Vertebrate Arylsulfatase K (ARSK): Comparative and Evolutionary Studies of the Lysosomal 2-Sulfoglucuronate Sulfatase

Roger S Holmes*
Griffith Institute for Drug Discovery and School of Natural Sciences, Griffith University, Nathan, QLD, Australia

Abstract

Arylsulfatase K (ARSK) is one of 17 sulfatase gene family members encoded on the human genome for which a role has been recently identified as a lysosomal 2-sulfoglucuronate sulfatase. Vertebrate ARSK sequences shared 60-82% identity but only <27% identities with other arylsulfatase family members. Comparative enzyme structures were studied, including residues with predicted roles in forming N-glycosylation sites, Ca<sup>2+</sup> binding and active site residues. Vertebrate ARSK genes usually contained 8 coding exons. A human ARSK gene promoter comprised CpG61 and multiple TFBS, which may be involved in signal transduction, transcription activation or regulating entry into cell division. Phylogenetic analyses examined evolutionary changes for the vertebrate ARSK and the invertebrate SUL1 genes. In summary, a major role for this enzyme as a 2-sulfoglucuronate sulfatase is supported which has been conserved throughout vertebrate evolution.

Keywords: Vertebrates; Arylsulfatase K; Amino acid sequence; 2-sulfoglucuronate sulfatase; Chromosome 5

Abbreviations: ARSK: Arylsulfatase K; kbps: Kilobase Pairs; CpG Island: Multiple C (Cytosine)-G (Guanine) Dinucleotide Region; miRNA: microRNA Binding Region; BLAST: Basic Local Alignment Search Tool; BLAT: Blast-Like Alignment Tool; NCBI: National Center for Biotechnology Information; SWISS-MODEL: Automated Protein Structure Homology-Modeling Server

Introduction

Seventeen human sulfatase gene families and fourteen mouse sulfatase gene families encode sulfatases which catalyse the hydrolysis of a range of biological sulfate esters in the body [1-5]. The gene encoding arylsulfatase K (ARSK; EC 3.1.6.13) (ARSK in vertebrates; Arsk in rodents) was initially identified using bioinformatic methods through its conserved enzyme; exonic structures for this conserved enzyme; exonic structures for key structures for this conserved enzyme; exonic structures for vertebrate ARSK genes; potential sites for regulating human ARSK gene expression; alignment and evolutionary studies of vertebrate ARSK and invertebrate arylsulfatase (SUL1) and comparisons of ARSK structures with other sulfatase gene families.

Methods

Gene (ARSK) and protein (ARKS) structures

Vertebrate ARSK amino acid sequences were derived from BLAST studies using NCBI web tools (http://www.ncbi.nlm.nih.gov/) [9] using the human ARSK sequence [1,7] (Table 1). BLAT analyses were subsequently undertaken for each of the predicted ARSK amino acid sequences using the UC Santa Cruz (UCSC) Genome Browser to obtain predicted locations, exon boundaries and gene sizes for each of the vertebrate ARSK and SUL1 genes (Table 1) [10]. Structural features for the major human mRNA ARSK isoform were also examined using AceView [11].

Structures and properties of vertebrate ARSK

Predicted structures for vertebrate ARSK proteins were obtained using the SWISS-MODEL web-server (http://swissmodel.expasy.org/) [12] and a tertiary structure reported for a putative sulfatase from Bacteroides thetaiotaomicron (PDB:1b5q), with a modelling residue range of 35-475 for human ARSK [13]. Predicted signal peptide cleavage sites, N-glycosylation sites, molecular weights and theoretical isoelectric points for vertebrate ARSK proteins were obtained using Expasy web tools (http://au.expasy.org/tools/pi_tool.html).

Human ARSK tissue expression

Comparative tissue expression levels for human ARSK mRNA were obtained using a GTEx database [14] (Data Source: GTEx Analysis Release V6p (dbGaP Accession phs000424.v6.p1) (http://www.gtex.org).

*Corresponding author: Roger S Holmes, Griffith Institute for Drug Discovery and School of Natural Sciences, Griffith University, Nathan, QLD, Australia, Tel: +61 410583348; E-mail: r.holmes@griffith.edu.au

Received July 12, 2017; Accepted August 14, 2017; Published August 21, 2017

Citation: Holmes RS (2017) Vertebrate Arylsulfatase K (ARSK): Comparative and Evolutionary Studies of the Lysosomal 2-Sulfoglucuronate Sulfatase. J Data Mining Genomics Proteomics 8: 212. doi: 10.4172/2153-0602.1000212

Copyright: © 2017 Holmes RS. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

DOI: 10.4172/2153-0602.1000212
Amino acid sequence alignments, identities and phylogenetic analyses

Clustal Omega was used to generate amino acid sequence alignments and percentage identities for vertebrate ARSK and other human arylsulfatase sequences (Tables 1 and 2) [15]. Phylogenetic analyses used the http://www.phylogeny.fr/ bioinformatic portal [16] to reconstruct vertebrate ARSK evolutionary relationships with a C. elegans SUL1 gene and protein (Table 1).

Results and Discussion

ARKS and other human sulfatase genes and proteins: Proposed classification

Table 2 summarises a classification scheme for 17 human sulfatase genes and proteins previously proposed, based on phylogenetic and amino acid sequence comparisons for human SULF1 and SULF2 (extracellular sulfatases 1 and 2) [5] and the other 15 human sulfatases, including the human ARSK gene and protein. Seven groups of these genes and proteins were identified, including ARS group 5 for a single human ARSK gene, which encodes a distinct 526 amino acid sequence enzyme and shares <27% sequence identities with other human sulfatases. This is in contrast to the sequence identities observed for other human enzyme ARS groups: Group 1 (ARSA, ARSG and GALNS) (33-39% identical); group 2 (ARSB, ARSI and ARSJ) (54-59%) [2]; group 3 (ARSD, ARSE, ARSF and STS) (57-64%) [3]; and group 4 ARS sequences (SULF1, SULF2 and GNS) (42-67%) [5]. Groups 6 (SGSH) and 7 (IDS) [4] enzymes, however, exhibited distinct amino acid sequences which were 22% and 26% identical, respectively, with the human ARSK amino acid sequence (Table 2). With the exception of the group 3 gene cluster previously reported on the human X-chromosome for ARSD, ARSE, ARSF, ARSH and STS [3], other human ARS genes are separately located on the human genome, including ARSK on human chromosome 5 [6,7].

<table>
<thead>
<tr>
<th>Gene</th>
<th>Group</th>
<th>Organism</th>
<th>Species</th>
<th>Chromosome* Location</th>
<th>Coding Exons (strand)</th>
<th>Gene Size (bps)</th>
<th>GenBank ID*</th>
<th>UniProt ID</th>
<th>Amino acids</th>
<th>MW (pI)</th>
<th>% Identity ARSK</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARSK</td>
<td>1</td>
<td>Human</td>
<td>Homo sapiens</td>
<td>5:95,555,279-95,603,523</td>
<td>8 (+ve)</td>
<td>48,245</td>
<td>NM_198150</td>
<td>Q6UWY0</td>
<td>561,450 (9.0)</td>
<td>...22</td>
<td>25</td>
</tr>
<tr>
<td>ARSK</td>
<td>2</td>
<td>Baboon</td>
<td>Papio anubis</td>
<td>6:69,595,690-69,831,202</td>
<td>8 (+ve)</td>
<td>45,513</td>
<td>XP_003899879*</td>
<td>A0A96983M3D</td>
<td>561,368 (8.8)</td>
<td>...15</td>
<td>20</td>
</tr>
<tr>
<td>ARSK</td>
<td>3</td>
<td>Mouse</td>
<td>Mus musculus</td>
<td>13:76,062,259-76,098,499</td>
<td>8 (+ve)</td>
<td>38,241</td>
<td>NM_028847</td>
<td>Q8D2L1</td>
<td>556,130 (8.7)</td>
<td>...18</td>
<td>16</td>
</tr>
<tr>
<td>ARSK</td>
<td>4</td>
<td>Cow</td>
<td>Bos taurus</td>
<td>7:97,276,428-97,331,337</td>
<td>8 (+ve)</td>
<td>54,910</td>
<td>BC118338*</td>
<td>Q148F3</td>
<td>540,631 (9.1)</td>
<td>...15</td>
<td>13</td>
</tr>
<tr>
<td>ARSK</td>
<td>5</td>
<td>Tasmanian devil</td>
<td>Sarkoculis harissi *GL841273:981,226-1,016,664</td>
<td>8 (+ve)</td>
<td>35,439</td>
<td>XP_003758981*</td>
<td>G3WFMI</td>
<td>553,632 (8.3)</td>
<td>...21</td>
<td>17</td>
<td></td>
</tr>
<tr>
<td>ARSK</td>
<td>6</td>
<td>VOY</td>
<td>Gallus galus</td>
<td>Z:57,238,252-57,248,657</td>
<td>8 (+ve)</td>
<td>10,406</td>
<td>NM_001031415</td>
<td>Q5ZK90</td>
<td>535,631 (8.7)</td>
<td>...24</td>
<td>14</td>
</tr>
<tr>
<td>ARSK</td>
<td>7</td>
<td>Frog</td>
<td>Xenopus tropicalis *KB021649:36,197,268-36,213,144</td>
<td>8 (+ve)</td>
<td>15,877</td>
<td>XP_002833455*</td>
<td>Q0IHJ2</td>
<td>560,606 (6.5)</td>
<td>...24</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>ARSK</td>
<td>8</td>
<td>Zebra fish</td>
<td>Danio rerio</td>
<td>5:50,252,467-50,263,734</td>
<td>8 (+ve)</td>
<td>11,268</td>
<td>NM_001071757</td>
<td>Q8C6J7</td>
<td>523,59,470 (6.3)</td>
<td>...16</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Vertebrate ARSK and worm SUL1 genes and proteins. RefSeq: the reference amino acid sequence; *Predicted amino acid sequence; GenBank IDs are derived from NCBI http://www.ncbi.nlm.nih.gov/genbank/; UniProt refers to UniProtKB/Swiss-Prot IDs for individual ARSK and SUL1 proteins (http://kr.expasy.org); ^ refers to a scaffold; bps refers to base pairs of nucleotide sequences; pI refers to theoretical isoelectric points; the number of coding exons are listed.

Table 2: Proposed classification of ARSK and other human arylsulfatase genes and proteins. Based on a previous proposal for the proposed classification of human arylsulfatase genes and proteins into 7 groups [5]; ARSK data is highlighted in red; pI=isoelectric point; bps=Base pairs of nucleotide sequence; % identities with the human ARSK amino acid sequence are shown; ARSA: arylsulfatase A; ARSG: Arylsulfatase G; GALNS: N-acetylgalactosaminase-6-sulfatase; ARSB: Arylsulfatase B; ARSI: Arylsulfatase I; ARSJ: Arlysulfatase J; ARSD: Arylsulfatase D; ARSE: Arylsulfatase E; ARSF: Arylsulfatase F; ARSH: Arylsulfatase H; STS: Sterolsulfatase; SUL1: Extracellular sulfatase 1; SUL2: Extracellular sulfatase 2; GNS: N-acetylgalactosamine-6-sulfatase; SGSH: N-sulfatase N-sulfoglucosamine sulphohydrolase; IDS: Iduronate 2-sulfatase.
Vertebrate ARSK and SUL1 amino acid sequences

The amino acid sequence for human ARSK [6,7] and the deduced amino acid sequences for mouse (Mus musculus), chicken (Gallus gallus) and zebrafish (Danio rerio) ARSK are shown in Figure 1 (Table 1). These and other vertebrate ARSK sequences were more than 60% identical, suggesting that they are members of the same arylsulfatase gene family. In contrast, the human ARSK protein sequence was <27% identical with other vertebrate arylsulfatase families (Table 2), confirming the separate family status for ARSK enzymes. Vertebrate ARSK enzymes contained 523-556 amino acids (Figure 1 and Table 1), including key residues and arylsulfatase domains previously reported [6,7]: Active site residues (human ARSK numbers used) binding calcium ions (Ca2+) (40Asp, 313Asp, 314His) or substrate (80Cys; 128Lys; 208His; 251His; 326Lys) were conserved, including 80Cys, which is post-translationally modified to form C(alpha)-formylglycine by sulfatase modifying factor 1 (SUMF1) within the active site of related sulfatase gene families [1].

The N-terminus leader peptide (1-22) revealed similarities in sequence, containing multiple hydrophilic residues. Seven N-glycosylation sites conserved for the mammalian ARSK sequences, including 108Asn, 166Asn, 193Asn, 262Asn, 375Asn, 413Asn and 498Asn, are designated as N-glycosylation sites 1-7 (Figure 1). These were found for all vertebrate ARSK sequences examined (human, chimp, gorilla, gibbon, squirrel monkey, baboon, rhesus monkey, tarsier and marmoset), although the orangutan (Pongo Abelii) ARSK sequence, lacked N-glycosylation site 1. Mouse and rat ARSK sequences also contained N-glycosylation sites 1-7, whereas the cow (Bos taurus), pig (Sus scrofa), horse (Equus caballus), sheep (Ovis aries), and dog (Canis lupus familiaris) ARSK sequences lacked N-glycosylation site 2 (data not shown). The chicken ARSK sequence contained seven N-glycosylation sites, although sites 1 and 2 were located in different positions within the protein. The zebra fish ARSK sequence contained six N-glycosylation sites, retaining sites 1, 4 and 6 as for the human sequence, but incorporating 3 other sites throughout the sequence. The predicted 3D structure for ARSK are presented in Figures 1 and 2, generated using the tertiary structures [17-20].

Multiple N-linked glycosylation sites were a feature of this enzyme ensuring that the vertebrate ARSK glycoprotein is suitably localized within the cell to perform its metabolic functions.

Vertebrate ARSK protein structures

Predictions of secondary and tertiary structures for human ARSK are presented in Figures 1 and 2, generated using the tertiary structure for a putative sulfatase from Bacteroides thetaiotaomicron (PDB:1b5qB) (modelling residue range of 35-475). A predicted human ARSK secondary structure for residues 476-536 was obtained using the SWISS-MODEL web-server [12]. The predicted 3D structure for ARSK shows a large active site cleft containing a metal (Ca2+) ion. The enzyme has 2 major domains, with the active site at the base of a cleft on the larger domain, corresponding to the N-terminal region of the enzyme, with the other domain predominantly comprising the C-terminal domain, containing a sequence of beta sheets (β12-14) and a large alpha helix (α13). The C-terminal α13 helix was not fully identified in the predicted ARSK 3D structure since the modeling range fell short of the complete C-terminal sequence, as were two other C-terminal α helices (α14 and α15). 3D structures for other human arylsulfatase sequences have been previously reported, revealing similar tertiary subunit structures, reflecting strong conservation in amino acid sequences and structures [17-20].

Human ARSK tissue expression

Figure 3 summarizes the tissues distribution profile for ARSK transcripts from human tissues. A previous study has reported a broad ARSK mRNA expression profile in human tissues suggesting that a ubiquitous biological arylsulfate substrate was available for ARSK physiologically [4], which has been recognized as lysosomal 2-sulfoglucuronate sulfate [7]. ARSK displayed highest expression levels in fibroblasts and tibial nerve cells, and modest expression levels in other tissues of the body but with very low whole blood expression levels.
Gene locations, exonic structures and regulatory sequences for vertebrate ARSK genes

The predicted chromosome and strand locations and exonic structures for vertebrate ARSK genes are presented in Table 1, together with a proposed ancestral invertebrate SUL1 gene. The predicted vertebrate ARSK genes were transcribed on the negative DNA strand (mouse, tasmanian devil (marsupial genome), chicken and frog genomes) or the positive DNA strand (human, baboon, cow and zebra fish genomes). Predicted exonic start sites for human, mouse, chicken and zebra fish ARSK genes (8 coding exons in each case) are shown in Figure 1, in identical or similar positions to those predicted for the human ARSK gene.

The human ARSK transcript was ~50 kbps in length with CpG61 and several Transcription Factor Binding Sites (TFBS) located in the 5'-untranslated promoter region of human ARSK on chromosome 5, and a 3'-untranslated region (UTR) lacking any detectable microRNA target sites (Figure 4). CpG61 contained 762 bps with a C plus G count of 463 bps, a C or G content of 61% and showed a ratio of observed to expect CpG of 0.87. This ARSK CpG Island may play a key role in gene regulation and may contribute to the broad gene expression observed in human tissues (Figure 3) [7,8,21,22]. Ten TFBS sites were colocated with CpG61 in the human ARSK promoter region which may contribute to the ubiquitous tissue expression. Of special interest among these identified ARSK TFBS were the following: SMARCA4 (2 sites), encoding transcriptional
Evolution of ARSK and other ARS sequences

A phylogenetic tree (Figure 5) was calculated by aligning vertebrate ARSK amino acid sequences, which were ‘rooted’ with a worm \((Caenorhabditis elegans)\) arylsulfatase (SUL1) sequence (Table 1). The phylogram showed clustering of the vertebrate ARSK sequences which were consistent with their evolutionary relatedness and separation of the ARSK protein group, which was distinct from the worm SUL1 sequence. It is apparent that ARSK is an ancient protein and separation of the ARSK protein group, which was distinct from the sequences which were consistent with their evolutionary relatedness (Table 1). The phylogram showed clustering of the vertebrate ARSK sequence \((C. elegans)\) arylsulfatase (SUL1) sequence. The tree is labeled with the ARSK and sulfatase-like (SUL1) name and the name of the organism with the worm \((Caenorhabditis elegans)\) SUL1 sequence. The tree was used to ‘root’ the tree. Note the major cluster corresponding to the vertebrate ARSK gene family. A genetic distance scale is shown. The number of times a clade (sequences common to a node or branch) occurred in the bootstrap replicates is shown. Replicate values of 0.9 or more, which are highly significant, are shown with 100 bootstrap replicates performed in each case.

Conclusion

Vertebrate ARSK genes and encoded proteins represent a distinct gene and protein arylsulfatase family which share key conserved sequences and domains reported for other arylsulfatase proteins previously studied \([2-5,17-20]\). The metabolic role for ARSK has recently been established and the natural substrate for this enzyme described as lysosomal 2-sulfoglucuronate, a key enzyme in the catabolism of heparin sulfate and dermatan sulfate \([8]\). A single gene \((ARSK/Arsk)\) encodes this enzyme among the vertebrate genomes studied (Table 1), which is moderately expressed in a wide range of human tissues, but with highest levels of expression in fibroblasts (Figure 4). ARSK usually contained 8 coding exons on the negative or positive strands, depending on the vertebrate genome (Table 1). The human ARSK gene contained a large CpG island within the promoter region, with several transcription factor binding sites collocated within the ARSK gene promoter region (Figure 4). Predicted secondary and tertiary structures for human ARSK showed similarities with reported 3D structures for other arylsulfatas \([17-20]\). Structural domains reported for human ARSK, included the N-terminal signal peptide; the active site (including a Ca\(^{2+}\) binding site), which is responsible for arylsulfatase activity; and seven predominantly conserved N-glycosylation sites, which ensured that ARSK was suitably micro localized within the cell as a glycoprotein. Phylogenetic studies suggested that the ARSK gene appeared early in evolution, prior to the appearance of bony fish.

Acknowledgements

I thank Dr. Laura Cox from the Texas Biomedical Research Institute for useful discussions.

Figure 4: Human ARSK gene: structure and major transcript. Derived from the AceView website http://www.ncbi.nlm.nih.gov/IEB/Research/Acembly/ [11]: shown with capped 5' and 3' ends for the predicted mRNA sequences; NM refers to the NCBI reference sequence; coding exons are in pink; the direction for transcription is shown as 5' → 3'; a large CpG61 island at the gene promoter is shown; predicted Transcription Factor Binding Sites (TFBS) for the human ARSK gene promoter.

Figure 5: Vertebrate ARSK phylogenetic tree with the \(C. elegans\) SUL1 sequence. The tree is labeled with the ARSK and sulfatase-like (SUL1) name and the name of the organism with the worm \((Caenorhabditis elegans)\) SUL1 sequence, which was used to ‘root’ the tree. Note the major cluster corresponding to the vertebrate ARSK gene family. A genetic distance scale is shown. The number of times a clade (sequences common to a node or branch) occurred in the bootstrap replicates are shown. Replicate values of 0.9 or more, which are highly significant, are shown with 100 bootstrap replicates performed in each case.
References


