDNA encoding a fusion of the signal peptide of preprolactin followed by N-terminal fragments of mature ASA comprising the Cys69 to be modified. It has been shown previously that residues 58–80 of ASA are sufficient to determine FGly formation even after inserting this part of ASA into a heterologous protein (Dierks *et al.*, 1997). Modification is cysteine-specific and depends on the correct position of the cysteine within this sequence (Dierks *et al.*, 1997, 1998b).





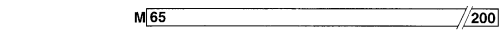
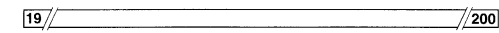


To narrow down the critical part of the ASA sequence 58–80 further, we transferred residues 65–80 of ASA to a reporter protein lacking methionines. For this purpose, residues 31–53 and 163–229 of preprolactin were fused to the N- and C-terminus, respectively, of the ASA sequence 65–80. Furthermore, the dipeptide arginine–methionine was introduced between residue 53 of preprolactin and residue 65 of ASA (see scheme in Figure 1A). The arginine generates a cleavage site for trypsin and the methionine facilitates labeling of the tryptic peptide 3 that comprises the methionine at its N-terminus and ASA residues 65–73, thus including the critical Cys69. The cDNA construct was used to program synthesis of the hybrid polypeptide *in vitro* in the presence of [<sup>35</sup>S]methionine and microsomes. The <sup>35</sup>S-labeled translation product imported into the microsomes was purified, mixed with unlabeled ASA protein, serving as a carrier, and subjected to carboxymethylation of cysteines and digestion with trypsin. The tryptic peptides were separated by reverse phase HPLC (RP-HPLC). The radioactivity was associated with two major fractions, designated as [<sup>35</sup>S]P3 and [<sup>35</sup>S]P3\* (Figure 1A). By radiosequencing, both peptides were identified as forms of the tryptic peptide 3 of the preprolactin–ASA hybrid (not shown). The unlabeled ASA added as a carrier for tryptic digestion served as an internal standard for peptide analysis. Here the carboxymethylated Cys69 and the FGly69 are part of the tryptic peptides P2 and P2\*, respectively, comprising residues 59–73 of ASA. The identification of these peptides (Figure 1A) is based on mass spectrometry, amino acid sequencing and aldehyde detection (see Materials and methods). P2 and P2\* at their N-terminus differ from their <sup>35</sup>S-labeled P3 counterparts of the preprolactin–ASA hybrid, thereby explaining the different positions in the HPLC chromatogram. However, the relative retention on the column of the modified as compared with the unmodified peptide was very similar for both peptide 2 of ASA and peptide 3 of the hybrid (Figure 1A). The two labeled fractions, [<sup>35</sup>S]P3\* and [<sup>35</sup>S]P3, were analyzed for the presence of a formyl group by reaction with the aldehyde-specific reagent dinitrophenyl hydrazine (DNP-hydrazine). Hydrazone formation was observed only for fraction [<sup>35</sup>S]P3\* but not for fraction [<sup>35</sup>S]P3 (Figure 1B and C). This indicates that the radioactivity designated as [<sup>35</sup>S]P3 represents the non-modified form of peptide 3 containing carboxymethylcysteine 69, while [<sup>35</sup>S]P3\* represents the modified, FGly69-containing form of peptide 3 (see Dierks *et al.*, 1997).

Quantitation of [<sup>35</sup>S]P3\* and [<sup>35</sup>S]P3 revealed that the modification occurred with an efficiency of 18%. This efficiency agrees well with values of 17–21% determined earlier (Dierks *et al.*, 1997) for preprolactin hybrids



**Fig. 1.** ASA residues 65–80 are sufficient for *in vitro* modification of Cys69. (A) A cDNA coding for preprolactin residues 1–53 and 163–229 joined by residues 65–80 of ASA (see scheme) was expressed *in vitro* in the presence of [<sup>35</sup>S]methionine and dog pancreas microsomes. In the hybrid protein, a dipeptide (RM) had been introduced N-terminal of ASA residue 65 in order to provide a tryptic cleavage site (R) and a methionine label (M) at the N-terminus of tryptic peptide 3 comprising ASA residues 65–73, i.e. including the critical Cys69. The translation product imported into the microsomes was purified and checked for the absence of non-imported precursor by SDS–PAGE and phosphoimaging (not shown). After mixing with unlabeled ASA carrier protein, the <sup>35</sup>S-labeled import product was subjected to reductive carboxymethylation of cysteines, digestion with trypsin and separation of tryptic peptides by RP-HPLC. In the chromatogram shown, the positions of the unmodified and modified peptide 2 of the carrier protein (P2 and P2\*, respectively) are indicated. The labeled peptides were localized by liquid scintillation counting (see bar graph) and identified by radiosequencing as derivatives of peptide 3 of the translation product (not shown). (B and C) To test for the presence of a formyl group, the two forms of labeled peptide 3, designated as [<sup>35</sup>S]P3\* and [<sup>35</sup>S]P3, were each subjected to reaction with DNP-hydrazine. The incubation mixture was loaded on an RP-HPLC column in order to separate the DNP-hydrazone product from the parent peptide and the reagent. Only [<sup>35</sup>S]P3\* gave rise to hydrazone formation, as indicated by the shift of 81% of the radioactivity to a higher retention time. [<sup>35</sup>S]P3\* thus is identified as the modified form of peptide 3, while [<sup>35</sup>S]P3 represents unmodified peptide 3 (compare B and C).

containing ASA residues 58–80 (see also Figure 2, hybrid 5) and for truncated sulfatases comprising ASA residues 19–200 or 65–200. ASA 19–200 and ASA 65–200 served as controls for the hybrid proteins harboring the FGly69-containing peptide of ASA in full-length (residues 59–73; hybrids 5–6 in Figure 2) or truncated form (residues 65–73; hybrids 1–4 in Figure 2), respectively. The truncated P3\* derived from hybrid 65–80 reacted more efficiently with DNP-hydrazine in comparison with full-

ASA-PPL-Hybrid	[ <sup>35</sup> S]P3* (% of P3+P3*)	[ <sup>35</sup> S]P3*-hydrazone (% of total P3*)	Relative modification (%)
(1) 	18.0	81	86
(2) 	9.3	63	35
(3) 	7.9	26	12
(4) 	4.4	29	8
M65 	19.5	87	100
19 	18.3	53	100
(5) 	20.6	41	87
(6) 	7.1	9	7

**Fig. 2.** ASA residues 65–80 are necessary for efficient *in vitro* modification of Cys69. *In vitro* modification efficiencies for the indicated preprolactin–ASA hybrids are expressed as the percentage of [<sup>35</sup>S]P3\* of the sum of labeled P3 plus P3\*, as determined after RP-HPLC of tryptic peptides (see Figure 1A). In addition, the fraction of [<sup>35</sup>S]P3\* that was converted to the corresponding DNP-hydrazone (see Figure 1B) is given. These two values were used to calculate relative modification efficiencies (see Materials and methods) that are given as a percentage of the control modification observed for an ASA fragment comprising either residues 65–200 (control for hybrids 1–4) or residues 19–200 (control for hybrids 5–6). All values are means of duplicate or triplicate independent determinations. For explanation of the schemes, see Figure 1. The tryptic peptide 3 of the hybrid proteins is indicated. The modified form of this peptide reacts much more efficiently with DNP-hydrazine when ASA residues 59–64 are absent (compare also the two controls), as was observed earlier (Dierks *et al.*, 1997).

length P3\* (Figure 2). The same result was observed for ASA 65–200 and may be explained by a higher electrophilia of the formyl carbon or by the absence of an intramolecular Schiff base formed with the N-terminus only in the non-truncated P3\*. Taken together, these results demonstrate that residues 65–80 are both necessary and sufficient to direct modification with the maximum efficiency attainable under the given *in vitro* conditions.

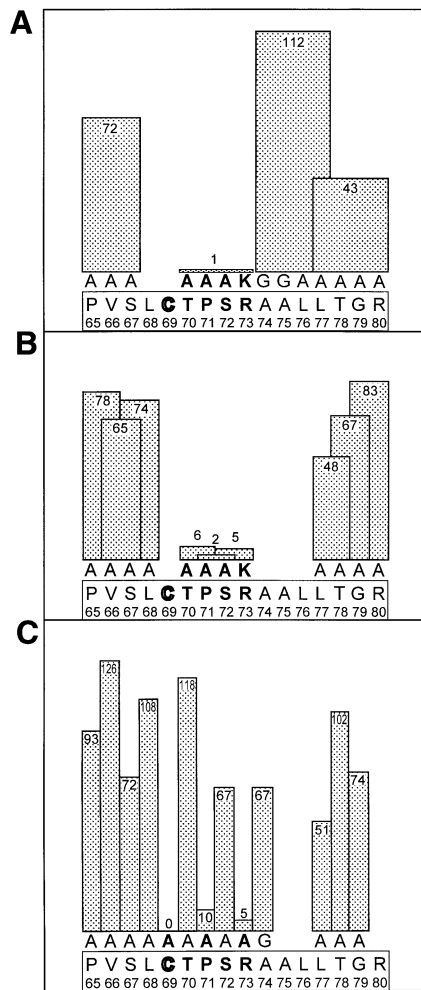
Progressive C-terminal truncations of the ASA part of the preprolactin hybrids from residue 80 down to residue 73 led to concomitant reduction of modification efficiency (Figure 2, hybrids 1–4). A low but significant modification was observed even when only residues 65–73 of ASA were transferred to the preprolactin context, which agrees with the earlier observation of a low-level modification of an ASA fragment terminating at position 73 (Dierks *et al.*, 1997). Thus, the short 65–73 sequence obviously fulfills all basic requirements needed for interaction with the modifying system. This interaction, however, is markedly improved by ASA residues 74–80, but not by unrelated sequences such as that of preprolactin. On the contrary, N-terminal extension of the ASA sequence by seven residues did not improve the modification. Not only the modification efficiencies of ASA 65–73 (Figure 2, hybrid 4) and ASA 58–73 (hybrid 6), but also those of ASA 65–80 (hybrid 1) and ASA 58–80 (hybrid 5), were comparable. This is noteworthy because the modified and unmodified forms of peptide 3 derived from constructs 5 and 6 are identical in sequence to P2\* and P2 of native ASA and could, therefore, be identified by their co-elution with the respective unlabeled ASA peptides during HPLC. This lends further support to the identification of the labeled peptides as P3\* and P3 (see Figure 1), which otherwise is based on radiosequencing and DNP-hydrazone formation (see above).

A detailed alanine/glycine scanning mutagenesis study was performed analyzing residues 65–80 of ASA in the authentic N- and C-terminal context. It has to be noted that Arg73 was mutated into a lysine in order to conserve the tryptic cleavage site at the C-terminus of peptide 2. In a first round, triple or quadruple substitutions were

introduced *en bloc* in positions 65–67, 70–73, 74–77 or 77–80. Only the modification of mutant 70–73 was abolished, whereas the modification of mutant 74–77 was normal, and modification of mutants 65–67 and 77–80 was reduced to 72 or 43%, respectively (Figure 3A). Similar results were obtained when double substitution mutants were investigated (Figure 3B). ASA constructs carrying double substitutions in the central 70–73 region, i.e. mutants 70–71, 71–72 and 72–73, were modified with very low efficiency (2–6% relative modification). Double substitution mutants in the N- or C-terminal part of the 65–80 sequence were only moderately affected, the strongest inhibition being observed for the 77–78 mutant (48% relative modification, Figure 3B).

When a series of single substitution mutants was analyzed, it was found that apart from the C69A substitution, which abolished modification completely, the modification of mutants P71A and R73A was severely inhibited (10 and 5% relative modification, respectively; Figure 3C). On the other hand, modification of mutants T70A or S72A was not affected. The importance of Cys69, Pro71 and Arg73 was studied in more detail by introducing conservative or non-conservative substitutions. It was shown that Cys69 and Pro71 did not tolerate any of the substitutions tested (see Figure 4), while Arg73 tolerated conservative substitution by a lysine (68% relative modification). Although the single mutations T70A, S72A or R73K did not lead to severe effects on modification, the importance of all three positions became apparent when two of them were substituted simultaneously. All double mutants tested, namely S72A/R73K, S72T/R73K and T70A/S72A, showed very low modification efficiencies (5, 11 and 18%, respectively). In conclusion, the core motif critical for FGly formation comprises residues 69–73 (CTPSR) with the key residues Cys69, Pro71 and Arg73.

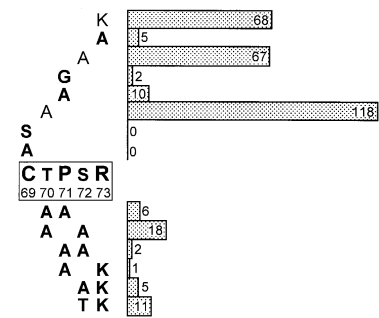
If the modifying enzyme recognizes a linear sequence motif, it should be possible to inhibit FGly formation using related peptides, which are added to the *in vitro* system and which are taken up by the microsomes via the TAP transporters (Androlewicz *et al.*, 1993; Neeffjes *et al.*, 1993; Shepherd *et al.*, 1993). *In vitro* modification of



**Fig. 3.** Effect of multiple or single substitution mutations in positions 65–80 on modification efficiency. In an ASA fragment comprising residues 19–200, three to four (A), two (B) or single residues (C) in positions 65–80 were substituted by alanine, glycine or lysine, as indicated. The mutant proteins carrying the preprolactin signal peptide were expressed *in vitro* in the presence of microsomes and analyzed for modification of Cys69 as described in Figure 1. Relative modification efficiencies (see Figure 2) are given as a percentage of the wild-type ASA 19–200 control and are means of at least two independent determinations. All triple and quadruple mutations (A) and the single mutations C69A, P71A and R73A (C) were also transferred to ASA fragments comprising either residues 19–80 or 65–200. The modification efficiencies determined for these additional mutant constructs confirmed the values given in the figure, thereby excluding unspecific long-distance interactions of N- or C-terminal regions with the mutated sequence as being responsible for the effects observed.

ASA 19–200 and ASA 65–200 was studied in the presence of various synthetic peptides. Peptide 65–80, with an amino acid sequence corresponding to ASA residues 65–80, strongly inhibited FGly formation. This inhibition was dependent on the peptide concentration (not shown), reaching ~98% inhibition at a concentration of 0.35 mM (Figure 5). Other peptides of similar size and composition but unrelated to ASA (see Materials and methods) were not inhibitory.

The experiment was repeated with fragments of peptide 65–80 comprising residues 65–73, 74–80 or 68–80. Of these, only peptide 68–80 showed strong inhibition, whereas even the simultaneous presence of peptides 65–73



**Fig. 4.** Effect of conservative and non-conservative substitutions in positions 69–73 on modification efficiency. In an ASA fragment comprising residues 19–200, one (top panel) or two residues (bottom panel) in positions 69–73 were mutated as indicated. For determination and calculation of relative modification efficiencies, see Figure 3. The values given are means of at least two independent determinations.

Peptide name	Peptide sequence	Relative modification (% of control)
65-80	PVSLCTPSRAALLTGR	2
65-73	PVSLCTPSR	127
74-80	AALLTGR	80
65-73 + 74-80	PVSLCTPSRAALLTGR	83
68-80	LC TPSRAALLTGR	9
non-ASA 17mer	ADG CDFVCRSKPRNVPA	82

**Fig. 5.** Inhibition of ASA modification by synthetic peptides. An ASA fragment comprising residues 19–200 and carrying the preprolactin signal peptide was expressed *in vitro* in the presence of microsomes and various synthetic peptides at a concentration of 0.35 mM. The peptides were added 5 min prior to the start of translation. The peptide sequences are given and correspond to ASA residues 65–80, 65–73, 74–80 or 68–80, as indicated by the hatched area. The non-ASA 17mer contains cysteines, prolines and arginines, but is unrelated to ASA. The relative modification efficiencies are given as a percentage of the modification determined in the absence of any peptide.

and 74–80 did not produce significant inhibition (Figure 5). Thus, in order to compete with the polypeptide substrate for binding by the modifying enzyme, the inhibitory peptide has to comprise residues 68–80 of ASA. This finding agrees with results obtained by mutational studies, in which an ASA fragment terminating at position 80 (ASA 19–80) was fully modified, whereas modification of an ASA 19–73 fragment was rather inefficient (Dierks *et al.*, 1997). As also shown by the preprolactin–ASA hybrids (Figure 2), the modifying enzyme needs a contiguous linear sequence extending beyond Arg73 in order to recognize and bind to the CTPSR motif efficiently.

## Discussion

The sequence information directing conversion of cysteine to FGly in sulfatases is confined to residues –1 to +11 with respect to the position of the cysteine to be modified. This became evident by three different approaches. Formation of FGly69 was observed with the same efficiency as that of the control after transferring residues 65–80 of ASA to a heterologous protein. Secondly, the modification could be inhibited by a synthetic peptide equivalent to ASA residues 68–80. Finally, even triple alanine substitutions of residues 65–67 or deleting the entire region N-terminal of

			Length (residues)	Signal peptide	FGly	Accession No.
<b>A Human sulfatases:</b>						
Arylsulfatase A	C T P S R A A L L T G R	(Pos. 69-80)	507	+	+	X52151
Arylsulfatase B	C T P S R S Q L L T G R	(Pos. 91-102)	533	+	+	J05225
Arylsulfatase C (Steroid sulfatase)	C T P S R A A F M T G R	(Pos. 83-94)	583	+	n.d.	J04964
Arylsulfatase D	C T P S R A A F L T G R	(Pos. 89-100)	593	+	n.d.	X83572
Arylsulfatase E	C T P S R A A F L T G R	(Pos. 86-97)	589	+	n.d.	X83573
Arylsulfatase F	C S P S R S A F L T G R	(Pos. 79-90)	591	+	n.d.	X97868
N-Acetylgalactosamine 6-sulfatase	C S P S R A A L L T G R	(Pos. 79-90)	522	+	n.d.	U06088
N-Acetylglucosamine 6-sulfatase	C C P S R A S I L T G K	(Pos. 91-102)	552	+	n.d.	Z12173
Iduronate sulfatase	C A P S R V S F L T G R	(Pos. 84-95)	550	+	n.d.	M58342
Sulfamidase	C S P S R A S L L T G L	(Pos. 70-81)	502	+	n.d.	U30894
<b>B Lower eukaryotic sulfatases:</b>						
<i>Hemicentrotus pulcherrimus</i>	C T P S R S A I M T G R	(Pos. 100-111)	551	+	n.d.	X16679
<i>Strongylocentrotus purpuratus</i>	C T P S R S A I V T G R	(Pos. 115-126)	567	+	n.d.	M28404
<i>Heliocidaris erythrogramma</i>	C T P S R S A I M T G R	(Pos. 106-117)	559	+	n.d.	AF013158
<i>Volvox carteri</i>	C C P S R T N L W R G Q	(Pos. 72-83)	649	+	+	X77214
<i>Chlamydomonas reinhardtii</i>	C C P S R T N L C A A S	(Pos. 73-84)	646	+	n.d.	X16180
<i>Neurospora crassa</i>	C C P A R V S L W T G K	(Pos. 89-100)	639	+	n.d.	U89492
<b>C Prokaryotic sulfatases:</b>						
<i>Klebsiella pneumoniae</i>	S A P A R S M L L T G N	(Pos. 72-83)	577	+	+	M31938
<i>Pseudomonas aeruginosa</i>	C S P T R S M L L T G T	(Pos. 51-62)	533	-	+	Z48540
<i>Sinorhizobium meliloti</i>	C A P A R A S F M A G Q	(Pos. 54-65)	512	-	n.d.	U39940
<i>Burkholderia caryophylli</i> (phosphonate monoester hydrolase)	C G P A R A S L L T G L	(Pos. 57-68)	514	-	n.d.	U44852

**Fig. 6.** The motif directing FGly formation *in vitro* is highly conserved among all members of the sulfatase family. Sequences homologous to residues 69–80 of ASA are shown for all biochemically characterized sulfatases, except for seven mammalian sulfatases, which in the given sequence are identical or almost identical to their human counterparts. The CXPSR motif is found in all human/mammalian sulfatases (A), in five out of six lower eukaryotic sulfatases (B) and also in two potential sulfatases of *Caenorhabditis elegans* (accession nos U53180 and U43375). The key residues within this motif (CXPXR) are fully conserved in all sulfatases including the prokaryotic ‘cysteine-type’ sulfatases of *Pseudomonas aeruginosa* and *Sinorhizobium meliloti* (C), *Escherichia coli* and *Mycobacterium tuberculosis* (Schirmer and Kolter, 1998) and also in the phosphonate monoester hydrolase of *Burkholderia caryophylli* (C). In the ‘serine-type’ sulfatase of *Klebsiella pneumoniae*, the FGly is generated by modification of a serine (Miech *et al.*, 1998). This sulfatase (C), like the other two known ‘serine-type’ sulfatase sequences of *E.coli* (Schirmer and Kolter, 1998), also carries the proline and arginine (SXPXR). The auxiliary element (see text) directly following the key residues comprises a LTGR tetrapeptide, out of which at least three residues (LTG or TGR) are present in 16 of the 20 depicted sequences (A–C). All sequences shown are located within the N-terminal fifth of the respective proteins, which mostly comprise 500–600 amino acid residues, as indicated. The length of the *K.pneumoniae* sulfatase was derived from our own DNA sequencing (Szameit *et al.*, 1999). All eukaryotic sulfatases or hypothetical sulfatases carry a signal peptide, as predicted by the von Heijne method (Nielsen *et al.*, 1997), whereas in bacteria only the ‘serine-type’ sulfatases are secretory proteins. Conversion of the first of the shown residues to FGly was demonstrated for five different sulfatases, as indicated (n.d., not determined). In addition, the database accession numbers are given (for references, see text; Selmer *et al.*, 1996; von Figura *et al.*, 1998).

residue 65 or C-terminal of residue 80 (Dierks *et al.*, 1997) did not affect FGly generation significantly.

Within the –1 to +11 region, two sequence motifs directing FGly formation were identified by a mutational analysis. The core motif, which is absolutely required, comprises the pentapeptide CTPSR starting with the cysteine to be modified. The most crucial residues besides Cys69 are Pro71 and Arg73. Substitution of the proline by alanine or glycine and of the arginine by alanine, but not by lysine, nearly abolished cysteine modification *in vitro*. In addition, Thr70 and Ser72 contribute to the sequence motif. While each of the two residues could be changed individually into alanine without reducing FGly generation, a conclusion also drawn from *in vivo* studies (Knaust *et al.*, 1998), the simultaneous substitution of both residues or of one of them in combination with the ‘silent’ R73K mutation severely reduced the modification efficiency. Such a synthetic phenotype was observed in all combinations of two single mutations introduced into positions 70–73 (Figure 4). Thus, all five residues 69–73 (CTPSR) could be identified as constituents of the

sequence motif searched for. In line with this conclusion, the three key residues cysteine, proline and arginine are totally conserved in all eukaryotic sulfatases. Ser72 is substituted only in the *Neurospora crassa* sulfatase by alanine (Figure 6A and B). The position of Thr70 shows limited variability, being occupied by either threonine (13 out of 23 sulfatases), serine (3/23), cysteine (5/23) or alanine (2/23).

A second ‘auxiliary’ motif is located directly C-terminal of the core CTPSR motif and comprises a stretch of seven amino acids (AALLTGR, residues 74–80), which include the highly conserved tetrapeptide LTGR (Figure 6). This second motif is likely to assist in the presentation of the cysteine to the modifying enzyme, as became most obvious in the peptide inhibition experiments showing that peptide 65–80, but not a mixture of peptides 65–73 and 74–80, competed efficiently with the *in vitro* substrate (Figure 5). Accordingly, efficient modification of ASA (Dierks *et al.*, 1997) or of the ASA–preprolactin hybrids (Figure 2) was observed only when residues 74–80 were present. Obviously the modifying enzyme has to get a hold on the

substrate polypeptide, which is not possible when residues 74–80 are deleted or substituted by the unrelated preprolactin sequence IFGQVIP. In the mutational analysis (Figures 3 and 4), we observed an ~50% inhibition of FGly formation when residues 77–80, 77–78 or even 77 alone were substituted by alanines (Figure 3). On the contrary, mutating residues 74–77 did not affect the modification. Furthermore, Leu77 and also the even more highly conserved Thr78 or Gly79 could be substituted individually by alanine without clear inhibition of ASA modification after expression in fibroblasts (Knaust *et al.*, 1998). Taken together, the auxiliary effect of residues 74–80 on modification, as observed *in vitro*, seems to rely not on the primary structure but on the secondary structure, which this stretch of amino acids adopts shortly after synthesis and translocation of the N-terminal 120 residues, where this sequence is located in all known sulfatases (Figure 6).

In a previous study (Knaust *et al.*, 1998), a large series of single alanine/glycine substitution mutants covering ASA residues 68–86 were expressed in fibroblasts. Thirteen of these mutant sulfatases could be analyzed for the presence of FGly, while others, among them R73A, were unstable, thus preventing further analysis. Interestingly, apart from Cys69, no other residue was essential for the modification to occur. The strongest reduction of modification was observed for the P71A mutant (35% of the wild-type control), which basically agrees with the present *in vitro* data (see above). Substitution of Leu68 and Ala74 was associated with a partial reduction of FGly formation (Knaust *et al.*, 1998). This reduction was not observed in the present study and suggests that it was caused *in vivo* by a high expression, which can saturate the modification machinery (Schmidt *et al.*, 1995). Taken together, the findings of the earlier *in vivo* study and the present *in vitro* study agree with each other and show that apart from the critical cysteine, no further residue is strictly essential. In spite of this, several residues within the two motifs are highly conserved (Figure 6). At least for some of them this must be explained by the role these residues fulfill in the native molecule, where the FGly has to be positioned in the active site at the top of an  $\alpha$ -helix formed by residues 69–78 (Lukatela *et al.*, 1998; Waldow *et al.*, 1999).

FGly formation in eukaryotic sulfatases occurs in the ER either during or shortly after protein translocation (Dierks *et al.*, 1997). The modifying enzyme acts on the polypeptide while the latter is at least partially unfolded, as concluded from studies of translocation intermediates (Dierks *et al.*, 1997) and of ASA–preprolactin chimeras, which cannot be assumed to fold into any ASA-like structure. The finding that a linear amino acid sequence is the core motif directing FGly formation agrees with this view. Occlusion of this motif within higher ordered structures can be prevented if modification occurs already co-translationally and/or if another component interacts with the polypeptide in a chaperone-like manner. Here the auxiliary sequence element (see above) may function as a ‘pausing’ sequence leading to arrest of translocation and/or folding. Such a model is also supported by the previous observation that signal peptide cleavage is delayed in nascent ASA polypeptide chains (Dierks *et al.*, 1997).

The CXPSRXXXL/MTG sequence is found in all

human (Figure 6A) and other mammalian sulfatases, but not in any non-sulfatase protein or GenBank translation. The core CXPSR motif is present in 82 non-redundant protein sequences deposited in the Swissprot and PIR databases. By restricting position X to threonine, serine, cysteine or alanine, as is found in all 23 known eukaryotic sulfatases (see above), four prokaryotic and 12 eukaryotic non-sulfatase sequences are also retrieved. Among the latter, only two carry a signal peptide directing translocation into the ER. One is the T cell receptor  $\delta$  chain with a CAPSR segment located in a hypervariable region (Yang *et al.*, 1995). The other protein is the insulin-like peptide  $\beta$ -type 3 of *Caenorhabditis elegans*, which consists of 107 amino acids terminating with CCPSR. Modification directed by this sequence is highly unlikely due to the absence of auxiliary downstream sequences (see above), and in fact does not occur, as indicated by the disulfide bond formation of the two cysteines (Duret *et al.*, 1998). In conclusion, the C-T/S/C/A-P-S-R motif is highly specific for eukaryotic sulfatases.

The key residues within this motif, CXPXR (or SXPXR for the prokaryotic sulfatases undergoing FGly formation from a serine; see Introduction), are fully conserved in all known or potential members of the sulfatase family. These include the biochemically characterized pro- and eukaryotic sulfatases shown in Figure 6 and also 10 hypothetical sulfatases (see legend to Figure 6). Another CXPXR protein from the eubacterium *Burkholderia caryophylli* (Figure 6C) was characterized as a phosphonate monoester hydrolase (Dotson *et al.*, 1996). However, since this protein shows 83% identity to human sulfamidase in the stretch of amino acids given in Figure 6 and an overall identity of up to 24% to human sulfatases, it is a good candidate for a protein undergoing FGly modification and may well exhibit sulfatase activity.

## Materials and methods

### Site-directed mutagenesis and truncations

Mutagenesis of plasmid pTD3 (Dierks *et al.*, 1997) coding for a fusion of the signal peptide of preprolactin followed by ASA-F59M residues 19–200 was carried out by PCR methods using appropriate primers. Subcloning of the PCR products was achieved using, on the one hand, the *Bsa*AI, *Eag*I, *Sma*I or *Bst*EII site in the coding region of ASA and, on the other hand, the 5' *Hind*III or the 3' *Eco*RI site of the polylinker of the vector. Substitution of methionine codons 85, 87 and 120 by threonine (position 85) or leucine codons (positions 87 and 120) gave rise to pTD17. To fuse the signal peptide of preprolactin followed by the tetrapeptide TPDM to residues 65–200 of ASA, an oligonucleotide (ACTCCGGACATG) comprising a *Bsp*EI site was added 5' of codon 65 and the *Bsp*EI–*Eco*RI fragment was cloned back into pTD17 in-frame with the signal peptide, thereby yielding pTD31. The added methionine in the TPDM tetrapptide is referred to as Met64. All relevant mutations (see Figures 3 and 4) were transferred from the pTD17 background to pTD31 and also to pTD21, which encoded the preprolactin signal peptide fused to residues 19–80 of ASA-F59M due to a stop codon that was added 3' of codon 80. All mutations and constructs were analyzed by DNA sequencing in order to preclude any PCR-derived errors.

### Construction of preprolactin–ASA hybrids

To fuse residues 1–53 of preprolactin-V33D, encoded by pTD 1 (Dierks *et al.*, 1997), to residues 58–200 of ASA, an oligonucleotide (AGACACAGGCACGTAGAAGTCTGTCCG) comprising a *Bsa*AI site and coding for ASA-F59M residues 58–63 was added 3' of codon 53 of preprolactin. The PCR product was cloned as a *Hind*III–*Bsa*AI fragment into pTD17, thereby yielding pTD48. To fuse residues 163–229 of preprolactin to residues 58–80 of ASA, an oligonucleotide

(CTCCTGACCGGCCGG) comprising an *EagI* site and coding for ASA residues 76–80 was added 5' of codon 163 of pTD1. The PCR product was cloned as an *EagI*–*EcoRI* fragment into pTD48, thereby yielding pTD49. For replacing in this pTD49-encoded construct residues 58–64 of ASA by the dipeptide RM, which provides a tryptic cleavage site (R) and a residue analogous to Met64 (see above), the QuikChange method (Stratagene) was applied according to the instructions of the manufacturer; the coding sequence of the complementary primers was CCTGTTTGACCGGGCATCTAGAATGCTGTGTCTCTGTGC. The ASA part (residues 65–80) of the resulting hybrid, encoded by pTD57, was trimmed C-terminally by adding 3' to ASA codons 73, 74 or 76 an oligonucleotide that encoded preprolactin residues 133–137 and comprised an *MseI* site, which allowed fusion of the respective ASA codon to preprolactin codon 133.

### Protein expression and purification

ASA-F59M, which served as a carrier protein during peptide analysis (see below), was expressed in mouse embryonic fibroblasts deficient in both mannose-6-phosphate receptors, as described previously (Dierks *et al.*, 1997). The expressed ASA-F59M protein was purified from the secretions of the cells by affinity chromatography (Sommerlade *et al.*, 1994). It contained mainly an FGly residue in position 69. Due to overexpression, a minor fraction was unmodified, carrying Cys69 (Schmidt *et al.*, 1995).

*In vitro* synthesis of ASA-derived proteins was carried out in a coupled transcription–translation system (TNT, Promega) as described (Dierks *et al.*, 1997). Rough microsomes from dog pancreas were added at 7.5 equivalents (Walter and Blobel, 1983) per 50  $\mu$ l of translation mixture. Purification of translation products imported by the microsomes using differential centrifugation and proteinase K digestion also was described earlier (Dierks *et al.*, 1997). Aliquots (4%) were analyzed by SDS–PAGE on high-Tris gels (Dierks *et al.*, 1996) and phosphoimaging. The remaining 96% were used for peptide analysis.

### Inhibition by synthetic peptides

Peptides corresponding to ASA residues 65–80 (PVSCLTPSRAALL-TGR) or 68–80, and the cysteine-containing non-ASA peptides ADG-CDFVCRSKPRNVPA (Figure 5) and AFWQDLGNLVDGCD were synthesized on a 9050 peptide synthesizer (Millipore) using amino acids protected with 1-fluorenylmethoxycarbonyl (Fmoc) groups and activated with benzotriazol-1-yl-oxy-tris(pyrrolidino)-phosphonium hexafluorophosphate (ByBOP). After cleavage from the resin and the protecting groups, the peptides were purified by RP-HPLC on a Delta Pac C-18 column (Millipore). Purity was confirmed by analytical RP-HPLC, UV spectrometry and mass spectrometry. All peptides were quantitated by amino acid sequencing. Peptides corresponding to ASA residues 65–73 or 74–80 were generated by digestion of peptide 65–80 with trypsin (sequencing grade, Boehringer Mannheim) at 1% (w/w) of total peptide, which cleaves C-terminal of Arg73. The two fragments were purified by RP-HPLC and checked for purity by mass spectrometry and amino acid sequencing. The peptides were added to the reticulocyte lysate at a concentration of 0.35 mM and incubated with the microsomes for 5 min at 30°C prior to addition of the template DNA.

### Peptide analysis

Purified *in vitro* translation/translocation products were mixed with 30–40  $\mu$ g of unlabeled ASA-F59M carrier protein, which served as an internal standard, and subjected to peptide analysis. Reductive carboxymethylation and generation of tryptic peptides were carried out as described (Dierks *et al.*, 1997). Separation of tryptic peptides by RP-HPLC, mass spectrometry and sequencing of peptides also have been described earlier (Schmidt *et al.*, 1995). [<sup>35</sup>S]P (P2 or P3) and [<sup>35</sup>S]P\* (P2\* or P3\*) were identified by radiosequencing and, if possible (see Results), by their co-elution with the corresponding unlabeled peptides deriving from the carrier protein. To assay for the presence of an aldehyde group, purified [<sup>35</sup>S]P\* was subjected to reaction with DNP-hydrazine (see Dierks *et al.*, 1997). Unreacted P\* and its DNP-hydrazine derivative were separated by RP-HPLC on a C2/C18- $\mu$ Peak column (Pharmacia). The labeled peptides and their derivatives were quantitated by liquid scintillation counting. Relative modification efficiencies were calculated [P\*(% of total P) × P\*-hydrazine (% of total P\*)] and expressed as a percentage of the control given in each figure legend.

## Acknowledgements

We thank A.Voigt and J.Fey for assistance in cloning and peptide analysis, respectively, and K.Neifer for peptide synthesis and sequencing.

We also appreciate the help of E.Hartmann in database screening, and the advice of J.Neeffes (Amsterdam) regarding peptide inhibition. This work was supported by the Deutsche Forschungsgemeinschaft and the Fonds der Chemischen Industrie.

## References

- Androlewicz,M.J., Anderson,K.S. and Cresswell,P. (1993) Evidence that transporters associated with antigen processing translocate a major histocompatibility complex class I-binding peptide into the endoplasmic reticulum in an ATP-dependent manner. *Proc. Natl Acad. Sci. USA*, **90**, 9130–9134.
- Bond,C.S., Clements,P.R., Ashby,S.J., Collyer,C.A., Harrop,S.J., Hopwood,J.J. and Guss,J.M. (1997) Structure of a human lysosomal sulfatase. *Structure*, **5**, 277–289.
- Dierks,T. *et al.* (1996) A microsomal ATP-binding protein involved in efficient protein transport into the mammalian endoplasmic reticulum. *EMBO J.*, **15**, 6931–6942.
- Dierks,T., Schmidt,B. and von Figura,K. (1997) Conversion of cysteine to formylglycine: a protein modification in the endoplasmic reticulum. *Proc. Natl Acad. Sci. USA*, **94**, 11963–11968.
- Dierks,T., Miech,C., Hummerjohann,J., Schmidt,B., Kertesz,M.A. and von Figura,K. (1998a) Formation of formylglycine in prokaryotic sulfatases by oxidation of either cysteine or serine. *J. Biol. Chem.*, **273**, 25560–25564.
- Dierks,T., Lecca,M.R., Schmidt,B. and von Figura,K. (1998b) Conversion of cysteine to formylglycine in eukaryotic sulfatases occurs by a common mechanism in the endoplasmic reticulum. *FEBS Lett.*, **423**, 61–65.
- Dotson,S.B., Smith,C.E., Ling,C.S., Barry,G.F. and Kishore,G.M. (1996) Identification, characterization, and cloning of a phosphonate monoester hydrolase from *Burkholderia caryophylli* PG2982. *J. Biol. Chem.*, **271**, 25754–25761.
- Duret,L., Guex,N., Peitsch,M.C. and Bairoch,A. (1998) New insulin-like proteins with atypical disulfide bond pattern characterized in *Caenorhabditis elegans* by comparative sequence analysis and homology modeling. *Genome Res.*, **8**, 348–353.
- Franco,B. *et al.* (1995) A cluster of sulfatase genes on Xp22.3: mutations in chondrodysplasia punctata (CDPX) and implications for warfarin embryopathy. *Cell*, **81**, 15–25.
- Knaust,A., Schmidt,B., Dierks,T., von Bulow,R. and von Figura,K. (1998) Residues critical for formylglycine formation and/or catalytic activity of arylsulfatase A. *Biochemistry*, **37**, 13941–13946.
- Kolodny,E.H. and Fluharty,A.L. (1995) Metachromatic leukodystrophy and multiple sulfatase deficiency: sulfatide lipidosis. In Scriver,C.R., Beaudet,A.L., Sly,W.S. and Valle,D. (eds), *The Metabolic and Molecular Bases of Inherited Disease*. McGraw-Hill, New York, pp. 2693–2741.
- Lukatela,G., Krauss,N., Theis,K., Selmer,T., Gieselmann,V., von Figura,K. and Saenger,W. (1998) Crystal structure of human arylsulfatase A: the aldehyde function and the metal ion at the active site suggest a novel mechanism for sulfate ester hydrolysis. *Biochemistry*, **37**, 3654–3664.
- Miech,C., Dierks,T., Selmer,T., von Figura,K. and Schmidt,B. (1998) Arylsulfatase from *Klebsiella pneumoniae* carries a formylglycine generated from a serine. *J. Biol. Chem.*, **273**, 4835–4837.
- Neeffes,J.J., Momburg,F. and Hammerling,G.J. (1993) Selective and ATP-dependent translocation of peptides by the MHC-encoded transporter. *Science*, **261**, 769–771.
- Nielsen,H., Brunak,S. and von Heijne,G. (1997) Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Engng*, **10**, 1–6.
- Parenti,G., Meroni,G. and Ballabio,A. (1997) The sulfatase gene family. *Curr. Opin. Genet. Dev.*, **7**, 386–391.
- Recksiek,M., Selmer,T., Dierks,T., Schmidt,B. and von Figura,K. (1998) Sulfatases: trapping of the sulfated enzyme intermediate by substituting the active site formylglycine. *J. Biol. Chem.*, **273**, 6096–6103.
- Schirmer,A. and Kolter,R. (1998) Computational analysis of bacterial sulfatases and their modifying enzymes. *Chem. Biol.*, **5**, R181–R186.
- Schmidt,B., Selmer,T., Ingendoh,A. and von Figura,K. (1995) A novel amino acid modification in sulfatases that is defective in multiple sulfatase deficiency. *Cell*, **82**, 271–278.
- Selmer,T., Hallmann,A., Schmidt,B., Sumper,M. and von Figura,K. (1996) The evolutionary conservation of a novel protein modification, the conversion of cysteine to serinesialdehyde in arylsulfatase from *Volvox carteri*. *Eur. J. Biochem.*, **238**, 341–345.

- Shepherd,J.C., Schumacher,T.N., Ashton-Rickhardt,P.G., Imaeda,S., Ploegh,H.L., Janeway,C.A.,Jr and Tonegawa,S. (1993) TAP1-dependent peptide translocation *in vitro* is ATP dependent and peptide selective. *Cell*, **74**, 577–584.
- Sommerlade,H.J., Selmer,T., Ingendoh,A., Gieselmann,V., von Figura,K., Neifer,K. and Schmidt,B. (1994) Glycosylation and phosphorylation of arylsulfatase A. *J. Biol. Chem.*, **269**, 20977–20981.
- Szameit,C., Miech,C., Balleininger,M., Schmidt,B., von Figura,K. and Dierks,T. (1999) The iron sulfur protein AtsB is required for post-translational formation of formylglycine in the *Klebsiella* sulfatase. *J. Biol. Chem.*, **274**, in press.
- von Figura,K., Schmidt,B., Selmer,T. and Dierks,T. (1998) A novel protein modification generating an aldehyde group in sulfatases: its role in catalysis and disease. *BioEssays*, **20**, 505–510.
- Waldow,A., Schmidt,B., Dierks,T., von Bulow,R. and von Figura,K. (1999) Amino acid residues forming the active site of arylsulfatase A: role in catalytic activity and substrate binding. *J. Biol. Chem.*, **274**, in press.
- Walter,P. and Blobel,G. (1983) Preparation of microsomal membranes for cotranslational protein translocation. *Methods Enzymol.*, **96**, 84–93.
- Yang,Y.G., Ohta,S., Yamada,S., Shimizu,M. and Takagaki,Y. (1995) Diversity of T cell receptor  $\delta$ -chain cDNA in the thymus of a one-month-old pig. *J. Immunol.*, **155**, 1981–1993.

Received December 10, 1998; revised and accepted February 18, 1999