1. Introduction

Multiple sulfatase deficiency (MSD) is a rare but fatal, congenital metabolic disorder that belongs to the group of lysosomal storage diseases, characterized by a drastic reduction in the activity of the sulfatase enzyme family. The elucidation of the molecular basis of this reduced activity was achieved by the detection of a hitherto unknown posttranslational amino acid modification of all sulfatases. The non-proteinogenic amino acid formylglycine (FGly) was detected in the active site of multiple sulfatases (Schmidt *et al*., 1995) and was shown to be essential for the catalytic activity of sulfatases. After decades of medical and molecular biological research to elucidate the molecular pathobiochemistry of MSD, the formylglycine-generating enzyme (FGE) was discovered in 2003 by two independent biochemical and genetic approaches (Dierks *et al*., 2003; Cosma *et al*., 2003). These seminal studies established FGE as the enzyme responsible for activation of sulfatases. FGE is a unique protein with novel biochemical and structural properties. Elucidation of the X-ray crystal structure of FGE (Dierks *et al*., 2005) revealed a unique structure (termed FGE fold), which laid the basis for the proposal that FGE could function as a monooxygenase using a novel redox-dependent catalytic mechanism. Moreover, secreted FGE enters neighbouring cells via an unknown endocytic route, to perform its paracrine function (Zito *et al*., 2007). The overall function and regulation of FGE are ill defined and yet to be characterized.

1.1 Sulfatases and Single Sulfatase Deficiency

Sulfatases represent a large family of hydrolytic enzymes essential for the hydrolysis, degradation and remodeling of sulfate esters. Sulfatase enzymes are found in prokaryotes and eukaryotes, but have so far not been detected in plants (Landgrebe *et al*., 2003). Sulfatase substrates include a range from small cytosolic steroids, such as estrogen sulfate, to complex cell surface carbohydrates such as the glycosaminoglycans. In mammals, sulfatases are involved in the hydrolysis of various sulfated substrates such as glycosaminoglycans (heparin, heparin sulfate, chondroitin sulfate, keratan sulfate), steroid hormones (e.g dehydroepiandrosteron 3-sulfate) and sulfolipids (e.g cerebroside 3-sulfate) (Hopwood *et al*., 2001, Diez-Rox *et al*., 2005). The transformation of these molecules has been linked with important cellular functions including hormone regulation, cellular degradation and modulation of signaling pathways (Lamanna *et al*., 2008, Wang *et al*., 2008). Biochemically, thirteen out of 17 sulfatases encoded in the human genome have been characterized (Sardiello *et al*., 2005). According to their subcellular localization they can be categorized into lysosomal and non-lysosomal enzymes. Most of the sulfatases are localized in lysosomes and act at acidic pH (Hanson *et al*., 2004). Non-lysosomal sulfatases such as Sulf1 and Sulf2 are present at cell surface, others like arylsulfatses C, D and F are present in endoplasmic reticulum whereas arylsulfatase E is present in Golgi apparatus. Non-lysosomal enzymes act at neutral pH. The eight currently known and described metabolic disorders caused by low sulfatase activities

clearly show the importance of sulfatases in cell homeostasis and specificity of their activities. In these monogenic diseases, the lack of specific desulfation of individual substrates leads to the accumulation of metabolites in the lysosomes and extracellular fluids of affected patients (Ballabio and Gieselmann, 2009). Reduced sulfatase activity is linked either to severe lysosomal storage disorders such as mucopolysaccharidoses (MPS) and metachromatic leukodystrophy (MLD) or to non-lysosomal disorders such as X-linked ichthyosis and chondrodysplasia punctata (Hopwood and Ballabio, 2001). The accumulation of metabolic intermediates resulting in loss of function of lysosomes as in autophagy, leads to apoptosis and causes an immune response in the affected individuals (Settembre *et al*., 2008b). The onsets of these disease conditions appear frequently after birth or in early childhood, but some are manifested only in adolescence or in adults. Many sulfatase defects show the clinical symptoms of mucopolysaccharidosis like skeletal and facial deformities, developmental delays and can have severe neurological and motor disorders in terms of a progressive disease (Futerman and van Meer, 2004). An overview of the known sulfatases, their localization and physiological substrates as well as the associated genetic diseases is described in Table 1.1.

Sulfatases	Localization	Substrate	Genetic disorder
Arylsulfatase A	Lysosomal	Cerebroside-3-sulfate	Metachromatic leucodystrophy
Arylsulfatase B	Lysosomal	CS/DS	Marotaux-lamy syndromy
Arylsulfatase C	ER/microsomal	–	X-linked ichthyosis
Arylsulfatase D	ER	–	–
Arylsulfatase E	Golgi	–	Chondrodysplasia punctate 1
Arylsulfatase F	ER	–	–
Arylsulfatase G	Lysosomal	–	–
Arylsulfatase H	–	–	–
Arylsulfatase I	–	–	–
Arylsulfatase J	–	–	–
Arylsulfatase K	–	–	–
Galactosamine 6-sulfatase	Lysosomal	CS, KS	Morquio A syndrome
Glucosamine 6-sulfatase	Lysosomal	HS, KS	Sanfilippo D syndrome
Heparan 6-sulfatase	Lysosomal	HS	Sanfilippo A syndrome
Iduronate 2-sulfatase	Lysosomal	HS, DS	Hunter syndrome
Sulfatase 1	Cell surface	HS	–
Sulfatase 2	Cell surface	HS	–

Table: 1.1. Overview of the human sulfatase family.

The table contains various known sulfatases along with their subcellular localization, physiological substrates and associated diseases present in human. MPS: mucopolysaccharidosis, DS: dermatan sulfate, CS: chondroitin sulfate, HS: heparan sulfate; '-': unknown. (Compiled from Dierks *et al*., 2009; Oshikawa *et al*., 2009; E. Westendorf, 2011 and Kowalewski *et al*., 2012).

The sulfatases listed in the table 1.1 belong to the so-called human sulfatase type I. They contain the common non-proteinogenic amino acid formylglycine (FGly, 2-amino-3 oxopropionic acid) at their active site. The FGly residue positioned within the active site is thought to undergo hydration to a geminal diol, after which one of the hydroxyl groups acts as a catalytic nucleophile to initiate the sulfate ester cleavage, as illustrated in Fig. 1.1 (Lukatela *et. al*., 1998 and Boltes *et. al*., 2001).

There are only cysteine-type sulfatases present in eukaryotes, while cysteine and serine-type sulfatases are present in bacteria, i.e., a cysteine or serine residue can be modified to a FGly

residue. Both of these types represent the FGly-dependent sulfatase and are referred as Type I sulfatases (Hagelueken *et al*., 2006).

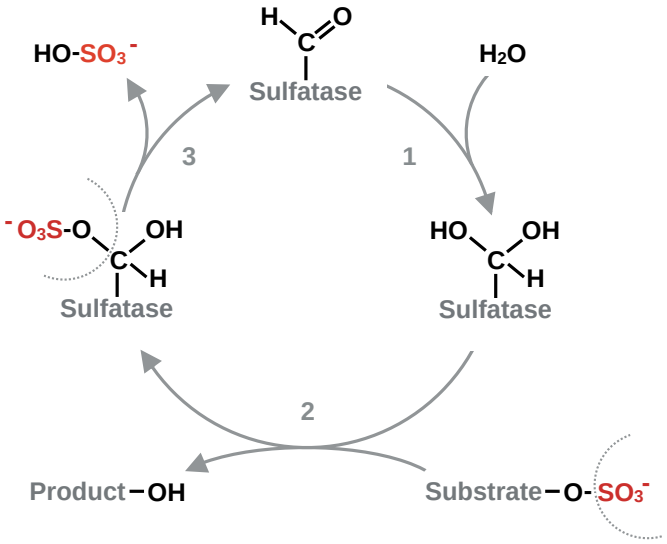

Figure 1.1: FGly-mediated sulfate ester hydrolysis.
The active site of the sulfatase contains the formylglycine (FGly) residue (top), which upon hydration (1) forms the aldehyde hydrate (right). This aldehyde hydrate undergoes a nucleophilic attack on the sulfate ester of the substrate by one of its germinal hydroxyls. This leads to trans-esterification (2) of the sulfate group onto the enzyme forming a sulfated FGly hemiacetal intermediate (left). The hydroxyl group of this intermediate reacts to eliminate the sulfate (3) undergoing cleavage of the C–O bond and finally regenerating the aldehyde (top). (Adapted from Dierks *et al*., 2009)

In addition, FGly-independent sulfatases are also present in prokaryotes. There is a Fe^{2+}-α-ketoglutarate-dependent dioxygenase-sulfatase (type II, Muller *et al*., 2004) and a metallo-beta lactamase-sulfatase (type III), which are characterized by a Zn^{2+} binding motif (Hagelueken *et al*., 2006). The function of prokaryotic sulfatases is far less studied. Unlike the human enzymes, the primary role of sulfatases is to supply sulfur for its growth by the release of inorganic sulfate from sulfate esters (Kertesz, 2000).

1.2 Multiple Sulfatase deficiency

The multiple sulfatase deficiency (MSD) is a very rare autosomal recessive disorder in humans, with a prevalence of 1 in 1.4 million births (Hopwood and Ballabio, 2001). In MSD, the activity of all sulfatases is drastically reduced and thus shows combined characteristic features of individual sulfatase deficiencies. Patients with MSD show a neurodegenerative loss of sensoric and motoric abilities and neurological deterioration. In addition, mental retardation, hepatospleenomegaly, shortening of stature and corneal clouding are commonly observed in MSD (Hopwood and Ballabio, 2001, Nyhan and Ozand, 1998, van der Knaap *et al*. 2005). Patients also show abnormal facial features and skeletal deformation. Most often, the disease leads to death in infancy (Burch *et al*., 1986, Vamos *et al*., 1981, Busche *et al*., 2009).

According to characteristic symptoms and their age of manifestation, MSD can be divided into different levels of severity. Very severe neonatal course forms can be distinguished from the late infantile and juvenile forms (Hopwood and Ballabio, 2001, Mancini *et al*., 2001, Loffeld *et al*., 2002, Cosma *et al*., 2004, Schlotawa *et al*., 2008, Annunziata *et al*., 2007).
MSD affected patients have mutations in the gene *SUMF1* (sulfatase modifying factor 1) which encodes the protein FGE. The proof of this being indeed a monogenic hereditary disease was achieved by recombinant expression of FGE in patient fibroblast when the normal sulfatase activity was restored (Dierks *et al*., 2003, Cosma *et al*., 2003). It has been shown that the activated sulfatases possess an essential FGly residue at its active site that is formed by the modification of cysteine residue catalyzed by FGE.
Currently, about 30 different disease-causing mutations in the gene SUMF1 in MSD patients are known. Most of these are missense mutations, which are hypomorphic in nature and vary to different degrees influencing the structure stability, activity and substrate affinity (Dierks *et al*., 2005 Schlotawa *et al*., 2008). Functional analysis of these individual mutations in FGE led to a reliable genotype-phenotype correlation for MSD (Schlotawa *et al*., 2011.). This study showed that unstable proteins with low residual activity led to severe forms of this disease, stable proteins with low residual activity or unstable FGE variants with high residual activity result in milder gradients of MSD. On the basis of these discoveries, a prognosis of the disease process in the present classification of the mutation is possible.
So called null mutations that are associated with a complete loss of function of the catalytic activity of FGE or do not allow protein expression, are considered to be embryonically lethal, because they have only been found in heterozygous MSD patients. In contrast, the establishment of a viable SUMF1 gene trap mouse line (Settembre *et al*., 2007) was successful. This mouse model allowed the study of the pathophysiology of lysosomal storage in MSD (Settembre *et al*., 2008a) and can be used for therapeutic trials. Currently, there is no effective therapy available for multiple sulfatase deficiency.
From a biochemical point of view, the metabolic disorder emphasizes two features. Firstly, MSD represents a monogenic hereditary disease, though it is caused by drastic reduction in the activity of a variety of enzymes (all known sulfatases). Secondly, this lysosomal storage disease is ultimately based on the lack of a non-lysosomal enzyme activity (Dierks *et al*., 2009).

1.3 Formylglycine Generating Enzyme

Two independent research groups identified the gene encoding the FGly-modifying factor in 2003 (Dierks *et al*., 2003, Cosma *et al*., 2003). The two groups agreed to name the gene sulfatase-modifying factor 1 (*SUMF1*) and the encoded protein was denoted by formylglycine-generating enzyme (FGE).
The gene *SUMF1*, encoding FGE, is located on chromosome 3p26 in humans and 1.06 kb in size, comprising nine exons. *SUMF1* is highly conserved in higher eukaryotes and is also

present in some prokaryotes, however, it is absent in *S. cerevisiae, C. elegans, E. coli,* and plants (Dierks *et al*., 2003, Landgrebe *et al*., 2003).

1.3.1 Preliminary biochemical characterization of FGE

Biochemical and structural studies have led to a very detailed characterization of human FGE. It is localized in the endoplasmic reticulum (ER), where it activates the sulfatases cotranslationally by catalyzing the modification of the conserved cysteine residue to the Cα-FGly (Dierks *et al*., 1997, 1998a, Fey *et al*., 2001). The human FGE consists of 374 amino acids. Out of these, 33 amino acids form the N-terminal signal sequence, which is involved in targeting FGE into ER. After signal peptide cleavage, mature intracellular FGE has a molecular mass of 41 kDa consisting of two domains. The largest part of FGE consisting of residues 91-374 forms the core domain, which is highly conserved in prokaryotes and eukaryotes. The N-terminal extension (residues 34-90), which probably forms a separate domain, is only found in eukaryotes, where it serves essential functions *in vivo.* FGE contains eight cysteine residues and intracellular FGE carries high mannose-type glycosylation at N141. When FGE is secreted, the carbohydrate side chain is modified to a complex-type and a stretch of 39 amino acids (residues 34-72) is cleaved off from the N-terminus of majority of the secreted FGE by a furin-like protease in Golgi during its transportation (Preusser-Kunze *et al*., 2005, Ennemann *et al*., 2013). The full length-FGE as well as the N-terminally truncated FGE were shown to modify sulfatase-derived peptides *in vitro,* indicating that the globular domain of FGE contains the active site of the enzyme.

1.3.2 Crystal structure of FGE

The first crystal structure of FGE (secreted FGE, residues 73-374) was solved in 2003. It shows a unique globular structure with a very low (23%) secondary structure content (Dierks *et al*., 2005; Fig. 1.2). FGE does not contain redox-active metals or cofactors. The two Ca^{2+} ions are coordinated by highly conserved amino acid residues and stabilize the hydrophobic core of the structure (Fig. 1.2).

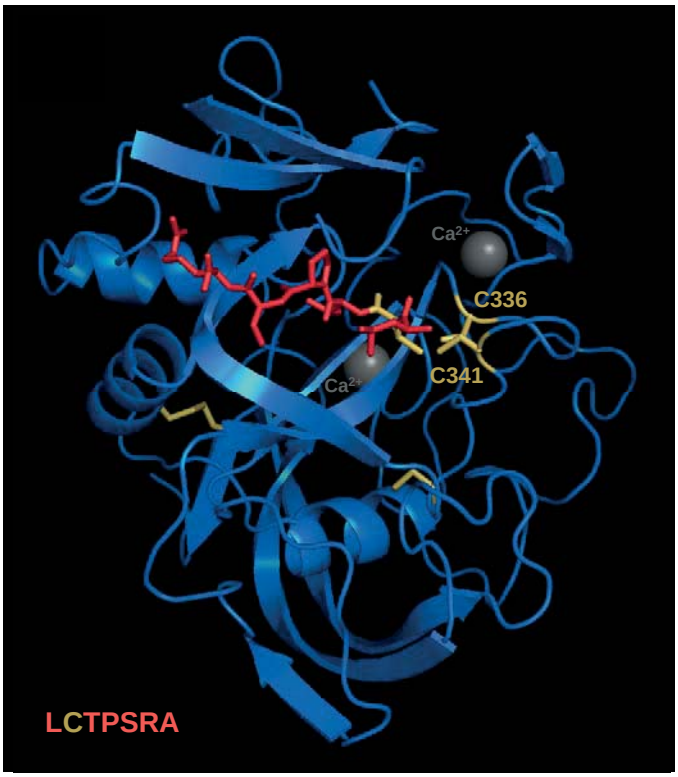

Figure 1.2: Crystal structure of human FGE.
The core domain of FGE is shown (as ribbon in blue) in compex with the sulfatase-derived peptide LCTPSRA (as sticks in red) as determined by Roeser *et al*., 2006. Cysteines are shown in yellow, the active site cysteines C336 and C341 are labelled. The two calcium ions are shown as grey balls. (Adopted from Dierks *et al.* 2009)

Ca^{2+} -binding site-I can be best described as a seven-coordinate distorted pentagonal bipyramidal structure that holds Ca^{2+}, coordinated by amino acids D273, N259, I260, F275 and two water molecules. Ca^{2+}-binding site-II shows an almost perfect octahedral coordination of Ca^{2+} by E130, E300, N293, G296, A298, and a water molecule (Fig. 1.3 right). The chemistry of eight cysteines present in FGE has become the focus of current research (Dierks *et al*., 2005, Mariappan *et al*., 2008a and b, Wachs 2009). The cysteine pairs C218-C365 and C235-C246 of the core domain are referred to as structural cysteines, since they are connected by disulfide bonds in the crystal structure providing structural stability. Meanwhile, there are indications that, under certain conditions at least the cysteine pair C218-C365 remain unbridged (Wachs, 2009).

The two cysteines of the active site, C336 and C341 are redox active and essential for the catalytic generation of FGly. C336 and C341 in the crystals were found reduced as well as in the oxidized disulfide-bridged state (Dierks *et al*., 2005). The electron density in different crystals at C336 also pointed to a sulfenic acid (including a water molecule) or a hydroperoxide, which suggested that C336 is highly redox active. The catalytic center of FGE is located on the surface of the structure in form of an oval shaped groove of 20 Å length, 10 Å depth and 12 Å width. This groove has two redox-sensitive cysteines (C336 and C341) on one end and a proline (P182) on the other end and can accommodate the binding of up to six amino

acids of nascent sulfatase polypeptide. The minimal recognition sequence of the sulfatase, C/S-[TSAC]-PXR, within the sulfatase signature sequence type I, is highly conserved. The complex of FGE and substrate peptide probably represents a general binding mechanism in which FGE is present in association with the still unfolded sulfatase polypeptide chain (Roeser *et al*., 2006). The efficiency of the conversion of the cysteine to FGly is also influenced by the sequence C-terminally of the minimal sequence motif (Dierks *et al*., 1999). The evaluation of the X-ray structure analysis of crystals in which substrate is bound to FGE, showed that especially the side chains of proline and arginine within the CTPSR motif of the peptide interacts with the active site of FGE (Roeser *et al*., 2006). This is consistent with the result that replacement of these residues leads to loss of FGly generating activity (Dierks *et al*., 1999).

A functional and structural study of prokaryotic FGE based on the crystal structure of *Streptomyces coelicolor* FGE and activity studies of FGE from *Mycobacterium tuberculosis* was carried out by Carlson *et al*., published in 2008. It showed that the structure of the prokaryotic FGE is remarkably similar to that of human FGE. It has the typical "FGE fold", with the characteristically low secondary structure content and the presence of a similar surface of the substrate binding pocket.

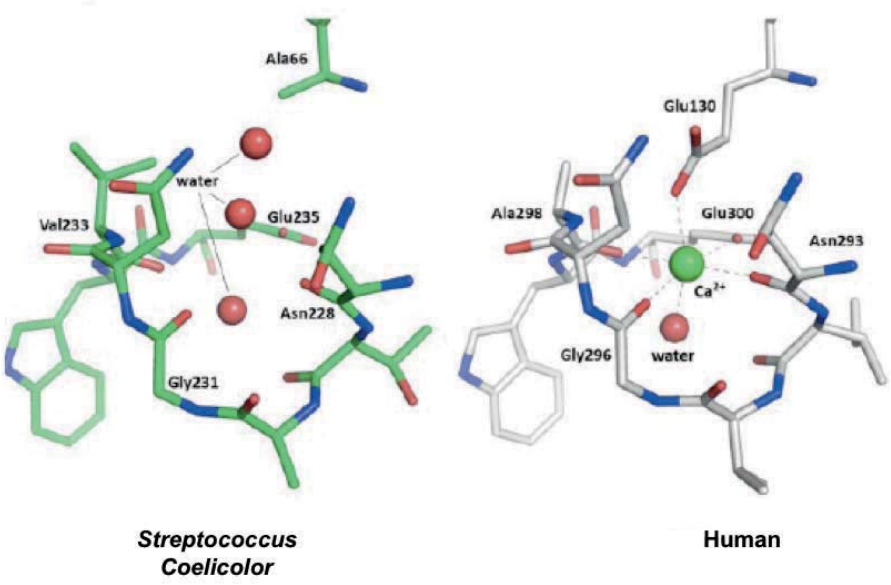

Figure 1.3: Structure of Ca^{2+} binding site II in Streptomyces FGE (left) and human FGE (right). Ca^{2+}-binding site-II in FGE from *Streptococcus coelicolor* and human are almost conserved except A66 in *S.* coelicolor which replaces E130 in human FGE. This leads to complete absence of Ca^{2+} at binding site-II which is replaced by two water molecules in *S.coelicolor* FGE. (Adopted from Dierks *et al*., 2005 and Carlson *et al*., 2008).

Prokaryotic FGE also acts as a monooxygenase, requires molecular oxygen and is independent of any organic cofactors or redox-active metals. Similar to the human FGE, prokaryotic FGE also contains the two essential catalytically active and redox-sensitive cysteines within a motif $C_{X4}C$. A striking contrast to human FGE is that FGE crystallized from *S. coelicolor* contains only one Ca^{2+} ion, whis is at position of Ca^{2+}-binding site I. A Ca^{2+} ion at Ca^{2+}-binding site II is

missing. Though the Ca^{2+} is absent at calcium binding site II, the site is conserved in prokaryotic FGE (Fig. 1.3). This might be probably due to A66 in *S. coelicolor* that disrupts an appropriate coordination environment. However, *M. tuberculosis* does not contain any Ca^{2+} ion in its structure. Unlike human FGE, the prokaryotic FGE do not require Ca^{2+} ions for crystallization. In addition, the prokaryotic FGE is even active in presence of EDTA (Carlson *et al*., 2008).

1.3.2 Role of the N-terminal domain

In comparison to prokaryotic FGE, eukaryotic FGE contains an N-terminal extension (residues 34-91), which is entirely encoded by exon 1 (Dierks *et al*. 2009). The major part of the N-terminal domain is cleaved off at R72, which is why this domain is missing in purified secreted FGE. Thus most of the structural and functional *in vitro* studies were determined for secreted purified N-terminally truncated FGE. The amino acid residues 34-374 comprise the full-length form of FGE denoted as fl-FGE (full-length FGE) and residues 73-374 the truncated form, denoted as Δ72-FGE.

Δ72-FGE alone is not able to activate sulfatases *in vivo*, but when coexpressed with pFGE (paralog of FGE, see section 1.4.1), to which the N-terminus of FGE had been fused, it activates an also coexpressed reporter sulfatase in cultured cells. This trans-activation indicates a cooperation between the N-terminal domain and the core domain of FGE during activity *in vivo* (Mariappan *et al*., 2008a).

The N terminus contains a fully conserved pair of cysteines (C50 and C52) arranged as a CXC motif and forming a disulfide bridge. Interestingly, this shows a functional mimicry to the thioredoxin domain of protein disulfide isomerase (PDI). C50 and C52 are either involved in an intra- or an intermolecular disulfide bridge. The intermolecular disulfide bond leads to formation of FGE homodimers and heteromers with other interacting partners of FGE (Preusser-Kunze *et al*., 2005; Mariappan *et al*., 2008b, Fraldi *et al*., 2008). It has also been shown that the N-terminal domain of FGE is also required for the retention of FGE in ER (Mariappan *et al*., 2008a).

It has been shown recently that furin-like proprotein convertase cleaves the RYSR motif present in the N-terminus of FGE. The CGC motif is crucial for the biological activity of FGE and highly conserved in eukaryotes, from humans to sponges. The N-terminal domain serves to adapt FGE to ER-based functioning in eukaryotes for two reasons: (i) ER retention of FGE is mediated by interaction with ERp44, an ER protein and (ii) for inter- or intramolecular activation of the catalytic domain of FGE by CGC. Furthermore, it has also been shown that cleavage of N terminus leads to inactivation of secreted human FGE, suggesting a possible role of N terminus in regulation of FGE function or activity (Ennemann *et al*., 2013).

1.4 Interacting partners of FGE

FGE is an ER-resident protein and it lacks a KDEL-like retention signal in its amino acid sequence. In addition, any cofactor or redox active metal ions associated with FGE function is not known. However, proteins that were identified to interact with FGE thereby fulfill a variety of functions.

1.4.1 pFGE is a paralog of formylglycine generating enzyme

pFGE and FGE share many structural and topological properties but they differ in their function. FGE catalyzes the oxidation of cysteine residue in the active site to FGly residue in sulfatases, whereas pFGE lacks such an activity *in vitro* as well as *in vivo.* pFGE is a soluble glycoprotein present in endoplasmic reticulum and is retained in the ER through its C-terminal PGEL sequence, a non-canonical variant of the classic KDEL-like retention signal (Gande *et al*., 2008). It contains a single high mannose-type oligosaccharide side chain intracellularly, which upon secretion gets converted to hybrid or complex oligosaccharide structures containing fucose and sialic acid residues (Zito *et.al.*, 2005, Mariappan *et.al*, 2005). pFGE is found only in vertebrates, and it shares 48% amino acid identity and 62% similarity with FGE (Cosma *et al*.,2003). pFGE is similar to FGE in its crystal structure and lacks secondary structural elements. It also contains the novel FGE fold with an interesting partitioning of the molecule into two halves, distinguishable by the amount of secondary structural elements they contain. A conserved disulfide bond formed between Cys-156 and Cys-290 and two Ca^{2+} ions present in the molecule stabilize the tertiary structure of pFGE. Similar to FGE, the positions of the two Ca^{2+} ions present in pFGE are conserved (Fig. 1.4 for Ca^{2+} binding site II).