Computational Biology and Chemistry 68 (2017) 71–77

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/compbiolchem

Research Article

Comparative and evolutionary studies of mammalian arylsulfatase and sterylsulfatase genes and proteins encoded on the X-chromosome

Roger S. Holmes

The Griffith Research Institute for Drug Discovery and School of Natural Sciences, Griffith University, Nathan, 4111 QLD, Australia

ARTICLE INFO

Article history:
Received 17 September 2016
Received in revised form 22 December 2016
Accepted 22 February 2017
Available online 24 February 2017

Keywords:
Arylsulfatase
Sterylsulfatase
STS
ARS
ARSD
ARSE
ARSF
ARSH
X-chromosome
Primates
Eutherian mammals
Marsupials
Evolution
Phylogeny
Primordial gene
Signal peptide
Transmembranes
Ca^{2+} binding
Active site
N-Glycosylation site

ABSTRACT

At least 19 sulfatase genes have been reported on the human genome, including four arylsulfatase (*ARS*) genes (*ARSD*; *ARSE*; *ARSF*; *ARSH*) and a sterylsulfatase (*STS*) gene located together on the X-chromosome. Bioinformatic analyses of mammalian genomes were undertaken using known human STS and ARS amino acid sequences to study the evolution of these genes and proteins encoded on eutherian and marsupial genomes. Several domain regions and key residues were conserved including signal peptides, active site residues, metal (Ca^{2+}) and substrate binding sequences, transmembranes and *N*-glycosylation sites. Phylogenetic analyses describe the relationships and potential origins of these genes during mammalian evolution. Primate ARSH enzymes lacked signal peptide sequences which may influence their biological functions. CpG117 and CpG92 were detected within the 5′ region of the human *STS* and *ARSD* genes, respectively, and miR-205 within the 3′-UTR for the human *STS* gene, using bioinformatic methods A proposal is described for a primordial invertebrate *STS*-like gene serving as an ancestor for unequal cross over events generating the gene complex on the eutherian mammalian X-chromosome.

1. Introduction

At least ninteen human sulfatase genes have been described, which encode enzymes with diverse metabolic roles in hydrolysing biological sulfate esters (Ratzka et al., 2010). These include a cluster of five arylsulfatase-like (*ARS*) genes located on the X-chromosome: sterylsulfatase (*STS*) (Yen et al., 1998); *ARSD* (Urbitsch et al., 2000); *ARSE* (Parenti et al., 1997); *ARSF* (Puca et al., 1997); and *ARSH* (Sardiello et al., 2005). Several X-linked diseases associated with variants for these genes have been described, including chronic lymphocytic leukemia for *ARSD* (Trojani et al., 2011), chrondrodysplasia punctata for *ARSE* (Franco et al., 1995), and X-linked ichthyosis for *STS* (Basler et al., 1992).

These arylsulfatases differ in their tissue and subcellular distribution, kinetic properties and biological roles. STS, also called steroid or estrone sulfatase (E.C.3.1.6.2), is associated with the endoplasmic reticulum membrane and is predominantly responsible for the catalytic conversion of inactive sulfated female sex hormone precursors to estrogens during pregnancy (Basler et al., 1992; Purohit et al., 2011). In contrast, the biological substrate (s) for the other X-linked arylsulfatases are not known, although roles in sphingolipid metabolism have been proposed, due to metabolic impacts of X-linked diseases (Franco et al., 1995; Trojani et al., 2011). ARSD (EC 3.1.6.1) is postranslationally glycosylated and

Abbreviations: ARS, arylsulfatase; STS, sterylsulfatase; ARSD, arylsulfatase D; ARSE, arylsulfatase E; ARSF, arylsulfatase F; ARSH, arylsulfatase H; UCSC, University of Santa Cruz California; EC, enzyme commission; BLAST, Basic Local Alignment Search Tool; BLAT, Blast-Like Alignment Tool; NCBI, National Center for Biotechnology Information; AceView, NCBI based representation of public mRNAs; TFBS, transcription factor binding sites; UTR, untranslated gene region; CpG, region of high density of guanine-cytosine dinucleotides; mRNA, messenger RNA.

E-mail address: r.holmes@griffith.edu.au (R.S. Holmes).

http://dx.doi.org/10.1016/j.compbiolchem.2017.02.009
1476-9271/

localized in lysosomes, and has been identified as a marker for chronic lymphocyte leukemia progression (Trojani et al., 2011). ARSE (EC 3.1.6.1) has been recognized clinically due to the impacts of mutations in the human *ARSE* gene resulting in bone dysplasia (X-linked recessive chrondrodysplasia punctata or CDPX1) (Parenti et al., 1997; Brunetti-Pierri et al., 2003; Matos-Miranda et al., 2013). Further sequencing of the human X-chromosome has enabled the identification of two other arylsulfatase genes, designated as *ARSF* and *ARSH*, which are localized within a 206 kb cluster on the X-chromosome (Puca et al., 1997; Sardiello et al., 2005). These arylsulfatase genes and enzymes have not been well characterized although sequence similarities with ARSD and ARSE are suggestive of related biological roles in lysosomal sphingolipid metabolism.

This study describes the predicted sequences, structures and phylogeny of mammalian *STS, ARSD, ARSE, ARSF* and *ARSH* genes and enzymes and compares these results with those previously reported for human and mouse *STS* genes and proteins (Stein et al., 1989; Hernandez-Guzman et al., 2003). Evidence is presented concerning the sequencing and properties of these genes and proteins from several mammalian species, and for distinct modes of gene regulation and expression with CpG islands identified for the human *STS* and *ARSD* gene promoters and miR-205 for the 3′-UTR region of the *STS* gene. Phylogenetic analyses also describe the relationships and potential origins of these *STS*-and *ARS*-like genes and enzymes during mammalian evolution and a proposal for generating these X-linked genes from an ancestral invertebrate *STS*-like gene.

2. Materials and methods

2.1. STS- *and* ARS-*like gene and enzyme identification*

STS and ARS-like amino acid sequences for mammalian and an invertebrate species were retrieved from databases (NCBI (http://blast.ncbi.nlm.nih.gov/Blast.cgi) and ExPASy (http://www.expasy.org) (Artimo et al., 2012)), using human, mouse and sea urchin (*Strongylocentrotus purpuratus*) STS and ARS-like sequences to seed searches. Identification of the genes was based on high predictive scores (>850) and sequence coverage (>98%) for STS- and ARS-like proteome sequences (as listed by NCBI) in each case (Tables 1 and S1). BLAT searches were performed using relevant protein sequences to confirm the gene and enzyme sequences among the species examined using the UCSC Genome Browser (Karolchik et al., 2009). Gene locations, predicted gene structures and protein subunit sequences were obtained for each gene and enzyme examined showing identity with the respective sequences (Tables 1 and S1). Representations of human *STS, ARSD, ARSE, ARSF* and *ARSH* gene structures were obtained using the AceView (Thierry-Mieg and Thierry-Mieg, 2006) (http://www.ncbi.nlm.nih.gov/IEB/Research/Acembly/) web browser. Identification of potential gene regulatory sites, including CpG islands and miRNA-binding sites within the respective 5′ and 3′ gene regions, were undertaken using the UCSC Human Genome Browser (Karolchik et al., 2009).

2.2. Predicted structures and properties of human ARSD, ARSE, ARSF and ARSH subunits

Predicted secondary and tertiary structures for human sequences were obtained using SWISS MODEL web tools (Schwede et al., 2003). The human STS tertiary structure (PDB:1P49) (Hernandez-Guzman et al., 2003) served as a reference for obtaining these structures, with modelled residue ranges of 23–575 for human STS; 37–588 for human ARSD; 34–585 for human ARSE; 26–576 for human ARSF; and 4–539 for human ARSH. Predicted transmembrane structures for human STS, ARSD, ARSE, ARSF and ARSH subunits were obtained using the web server (http://www.cbs.dtu.dk/services/TMHMM-2.0) provided by the Center for Biological Sequence Analysis of the Technical University of Denmark (Krogh and Larsson, 2001). SignalP 3.0 web tools were used to predict the presence and location of signal peptide cleavage sites (http://www.cbs.dtu.dk/services/SignalP); and the NetNGlyc 1.0 server used to predict potential *N*-glycosylation sites for human STS, ARSD, ARSE, ARSF and ARSH subunits (http://www.cbs.dtu.dk/services/NetNGlyc) (Emanuelsson et al., 2007).

2.3. Amino acid sequence alignments and phylogenetic analyses

Alignments of identified mammalian STS and ARS-like sequences were undertaken using Clustal Omega, a multiple sequence alignment program (Sievers and Higgins, 2014). Phylogenetic analyses used several bioinformatic programs, coordinated using the http://www.phylogeny.fr/bioinformatic portal, to enable alignment (MUSCLE), curation (Gblocks), phylogeny (PhyML) and tree rendering (TreeDyn), to reconstruct phylogenetic relationships (Dereeper et al., 2008). Sequences were identified as STS and ARS-like members, including a proposed primordial *STS* gene and STS protein identified for the sea urchin (*Strongylocentrotus purpuratus*) (Tables 1 and S1).

Table 1
Human ARSD, ARSE, ARSF, ARSH and STS genes and enzymes.

Gene	Enzyme ID	Enzyme Name	Chromosome location	Coding Exons (strand)	Gene Size bps	GenBank ID	UNIPROT ID	Amino acids	Subunit MW (pI)	Leader Peptide	Genetic Disease	Major Tissue Expression (x times average gene expression)
ARSD	3.1.6.1	arylsulfatase D	X:2,907,274-2,929,275	10 (-ve)	22,002	NM_009589	P51689	593	64,859 (6.8)	1.33	chronic lymphocytic leukemia	(x3.0): ovary, kidney, liver, lung, colon, testis, blood, brain, pancreas
ARSE	3.1.6.1	arylsulfatase E	X:2,934,835-2,958,434	10 (-ve)	23,600	NM_000047	P51690	589	65,669 (6.5)	1.31	chrodrodysplasia punctata	(x1.2): liver, kidney, colon, testis, intestine, pancreas
ARSH	3.1.6.1	arylsulfatase H	X:3,006,613-3,033,382	9 (+ve)	26,770	NM_001011719	Q5FYA8	562	63,525 (8.5)	na	na	(0.1): brain, testis, kidney
ARSF	3.1.6.1	arylsulfatase F	X:3,072,024-3,112,553	10 (+ve)	40,530	NM_001201538	P54793	590	65,940 (6.8)	1.22	na	(x0.3): kidney, brain, liver
STS	3.1.6.2	steryl sulfatase	X:7,253,194-7,350,258	10 (+ve)	97,065	NM_001320750	P08842	583	65,492 (7.6)	1.21	X-linked ichthyosis	(x2.5): brain, placenta, testis, liver, kidney, heart, lung, colon

na-not available; genetic diseases associated with these genes are shown; major tissues for expression are identified.