BMC Evolutionary Biology

RESEARCH ARTICLE Open Access

Insights into the evolution of sorbitol metabolism: phylogenetic analysis of SDR196C family

Agustín Sola-Carvajal, María I García-García, Francisco García-Carmona and Álvaro Sánchez-Ferrer*

Abstract

Background: Short chain dehydrogenases/reductases (SDR) are NAD(P)(H)-dependent oxidoreductases with a highly conserved 3D structure and of an early origin, which has allowed them to diverge into several families and enzymatic activities. The SDR196C family (http://www.sdr-enzymes.org) groups bacterial sorbitol dehydrogenases (SDH), which are of great industrial interest. In this study, we examine the phylogenetic relationship between the members of this family, and based on the findings and some sequence conserved blocks, a new and a more accurate classification is proposed.

Results: The distribution of the 66 bacterial SDH species analyzed was limited to Gram-negative bacteria. Six different bacterial families were found, encompassing α-, β- and γ-proteobacteria. This broad distribution in terms of bacteria and niches agrees with that of SDR, which are found in all forms of life. A cluster analysis of sorbitol dehydrogenase revealed different types of gene organization, although with a common pattern in which the SDH gene is surrounded by sugar ABC transporter proteins, another SDR, a kinase, and several gene regulators. According to the obtained trees, six different lineages and three sublineages can be discerned. The phylogenetic analysis also suggested two different origins for SDH in β-proteobacteria and four origins for γ-proteobacteria. Finally, this subdivision was further confirmed by the differences observed in the sequence of the conserved blocks described for SDR and some specific blocks of SDH, and by a functional divergence analysis, which made it possible to establish new consensus sequences and specific fingerprints for the lineages and sub lineages.

Conclusion: SDH distribution agrees with that observed for SDR, indicating the importance of the polyol metabolism, as an alternative source of carbon and energy. The phylogenetic analysis pointed to six clearly defined lineages and three sub lineages, and great variability in the origin of this gene, despite its well conserved 3D structure. This suggests that SDH are very old and emerged early during the evolution. This study also opens up a new and more accurate classification of SDR196C family, introducing two numbers at the end of the family name, which indicate the lineage and the sublineage of each member, i.e, SDR196C6.3.

Background

The short-chain dehydrogenase/reductase (SDR) superfamily consists of NAD(P)(H)-dependent oxidoreductases that are distinct from the medium-chain dehydrogenase/reductase (MDR) and aldo-keto reductase (AKR) superfamilies [1]. Our knowledge of these superfamilies initially emerged from observations made concerning alcohol dehydrogenases of *Drosophila* and mammalian liver, which were seen to be clearly different [2,3]. Insect and bacterial alcohol and polyol dehydrogenases initially received less attention, since these enzymes were found to be different, and were only considered to be of prokaryotic and lower eukaryotic origin [2,3]. However, the discovery of similarities between these latter enzymes and human or mammalian prostaglandin, hydroxyl-steroid and other dehydrogenases, changed that view dramatically [4-7]. In addition, in recent years, interest in SDR enzymes has increased, since they are useful in biotechnological and analytical processes. About 25% of all dehydrogenases belong to the SDR superfamily [8].

Common to all types of oxidoreductases is the occurrence of a Rossmann-fold dinucleotide cofactor-binding motif, which has been found to be one of the most common protein folds [9,10]. Among SDR, no high sequence

* Correspondence: alvaro@um.es
Department of Biochemistry and Molecular Biology-A, Faculty of Biology, Regional Campus of International Excellence "Campus Mare Nostrum", University of Murcia, Campus Espinardo, Murcia E-30100, Spain

identity between different members is observed (about 20–30%), but all of them display a highly similar 3D structure, typically folding into a simple one-domain architecture with distinct conserved motifs, including the cofactor binding site at the N-terminal, structure stabilizing motifs, the active center, catalysis-enhancing sites and the substrate binding site, located in the highly variable C-terminal region [1,11]. Such a degree of 3D structure conservation indicates that ancestral dehydrogenases existed within each SDR family, and after multiple events, these ancestral dehydrogenases gave rise to the present system of subfamilies and classes found within each family [1].

Given the early origin of SDR, the subsequent divergence has had time to become quite pronounced. Hundreds of SDR enzyme activities and their corresponding families have been detected. Based on the similar coenzyme-binding structure, their active-site relationship and repetitive patterns, five SDR superfamily types have been discerned from different data banks, named as "classical", "extended", "intermediate", "divergent" and "complex" SDR enzymes [12]. This divergence also includes different enzymatic activities, most of them dehydrogenases or reductases, but also lyases and some isomerases. The active-site Tyr residue, assisted by adjacent Lys, Asn and Ser residues, has been found to fit to the basic reaction mechanism in most cases, but also to reflect acid–base catalysis and proton transfers [1]. Thus, SDR proteins not only have a very distant origin, including a viral representation [1,13], presumably from a time when virally-mediated lateral gene transfer commonly occurred [14], but also show a wide range of activities, involving half of all enzyme activity types. Few gene/protein superfamilies exhibit this great divergence.

Recently, a sustainable and expandable nomenclature SDR database has been proposed, based on hidden Markov models (http://www.sdr-enzymes.org) [8,15]. This database has identified 314 SDR families, encompassing about 31,900 members [8]. Among them, the SDR196C family (http://www.sdr-enzymes.org/search) groups bacterial sorbitol dehydrogenases (L-iditol NAD + oxidoreductases, EC 1.1.1.14, SDH), which are of industrial interest for the specific determination of sorbitol (D-glucitol), a natural acyclic polyol found in food, and in pharmaceutical and cosmetic preparations [16].

In this study, we provide a comprehensive insight into the distribution, diversity, evolution and classification of the SDR196C superfamily in bacteria. The phylogenetic analysis revealed different lineages related to some sequence differences in the conserved SDR motifs and in the characteristic SDH blocks, allowing, for the first time, the classification of this SDR family (SDH family) into 6 different lineages and three sub lineages. This could permit a more efficient data curation, and a new nomenclature for the classification of incoming sequences into the SDR196C family.

Results and discussion

Distribution of SDH gene

The SDR database (http://www.sdr-enzymes.org) is a sustainable and expandable nomenclature database [8,15], which includes 127 bacterial sorbitol dehydrogenases (among characterized and putative) within the SDR196C family. Some identical sequences have been included two or more times in the database, representing different strains. In order to simplify the study, only one strain of these species was included in the analysis. The SDH gene was found in 66 bacterial species, all of them Gram-negative belonging to alpha (α)-, beta (β)- and gamma (γ)-Proteobacteria (see Additional file 1), with both low and high GC representatives (see Additional file 2). This distribution of SDH agrees with the prevalence of SDR enzymes, which are present in all forms of life [17] and, according to recent data from random genome screenings of microorganisms and viruses from sea water, are also among the most abundant genes in nature [13].

α-Proteobacteria, with six families, is the largest group with a total of 41 species, including members of the Acetobacteriace, Rhodospirillaceae, Rhodobacteriaceae, Brucellaceae, Phyllobacteriaceae and Rhizobiaceae families (see Additional file 1). Two of these species, *Ochrobactrum anthropi* and *O. intermedium* (members of the Brucellaceae family) are human pathogens that cause septicemia [18,19]. The Acetobacteriaceae family is represented by two members, *Acidiphilium cryptum* and *Gluconoacetobacter hansenii,* which are common in vinegar and used as iron contamination indicators [20]. Members of the Rhodospirilaceae, Phyllobacteriaceae and Rhizobiaceae families are usually nitrogen-fixing microorganisms found in soil and aquatic habitats, with the exception of *Agrobacterium tumefaciens*, *A. radiobacter* and *A. vitis* from the Rhizobiaceae family, which are well known plant pathogens, causing tumors. Rhodobacteriaceae is the most numerous family, with 22 members, including sea water microorganisms, photosynthetic bacteria (*Rhodobacter* sp., *R. capsulatus* and *R. sphaeroides* [21]) and two extremophiles (*Paracoccus denitrificans* and *Silicibacter lacuscaerulensis).*

The β-Proteobacteria group includes 15 members of the Burkholderiaceae family and 2 of the Commamonadaceae family (see Additional file 1). These two families are composed of soil and free-living microorganisms, which are usually nitrogen-fixing (*Burkholderia phymatum* and *B. xenovorans*) or symbionts (*B. graminis*, *B. phytofirmans*, *B. thailandensis*, *Acidovorax avenae* and *Variovorax paradoxus*). The Burkholderiaceae family also includes a commensal of the earthworm nephridia

(*Verminephrobacter eiseniae)*, a plant pathogen (*Ralstonia solanacearum*) and human pathogens from genera *Burkholderia,* which cause opportunistic infections in diseases, such as cystic fibrosis [22].

The γ-Proteobacteria group, which includes only 7 species from three different bacterial families, is the smallest group among the SDR196C family. Aquatic bacteria from the halophilic family Halomonadaceae (*Chromohalobacter salexigensis* and *Halomonas elongata*) and from the marine family Oceanospirillaceae (*Marinomonas* sp.) form part of this group. The third family of this group (Pseudomonaceae) includes a plant commensal (*Pseudomonas fluorescens)*, a plant pathogen (*P. savastanoi)* which cause olive knot [23], a saprophyte (*P. syringae),* and a soil bacterium (*Pseudomonas* sp.). Taken together, these results indicate that the SDH gene is widely distributed in nature, and is an important enzyme in the polyol metabolism of some bacteria, to use sorbitol as an alternative source of carbon and energy [21].

Genetic organization of SDH gene

The three main bacterial groups that contain the SDH gene, α-, β- and γ-proteobacteria, encode this gene in different polyol clusters with a different gene order (see Additional file 3). In general terms, the SDH gene in the polyol operon is usually surrounded by a transporter (mainly, ATP Binding Cassette [ABC] transporter, which translocate substrates across membrane *via* ATP hydrolysis), an SDR protein (normally, a mannitol dehydrogenase, MDH) and a sugar related kinase, such as ribitol kinase. However, the companion genes and the order in the cluster, vary between bacterial families, and, to a lesser extent, within families (see Additional file 3). Only one overall organization of SDH genes has been described for the polyol operon of *Rhodobacter sphaeroides* Si4, where *smoS* gene encodes for an SDH, *smoK* for an ABC transporter, and *mtlK* for a mannitol dehydrogenase [24].

Within the α-Proteobacteria, there are twelve variants of the SDH cluster, each family having its own gene order. The Rhodobacteriaceae family shows three variants of this polyol operon, in which a cluster of four genes related with ABC transporter are on one side of the SDH gene, and a dehydrogenase (mannitol or alcohol dehydrogenase) plus an extra gene (HAD, tRNA or FeoA protein) on the other side (see Additional file 3, variants 1–3). The Rhizobiaceae family has more diversity in its polyol cluster, which displays 5 different variants (see Additional file 3, variants 4–8), but still shows the pattern of at least three ABC genes on one side of the SDH gene, except for *Agrobacterium tumefaciens* and *Rhizobium etli,* in which two ABC genes are replaced by two sugar kinase genes (fructose kinase and tagatose 6-phosphate kinase) (see Additional file 3, variant 6). On the other side of the SDH gene, a dehydrogenase (MDH or a Zn^{2+} binding dehydrogenase) is also present, except for *Agrobacterium radiobacter,* which presents an *AraC* regulator, followed by metal-accepting chemotaxis sensory transducer MACST (see Additional file 3, variant 5). This microorganism also has a *LysR* gene, indicating a tight regulation of the SDH related genes in order to use this sugar and to control its metabolism under adverse conditions. This control is also seen in *A. tumefaciens* and *R. etli* (see Additional file 3, variant 6), with the presence of *LacI.*

The rest of the α-Proteobacteria families (Phyllobacteriaceae, Rhodospirillaceae, Acetobacteriaceae and Brucellaceae) have the common pattern of at least three ABC proteins on one side, but are more diverse on the other side, having not only kinases (hexokinase or fructose kinase) but also two singular enzymes in SDH clusters, which are related with phosphogluconate (2-dehydro-3-deoxyphosphogluconate aldolase and phosphogluconate dehydrogenase) (see Additional file 3, variant 11), or two consecutive dehydrogenases (mannitol dehydrogenase and alcohol dehydrogenase; see Additional file 3, variant 12). The *LysR* gene is also present, except in the Rhodospirillaceae family.

Among the β-proteobacteria group, the Commamonadaceae family has its own order, but with the presence of an intercalating MDH gene between SDH and ABC transporter genes (see Additional file 3, variant 13). The latter order is also similar to that of *Verminephrobacter eiseniae,* a member of the Burkholderiaceae family (Additional file 3, variant 14), which is in fact, quite different from the common pattern displayed for the rest of the Burkholderiaceae members (see Additional file 3, variants 15–17). The latter clusters show a ferric uptake regulator gene and its corresponding cation ABC transporter genes on one side, and sorbitol/mannitol ABC transport genes followed by HAD gene on the SDH gene side (see Additional file 3, variant 15), and sometimes interrupted by two sugar related genes (ribokinase and tagatose 1,6-biphosphate aldolase, variant 16; or 2-keto-3-deoxygluconate kinase and tagatose 6-phosphate kinase, see Additional file 3, variant 17).

Finally, the γ-Proteobacteria group has no specific pattern, and it is easy to differentiate the Halomonadaceae family (see Additional file 3, variants 18–19), with an haloacid dehalogenase gene intercalating the SDH and three ABC transporter proteins, from the Pseudomonadaceae family (see Additional file 3, variants 20–21), in which only one (or no) ABC transporter gene is present, together with an *AraC* gene. This latter family also lacks the second SDR gene, indicating that the SDH gene is not close to other the sugar-utilizing genes as it is in all of the above described families.

Signatures of horizontal gene transfer

Basically, there are two main methods to identify putative horizontal gene transfer events, phylogenetic methods and surrogate methods based on nucleotide composition. Also, the presence of transposases and/or integrases within a region may suggest another mode of transfer. However, no such enzymes genes were found in the proximity of any of the polyol clusters described above. On the other hand, the differences between the average GC content of whole genome (GC_g) and the GC content of the SDH genes (GC_{SDH}) (see Additional file 2), showed only one unknown possible horizontal gene transfer event (deviation from GC_g by +/− 5) in *Gluconoacetobacter hansenii* with a GC difference of −5.8 (see Additional file 2). In addition, *Loktanella vestfoldensis*, *Agrobacterium radiobacter*, *Chromohalobacter salexigensis*, *Pseudomonas savastanoi* and *Oceanicola batsensis* have a high, but not significant, GC difference (≈ +/− 4), which suggest the possibility of horizontal gene transfer for SDH in these species, too.

Phylogenetic analysis of SDH gene

In order to examine further the evolutionary history of the SDH gene, a multiple sequence alignment (MSA) was carried out with GUIDANCE [25,26], using the MSA algorithm PRANK [27], which gave an overall quality assessment exceeding 0.97 (1 corresponds to 100% certainty) (see Additional file 4). The phylogenetic analysis and the topology obtained were compared with that found for the species tree based on 16S rRNA sequences aligned with the above algorithm (Figure 1 and 2, respectively; see also Additional file 4). Phylogenetic analyses of SDH amino acid sequences resulted in a well-resolved tree, which was quite similar, regardless of the method used (see Additional file 5). Overall, the SDH genes in the three proteobacteria groups studied did not form three distinct lineages (Figure 1) as it does, in the 16S rRNA tree (Figure 2, see also Additional file 6). Indeed, the SDH tree could be subdivided into six main lineages (named 1, 2, 3, 4, 5 and 6) (Figure 1), lineage 6 being the most divergent, encompassing α- and γ-proteobacteria from five different families (Figure 1). Also within this lineage, three sublineages (Figure 3, see also Additional file 4) were found, all with a common origin. Lineages 6.1 and 6.2 were formed by species of the Rhodobacteriaceae family, except *Hoeflea phototrophobica*, which is member of the Phyllobacteriaceae family. Lineage 6.3 was the most divergent group, with members of α- and γ-proteobacteria from five different families: Phyllobacteriaceae, Pseudomonaceae, Rhodobacteriaceae, Rhodospirillaceae and Rhizobiaceae (Figure 3). This lineage 6 is also grouped the most SDH from the Rhodobacteriaceae family, except *R. bacterium* and *Thalassiobium* sp. Interestingly, the three members of genus *Rhodobacter* did not group as closely as might be expected, *R. capsulatus* being a member of lineage 6.1 and *R. sphaeroides* and *Rhodobacter* sp. members of lineage 6.3. This indicates a common ancestor, with a divergence in the time of SDH gene acquisition. Similarly, Pseudomonaceae members were grouped in lineage 6, except for *Pseudomonas* sp., which belongs to lineage 1, together with all members of Burkholderiaceae family (Figure 1 and 3). The distribution of the SDH from different *Pseudomonas* species in divergent branches of the tree (Figure 1) indicates that in these species, the SDH genes were acquired many times and from different sources.

Lineage 1 contained most of the members of β-Proteobacteria family, except for the Commamonadaceae representatives, which branches away in lineage 2. In addition, lineage 1 branches firmly away from the rest of the lineages, suggesting that the origin of SDH in lineage 1 is unique, and that there are two different origins for β-proteobacteria SDH. Lineage 2 is also a divergent group which includes, apart from Commamonadaceae SDH proteins, members of α- and γ-Proteobacteria from the Rhodobacteriaceae and Oceanospirillaceae families. The presence of members of the Rhodobacteriaceae family (*Rhodobacterales bacterium* and *Thalassiobium* sp) clearly separated from lineage 6, indicated the possibility of a horizontal gene transfer event, although this is not supported by the GC difference, with values of −1.6 and 2.6, respectively, or by the presence of transposases/integrases within its polyol cluster.

Lineages 3, 4 and 5 are well separated groups, of different but close origin. Members of Halomonadaceae (lineage 3), Acetobacteriaceae (lineage 4), Brucellaceae and Rhizobiaceae (lineage 5) families are included in these lineages. *Agrobacterium tumefaciens* and *Rhizobium etli* did not group with the rest of their family in lineage 6, indicating a divergent origin of SDH in this family. Interestingly, γ-proteobacteria had at least one member in four of the six lineages described (Lineages 1, 2, 3 and 6), which suggests a divergent origin of the SDH gene among γ-proteobacteria.

This widespread and variable origin of SDH detailed here is related to the distribution and evolution of SDR, which were mentioned above, occurs in all kingdoms of life [17]. However, this variability is not observed in the structure of these enzymes in all six lineages (see Additional file 7), which all share a common Rossmann-fold motif for dinucleotide cofactor binding, and a substrate binding site in the highly variable C-terminal region [1]. This variability in the distribution and the homogeneity in the structure, together with the recombinatorial formation of the catalytic subunit from building blocks, suggest that SDR, and consequently SDH, emerged early from α/β elements to a form a Rossman-fold domain in

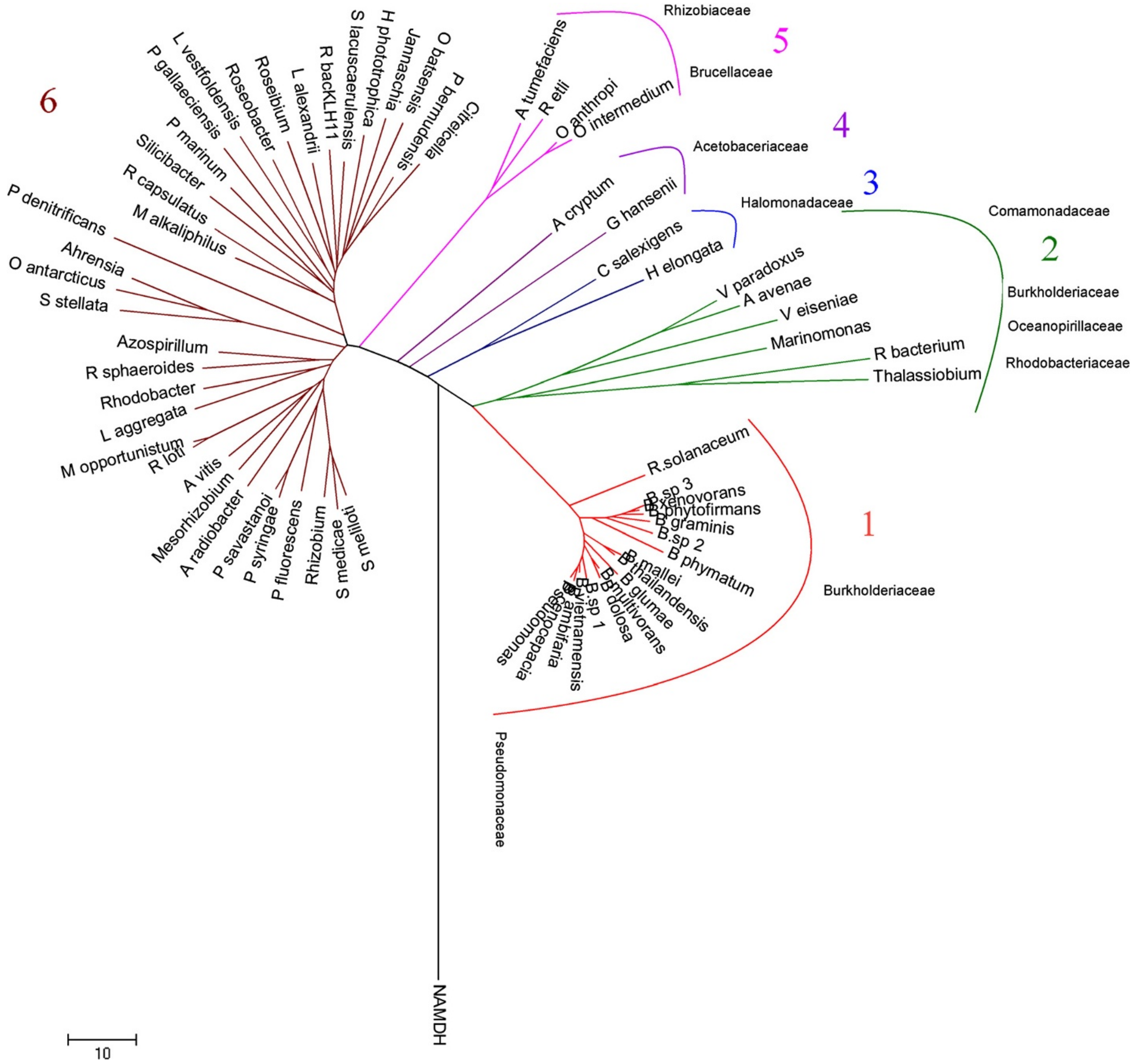


Figure 1 Phylogenetic tree of all species encoding SDH gene. The tree was obtained using Neighbor Joining (NJ) analysis in MEGA 4.0 [45]. 1,000 generations were used to build the consensus tree, as indicated in the methods section. The main inclusive taxonomic groups are indicated. The bacteria used are listed in Additional file 1.

the universal cellular ancestor prior to Darwinian evolution in cells of all kingdoms of life [17,28].

Sequence comparison of SDH lineages

With the aim of further exploring the evolution of SDH proteins, 11 characteristic SDH blocks (I-XI) were described using the sequence alignment of the six different lineages and the 3 sublineages of lineage 6. Overall, differences in the blocks between the six lineages were well defined, and it was possible to establish different consensus sequences for each lineage with Guidance scores above 0.97 (Figure 4), which supports the classification obtained from the phylogenetic analysis. Lineages 2 and 6 were the most variable of the lineages (see Additional file 8), as might be expected according to the phylogenetic analysis (Figure 1), although it was still possible to establish consensus sequences. The eleven SDH conserved blocks were also observed in the three sublineages of lineage 6 with Guidance scores above 0.97 (Figure 5).

Sequence alignment of lineage 1 showed highly conserved blocks (see Additional file 8) among its members.

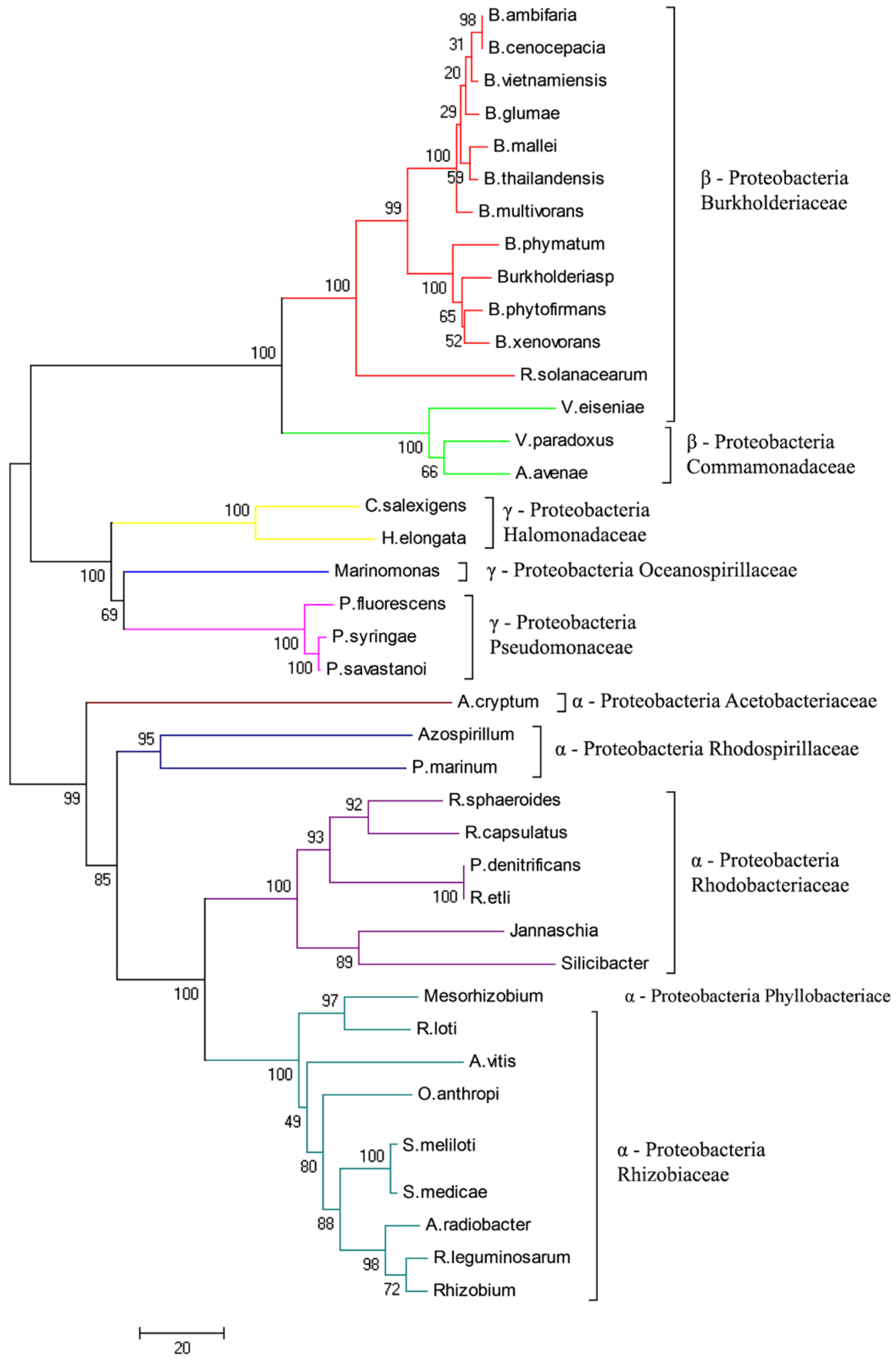


Figure 2 Phylogenetic tree of 16S rRNA of all families encoding SDH gene. Sequences for the analysis were obtained from GeneBank. The analysis was performed using the Neighbor Joining method, and the tree was implemented with MEGA 4.0 [45] after 1,000 generations. Families encoding SDH gene are indicated.

This high degree of sequence similarity was in agreement with that described in the phylogenetic analysis, since lineage 1 was basically composed of members of the Burkholderiaceae family, except for *Pseudomonas* sp. The specific block sequence for this lineage is indicated in Figure 4, and, interestingly, the sequence of block I (GEAVA), which is involved in NAD$^+$ binding [29-31] and the sequence of block X (DLTGA), which is related to NAD stabilization and tetramer formation, can be considered as fingerprints for this lineage, since these

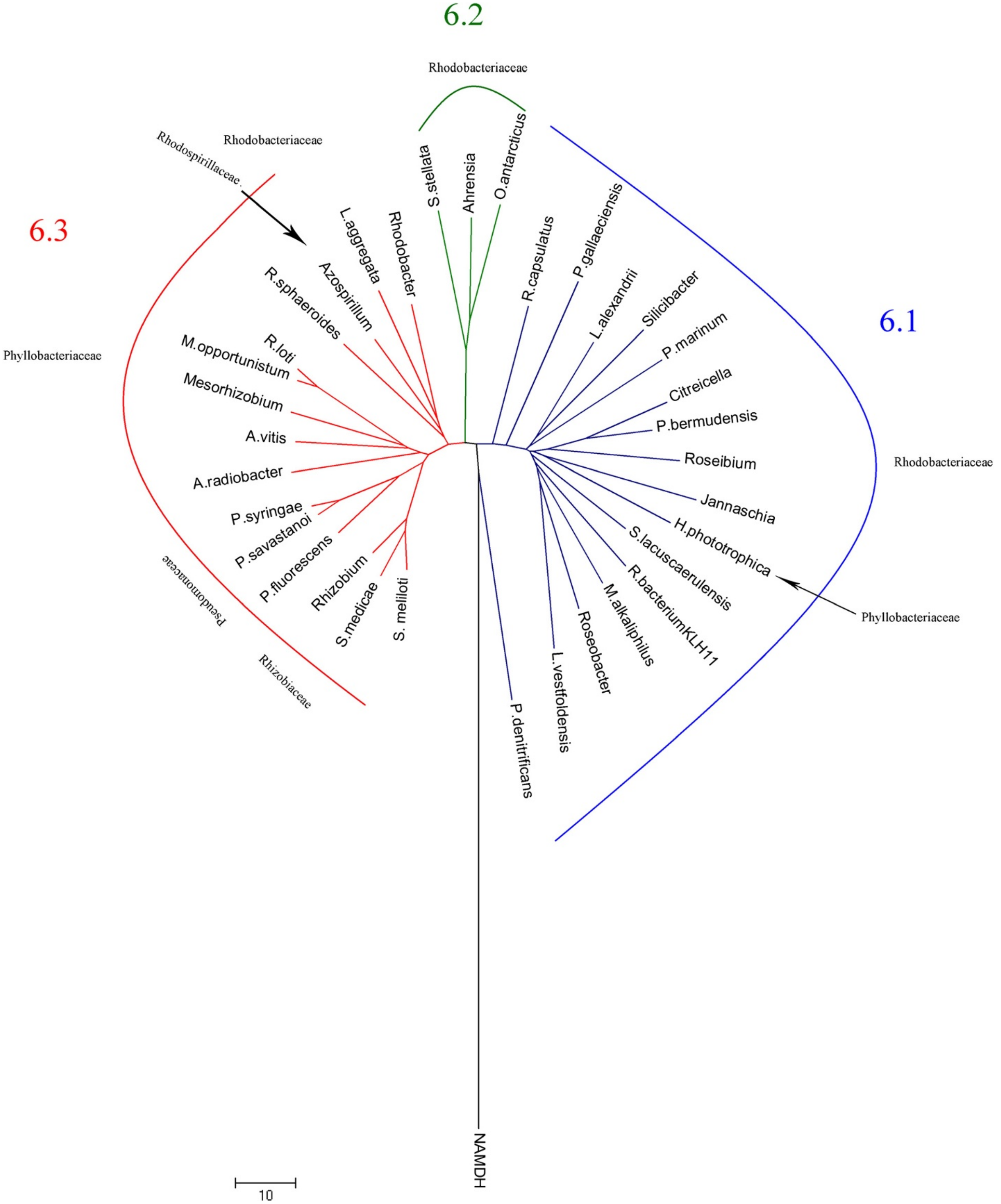
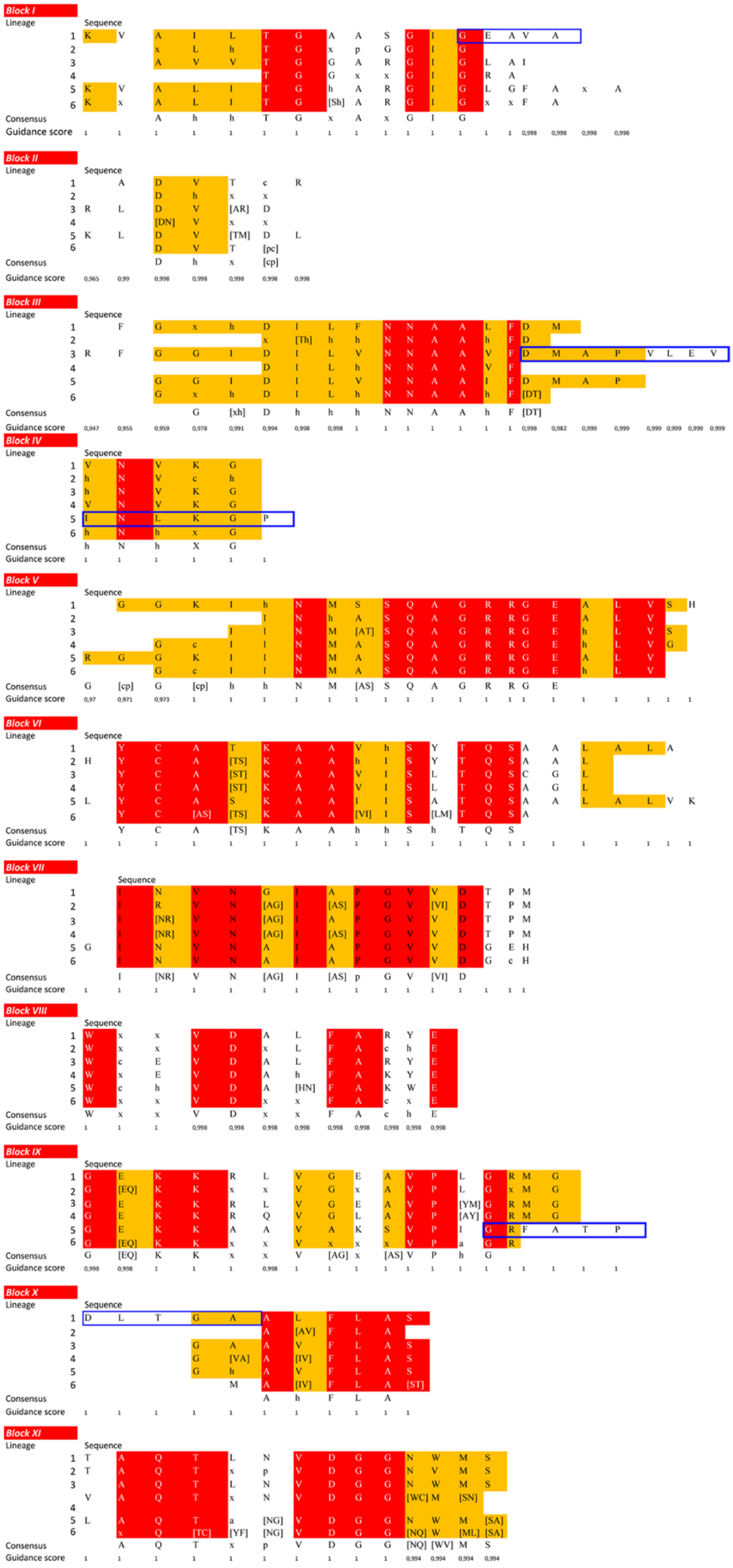


Figure 3 Phylogenetic tree of species forming lineage 6 of SDH. The tree was obtained using Neighbor Joining (NJ) analysis in MEGA 4.0 [45]. 1,000 generations were used to build the consensus tree, as indicated in the methods section. Main inclusive taxonomic groups are indicated.



Block I

Lineage	Sequence																		
1	K	V	A	I	L	T	G	A	A	S	G	I	G	E	A	V	A		
2			x	L	h	T	G	x	p	G	G	I	G						
3			A	V	V	T	G	G	A	R	G	I	G	L	A	I			
4						T	G	G	x	x	G	I	G	R	A				
5	K	V	A	L	I	T	G	h	A	R	G	I	G	L	G	F	A	x	A
6	K	x	A	L	I	T	G	[Sh]	A	R	G	I	G	x	x	F	A		
Consensus			A	h	h	T	G	x	A	x	G	I	G						
Guidance score	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0,998	0,998	0,998	0,998

Block II

Lineage	Sequence						
1		A	D	V	T	c	R
2			D	h	x	x	
3	R	L	D	V	[AR]	D	
4			[DN]	V	x	x	
5	K	L	D	V	[TM]	D	L
6			D	V	T	[pc]	
Consensus			D	h	x	[cp]	
Guidance score	0,965	0,99	0,998	0,998	0,998	0,998	0,998

Block III

Lineage	Sequence																						
1		F	G	x	h	D	I	L	F	N	N	A	A	L	F	D	M						
2						x	[Th]	h	h	N	N	A	A	h	F	D							
3	R	F	G	G	I	D	I	L	V	N	N	A	A	V	F	D	M	A	P	V	L	E	V
4						D	I	L	h	N	N	A	A	V	F								
5			G	G	I	D	I	L	V	N	N	A	A	I	F	D	M	A	P				
6			G	x	h	D	I	L	h	N	N	A	A	h	F	[DT]							
Consensus				G	[xh]	D	h	h	h	N	N	A	A	h	F	[DT]							
Guidance score	0,947	0,955	0,959	0,978	0,991	0,994	0,998	0,998	1	1	1	1	1	1	1	0,998	0,982	0,999	0,999	0,999	0,999	0,999	0,999

Block IV

Lineage	Sequence					
1	V	N	V	K	G	
2	h	N	V	c	h	
3	h	N	V	K	G	
4	V	N	V	K	G	
5	I	N	L	K	G	P
6	h	N	h	x	G	
Consensus	h	N	h	X	G	
Guidance score	1	1	1	1	1	1

Block V

Lineage	Sequence																					
1		G	G	K	I	h	N	M	S	S	Q	A	G	R	R	G	E	A	L	V	S	H
2						I	N	h	A	S	Q	A	G	R	R	G	E	A	L	V		
3					I	I	N	M	[AT]	S	Q	A	G	R	R	G	E	h	L	V	S	
4			G	c	I	I	N	M	A	S	Q	A	G	R	R	G	E	h	L	V	G	
5	R	G	G	K	I	I	N	M	A	S	Q	A	G	R	R	G	E	A	L	V		
6			G	c	I	I	N	M	A	S	Q	A	G	R	R	G	E	h	L	V		
Consensus	G	[cp]	G	[cp]	h	h	N	M	[AS]	S	Q	A	G	R	R	G	E					
Guidance score	0,97	0,971	0,973	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Block VI

Lineage	Sequence																					
1		Y	C	A	T	K	A	A	V	h	S	Y	T	Q	S	A	A	L	A	L	A	
2	H	Y	C	A	[TS]	K	A	A	h	I	S	Y	T	Q	S	A	A	L				
3		Y	C	A	[ST]	K	A	A	V	I	S	L	T	Q	S	C	G	L				
4		Y	C	A	[ST]	K	A	A	V	I	S	L	T	Q	S	A	G	L				
5	L	Y	C	A	S	K	A	A	I	I	S	A	T	Q	S	A	A	L	A	L	V	K
6		Y	C	[AS]	[TS]	K	A	A	[VI]	I	S	[LM]	T	Q	S	A						
Consensus		Y	C	A	[TS]	K	A	A	h	h	S	h	T	Q	S							
Guidance score	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Block VII

Lineage		Sequence														
1		I	N	V	N	G	I	A	P	G	V	V	D	T	P	M
2		I	R	V	N	[AG]	I	[AS]	P	G	V	[VI]	D	T	P	M
3		I	[NR]	V	N	[AG]	I	A	P	G	V	V	D	T	P	M
4		I	[NR]	V	N	[AG]	I	[AS]	P	G	V	V	D	T	P	M
5	G	I	N	V	N	A	I	A	P	G	V	V	D	G	E	H
6		I	N	V	N	A	I	A	P	G	V	V	D	G	c	H
Consensus		I	[NR]	V	N	[AG]	I	[AS]	p	G	V	[VI]	D			
Guidance score	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Block VIII

Lineage	Sequence											
1	W	x	x	V	D	A	L	F	A	R	Y	E
2	W	x	x	V	D	x	L	F	A	c	h	E
3	W	c	E	V	D	A	L	F	A	R	Y	E
4	W	x	E	V	D	A	h	F	A	K	Y	E
5	W	c	h	V	D	A	[HN]	F	A	K	W	E
6	W	x	x	V	D	x	x	F	A	c	x	E
Consensus	W	x	x	V	D	x	x	F	A	c	h	E
Guidance score	1	1	1	0,998	0,998	0,998	0,998	0,998	0,998	0,998	0,998	0,998

Block IX

Lineage	Sequence																		
1	G	E	K	K	R	L	V	G	E	A	V	P	L	G	R	M	G		
2	G	[EQ]	K	K	x	x	V	G	x	A	V	P	L	G	x	M	G		
3	G	E	K	K	R	L	V	G	E	A	V	P	[YM]	G	R	M	G		
4	G	E	K	K	R	Q	V	G	L	A	V	P	[AY]	G	R	M	G		
5	G	E	K	K	A	A	V	A	K	S	V	P	I	G	R	F	A	T	P
6	G	[EQ]	K	K	x	x	V	x	x	x	V	P	a	G	R				
Consensus	G	[EQ]	K	K	x	x	V	[AG]	x	[AS]	V	P	h	G					
Guidance score	0,998	0,998	1	1	1	0,998	1	1	1	1	1	1	1	1	1	1	1	1	1

Block X

Lineage	Sequence										
1	D	L	T	G	A	A	L	F	L	A	S
2						A	[AV]	F	L	A	
3				G	A	A	V	F	L	A	S
4				G	[VA]	A	[IV]	F	L	A	S
5				G	h	A	V	F	L	A	S
6					M	A	[IV]	F	L	A	[ST]
Consensus						A	h	F	L	A	
Guidance score	1	1	1	1	1	1	1	1	1	1	1

Block XI

Lineage	Sequence													
1	T	A	Q	T	L	N	V	D	G	G	N	W	M	S
2	T	A	Q	T	x	p	V	D	G	G	N	V	M	S
3		A	Q	T	L	N	V	D	G	G	N	W	M	S
	V	A	Q	T	x	N	V	D	G	G	[WC]	M	[SN]	
4														
5	L	A	Q	T	a	[NG]	V	D	G	G	N	W	M	[SA]
6		x	Q	[TC]	[YF]	[NG]	V	D	G	G	[NQ]	W	[ML]	[SA]
Consensus		A	Q	T	x	p	V	D	G	G	[NQ]	[WV]	M	S
Guidance score	1	1	1	1	1	1	1	1	1	1	0,994	0,994	0,994	0,994

Figure 4 (See legend on next page.)

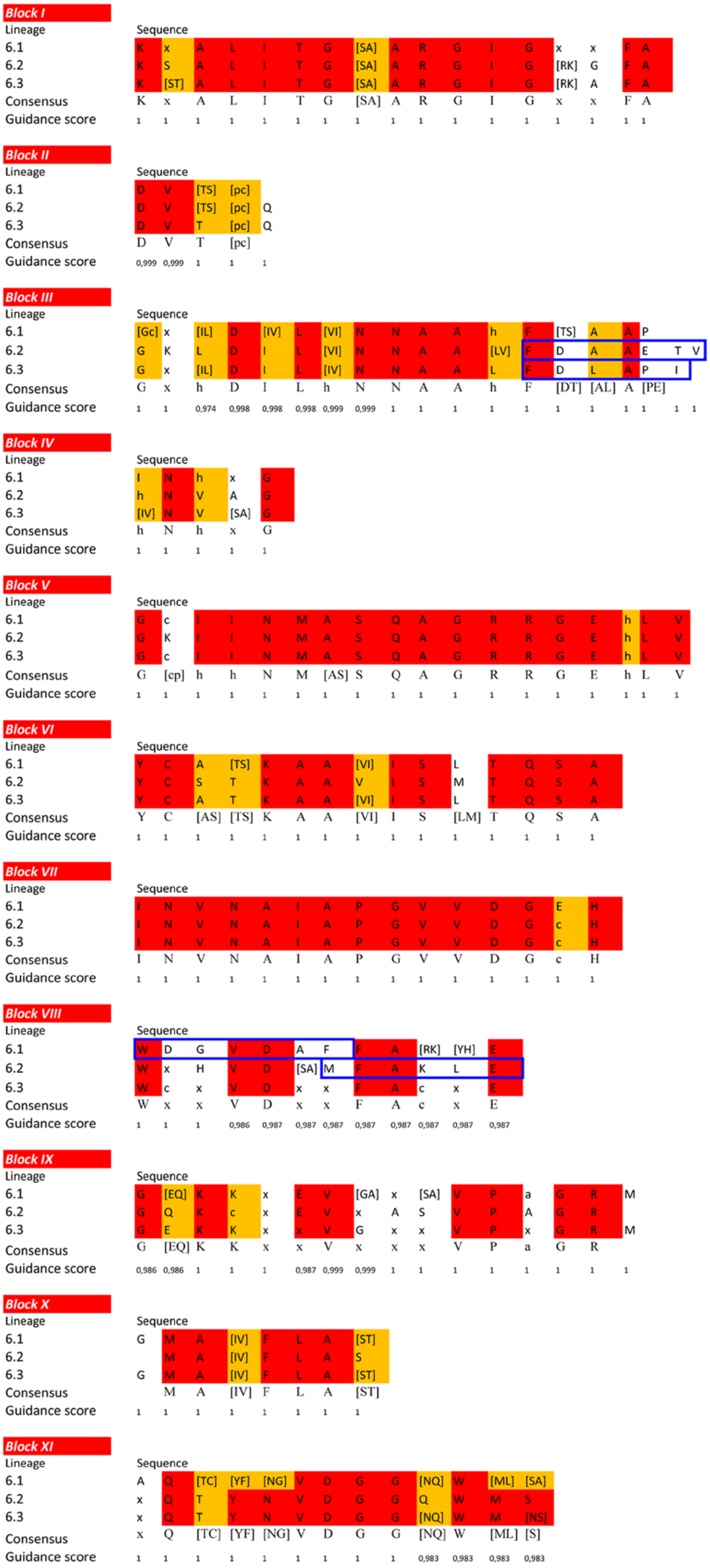
(See figure on previous page.)
Figure 4 Conserved SDH blocks in the six different lineages. "c", a charged residue; "h", a hydrophobic residue; "p", a polar residue and "x", any residue. Alternative amino acids at a given position are shown within brackets. Red background indicates strictly conserved amino acids, orange background indicates conserved amino acids and blue boxes indicate the specific blocks of each lineage. Guidance scores represent the degree of confidently aligned residues (1 corresponds to 100% certainty) [25,26].

sequences were only present in this lineage. Lineage 2, as described above, was highly divergent, as also shown in the sequence alignment (see Additional file 8), where only the blocks corresponding to the characteristic fingerprints for SDH (block V to VII) and some parts of the C-ter are conserved. Block XI and its sequence NVMS could be considered as its fingerprint. Lineage 3, which comprises only two members, showed a high sequence identity (see Additional file 8). Its long and conserved block III, whose function is to stabilize the central β-sheet [29], could be considered as its fingerprint, ending in the conserved sequence DMAPVLEV (Figure 4). The most notable feature of lineage 4 alignment is the absence of block I in *G. hansenii*, which is involved in cofactor recognition (see Additional file 8). Taking this into account, its assignation as an SDH or even as an SDR might be erroneous, although the remaining conserved domains of both SDR and SDH, including the catalytic tetrad N-Y-S-K, responsible of the oxidoreductase activity [29], are present. Thus, the absence of block I is an error which arises from wrongly determining the starting point of sequence. Lineage 5 was also highly conserved (see Additional file 8), and showed the eleven blocks of SDH with high identity. Specific fingerprints of this lineage are located in block IV (INLKGP), and at the end of block IX, which in the other lineages usually ends with GRMG, while in lineage 5 it ends in FATP (Figure 4). The heterogeneity of lineage 6 makes it difficult to establish a fingerprint, except for the consensus sequence of block V (see Additional file 8). However, when the three sublineages were aligned independently, a high degree of conservation was observed (Figure 5). Lineage 6.1 was the most divergent, showing identity only in the eleven conserved blocks (see Additional file 8). Although consensus sequences are shown for every block, only a specific sequence for this sublineage was observed at the beginning of block VIII (WDGVDAF, Figure 5). Similarly, in lineage 6.2, specific fingerprints could be found in block III (FAAAETV), and at the end of block VIII (MFAKLE). Lineage 6.3 had its own fingerprint at the end of block III (FDLAPI).

In addition, some lineages also presented highly conserved sequences between conserved blocks. Thus, lineages 1, 3 and 5 showed such a sequence between motifs I and II (CVLVD, EAGRV and EGAcFcIADI, respectively), and another between block IV and V in lineages 1, 3, 4 and 5 (FFLMQAVA, FFTLQAVAA, QAVAxQMI, and FMMKAVSNVMI, respectively). These latter residues are part of the large $\alpha_{5,}$ which is part of the sorbitol binding domain in SDH. Such a degree of conservation and the 3D proximity to blocks IV and V, which are part of the active site, suggests a role for these interblock sequences in making up the correct structure of the active site or the substrate binding site.

To expand the above analysis, a study of SDH family functional divergence was carried out to detect amino acid sites that have varying evolutionary conservation among member genes, using DIVERGE (DetectIng Variability in Evolutionary Rates among Genes) software [32,33]. The analysis grouped the amino acids residues responsible for altered functional constraints into two categories: (I) conserved in the first lineage, but variable in the second lineage; (II) conserved in the second lineage, but variable in the first lineage. A site-specific profile based on probability (Q_k) was used to identify critical amino acids [34], with a $Q_k > 0.75$ (see Additional file 9). Among the six lineages, only I, II, V and VI were relevant (see Additional file 10). In fact, when lineages I-II were compared, only 3 amino acids (IDD) were conserved in category II. Lineages I-V showed only one amino acid (R) in category I, and two (DR) in category II. The divergence was clearly more pronounced between lineage I and VI, with four amino acids (LPRE) in category I and 9 (IDAGIIAIG) in category II. This divergence pattern was also observed between lineages II and VI, with four amino acids (LDLD) in category I and 12 (FAIVIDAAGMRL) in category II. Finally, the divergence between lineage V and VI was reduced to only one amino acid (K) in category I.

To visualize these divergence sites, a 3D representation was carried out for each lineage (see Additional file 10) and for all sites together using *Rhodobacter sphaeroides* sequence and its crystal structure (pdb: 1k2w) [35] (Figure 6). At first sight, it was clear that divergent amino acids are basically outside the main conserved blocks, clearly indicating that the drift at these sites (shown by different colors in Figure 6A) is well tolerated by the structure with no loss of activity. The changes are outside block I (NADH-binding domain), block III (which stabilizes the central β-sheet), block VII (which determines the reaction direction) and blocks VIII and IX (involved in the cap domain, which defines the substrate channel, Figure 6B). However, as shown in Figure 6B, the main changes are located in and around β5, which contains one of the amino acids of the



Block I

Lineage	Sequence																
6.1	K	x	A	L	I	T	G	[SA]	A	R	G	I	G	x	x	F	A
6.2	K	S	A	L	I	T	G	[SA]	A	R	G	I	G	[RK]	G	F	A
6.3	K	[ST]	A	L	I	T	G	[SA]	A	R	G	I	G	[RK]	A	F	A
Consensus	K	x	A	L	I	T	G	[SA]	A	R	G	I	G	x	x	F	A
Guidance score	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Block II

Lineage	Sequence				
6.1	D	V	[TS]	[pc]	
6.2	D	V	[TS]	[pc]	Q
6.3	D	V	T	[pc]	Q
Consensus	D	V	T	[pc]	
Guidance score	0,999	0,999	1	1	1

Block III

Lineage	Sequence																		
6.1	[Gc]	x	[IL]	D	[IV]	L	[VI]	N	N	A	A	h	F	[TS]	A	A	P		
6.2	G	K	L	D	I	L	[VI]	N	N	A	A	[LV]	F	D	A	A	E	T	V
6.3	G	x	[IL]	D	I	L	[IV]	N	N	A	A	L	F	D	L	A	P	I	
Consensus	G	x	h	D	I	L	h	N	N	A	A	h	F	[DT]	[AL]	A	[PE]		
Guidance score	1	1	0,974	0,998	0,998	0,998	0,999	0,999	1	1	1	1	1	1	1	1	1	1	1

Block IV

Lineage	Sequence				
6.1	I	N	h	x	G
6.2	h	N	V	A	G
6.3	[IV]	N	V	[SA]	G
Consensus	h	N	h	x	G
Guidance score	1	1	1	1	1

Block V

Lineage	Sequence																	
6.1	G	c	I	I	N	M	A	S	Q	A	G	R	R	G	E	h	L	V
6.2	G	K	I	I	N	M	A	S	Q	A	G	R	R	G	E	h	L	V
6.3	G	c	I	I	N	M	A	S	Q	A	G	R	R	G	E	h	L	V
Consensus	G	[cp]	h	h	N	M	[AS]	S	Q	A	G	R	R	G	E	h	L	V
Guidance score	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Block VI

Lineage	Sequence														
6.1	Y	C	A	[TS]	K	A	A	[VI]	I	S	L	T	Q	S	A
6.2	Y	C	S	T	K	A	A	V	I	S	M	T	Q	S	A
6.3	Y	C	A	T	K	A	A	[VI]	I	S	L	T	Q	S	A
Consensus	Y	C	[AS]	[TS]	K	A	A	[VI]	I	S	[LM]	T	Q	S	A
Guidance score	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Block VII

Lineage	Sequence														
6.1	I	N	V	N	A	I	A	P	G	V	V	D	G	E	H
6.2	I	N	V	N	A	I	A	P	G	V	V	D	G	c	H
6.3	I	N	V	N	A	I	A	P	G	V	V	D	G	c	H
Consensus	I	N	V	N	A	I	A	P	G	V	V	D	G	c	H
Guidance score	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Block VIII

Lineage	Sequence											
6.1	W	D	G	V	D	A	F	F	A	[RK]	[YH]	E
6.2	W	x	H	V	D	[SA]	M	F	A	K	L	E
6.3	W	c	x	V	D	x	x	F	A	c	x	E
Consensus	W	x	x	V	D	x	x	F	A	c	x	E
Guidance score	1	1	1	0,986	0,987	0,987	0,987	0,987	0,987	0,987	0,987	0,987

Block IX

Lineage	Sequence															
6.1	G	[EQ]	K	K	x	E	V	[GA]	x	[SA]	V	P	a	G	R	M
6.2	G	Q	K	c	x	E	V	x	A	S	V	P	A	G	R	
6.3	G	E	K	K	x	x	V	G	x	x	V	P	x	G	R	M
Consensus	G	[EQ]	K	K	x	x	V	x	x	x	V	P	a	G	R	
Guidance score	0,986	0,986	1	1	1	0,987	0,999	0,999	1	1	1	1	1	1	1	1

Block X

Lineage	Sequence							
6.1	G	M	A	[IV]	F	L	A	[ST]
6.2		M	A	[IV]	F	L	A	S
6.3	G	M	A	[IV]	F	L	A	[ST]
Consensus		M	A	[IV]	F	L	A	[ST]
Guidance score	1	1	1	1	1	1	1	1

Block XI

Lineage	Sequence												
6.1	A	Q	[TC]	[YF]	[NG]	V	D	G	G	[NQ]	W	[ML]	[SA]
6.2	x	Q	T	Y	N	V	D	G	G	Q	W	M	S
6.3	x	Q	T	Y	N	V	D	G	G	[NQ]	W	M	[NS]
Consensus	x	Q	[TC]	[YF]	[NG]	V	D	G	G	[NQ]	W	[ML]	[S]
Guidance score	1	1	1	1	1	1	1	1	1	0,983	0,983	0,983	0,983

Figure 5 (See legend on next page.)

(See figure on previous page.)
Figure 5 Conserved SDH blocks in the three sublineages of lineage 6. "c", a charged residue; "h", a hydrophobic residue; "p", a polar residue and "x", any residue. Alternative amino acids at a given position are shown within brackets. Red background indicates strictly conserved amino acids, orange background indicates conserved amino acids and blue boxes indicate the specific blocks of each lineage. Guidance scores represent the degree of confidently aligned residues (1 corresponds to 100% certainty) [25,26].

catalytic tetrad (S139, 1k2w numbering), which is not affected. Thus, it could be concluded that the sequence alignment of the different lineages obtained according to the phylogenetic and functional divergence analysis showed specific fingerprints for each lineage, whose conserved sequences could be very useful for the future classification of SDH enzymes.

Conclusions

SDR196C family encompasses short chain SDH of prokaryotic origin. The distribution of this family is limited to Gram-negative bacteria, grouping members of the α-, β- and γ-Proteobacteria. This distribution is in agreement with the widespread nature of SDR, and indicates that sorbitol metabolism is of importance among these bacteria as an alternative source of carbon and energy. This variability is also observed in the genetic organization of polyol operon among the different species, and in the phylogenetic analysis, which clearly point to different origins for SDH in bacteria, although the 3D structure among groups is highly conserved, with the typical Rossmann fold motif. Such a degree of divergence in the origin but similarity in the structure suggests that SDHs are of extremely old, and emerged early in the evolution, giving rise to the six different lineages and the three sublineages observed in the phylogenetic analysis. This phylogenetic classification is supported by the sequence differences found in the conserved blocks of SDH, allowing for the first time, the classification of the SDR196C family into different subgroups, and introducing the possibility of expanding the actual classification of SDR enzymes by two extra numbers, the lineage and sublineage, separated by a point. As an example, *R. capsulatus* SDH could be classified as SDR196C6.3.

Methods

Sequence retrieval and cluster identification

Protein sequences were obtained from the SDR-enzyme database (http://www.sdr-enzymes.org), a sustainable and expandable nomenclature database based on hidden Markov models [8,15]. The DNA sequences of the 16S rRNA from species encoding the SDH gene were from GenBank. When several strains from the same species encoded the same SDH gene, only one representative strain was included, the strain from the first sequenced genome.

Sequence alignment

The sequences were aligned using GUIDANCE [25,26] with the MSA algorithm PRANK [27]. The alignments were further checked manually using Gene-Doc [36]. Large gaps and hyper variable sites were removed from the alignments; the same methodology was applied to gaps at the beginning and end of the alignment, which represent missing sequence data. Aligned sequences and their secondary structure are shown using ESPript [37].

Phylogenetic analysis

Prot test and model test (protein and DNA sequences, respectively) were used in order to choose the most appropriate method to calculate the distances [38]. WAG with invariable sites for the SDH protein sequences and GTR with invariable sites for 16S rRNA sequences were used [39,40]. Three different tree-building methods were used: Maximum Likelihood (ML), Bayesian analysis (BY) and Neighbor Joining (NJ), as implemented in PHYML, MrBayes 3.1.2, and MEGA 4, respectively [41-43]. The bootstrap values for ML and NJ trees were obtained after 1,000 generations. For the trees constructed using BY, the Markov chains were run for 1,000,000 generations. The burn-in values were set for 10,000 generations, and the trees were sampled every 100 generations. Splitstree and MEGA 4 tree viewer were used to visualize the trees and calculate confidence values [44,45].

Functional divergence analysis

Type I functional divergence was tested according to the previously described methods, using the DIVERGE software [33]. The alignment used for the phylogenetic analysis was also used for this application. The tree obtained by NJ was refined using the tool PROTTEST [46], to determine the best evolutionary model for the set of query proteins [47]. The crystal structure 1k2w was used to determine the location of divergent amino acids according to the analysis obtained. The test could not be applied to lineages III and IV, since DIVERGE needs at least 4 species to be considered a cluster. A site-specific profile based on probability (Q_k) was used to identify critical amino acids [34], with a $Q_k > 0.75$. The values obtained for the critical amino acids and

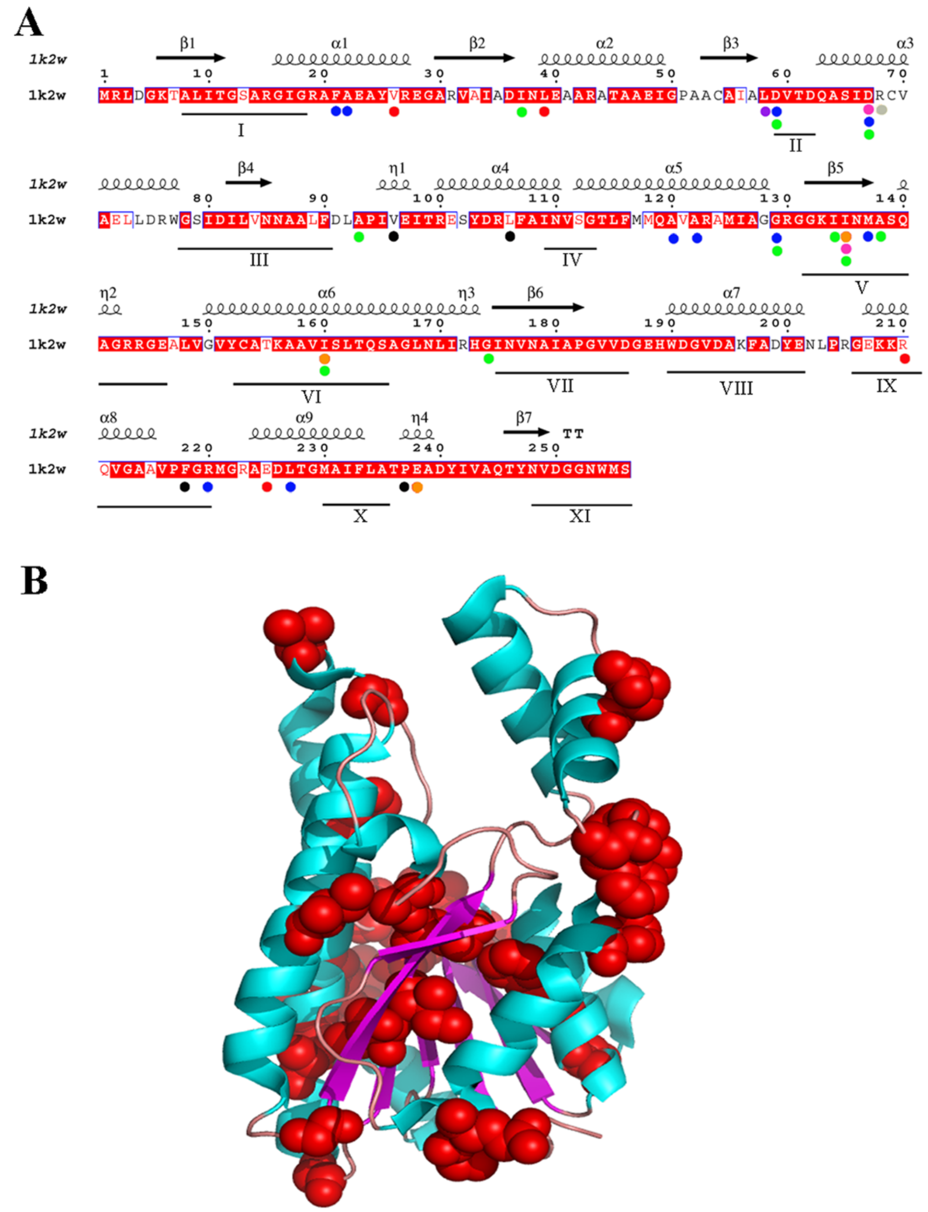

A
1k2w
β1 α1 β2 α2 β3 α3
1 10 20 30 40 50 60 70
1k2w MRLDGKTALITGSARGIGRAFAEAYVREGARVAIADINLEAARATAAEIGPAACAIALDVTDQASIDRCV
I
II
β4 η1 α4 α5 β5
80 90 100 110 120 130 140
1k2w AELLDRWGSIDILVNNAALFDLAPIVEITRESYDRLFAINVSGTLFMMQAVARAMIAGGRGGKIINMASQ
III
IV
V
η2 α6 η3 β6 α7
150 160 170 180 190 200 210
1k2w AGRRGEALVGVYCATKAAVISLTQSAGLNLIRHGINVNAIAPGVVDGEHWDGVDAKFADYENLPRGEKKR
VI
VII
VIII
IX
α8 α9 η4 β7
220 230 240 250
1k2w QVGAAVPFGRMGRAEDLTGMAIFLATPEADYIVAQTYNVDGGNWMS
X
XI
B

Figure 6 The site-specific profile fro predicting critical amino acid residues for the functional divergence between SDH families. A) Critical amino acid residues are represented by circles on the sequence of R. *sphaeroides* SDH (pdb: 1k2w). Symbols above the sequence represent the secondary structure, springs represent helices and arrows represent β-strands. Conserved blocks of bacterial SDH (I to XI) are indicated below the sequence alignment. Triangles represent the location of the active site. Circles represent the divergent amino acids according to the following color code: Grey; conserved amino acids in Lineage I and divergent in Lineage V, Red; conserved amino acids in Lineage I and divergent in Lineage VI, Orange; conserved amino acids in Lineage II and divergent in Lineage I, Black; conserved amino acids in Lineage II and divergent in Lineage VI, Pink; conserved amino acids in Lineage V and divergent in Lineage I, Purple; conserved amino acids in Lineage V and divergent in Lineage VI, Green; conserved amino acids in Lineage VI and divergent in Lineage I, Blue; conserved amino acids in Lineage VI and divergent in Lineage II. **B**) Representation of divergent amino acids in the structure of 1k2w. α-helices are indicated in cyan, β-sheets are indicated in purple, loops are indicated in light pink and the divergent amino acids as red spheres. Structure was rendered using PyMol [50].

their location in the alignment according to DIVERGE are shown in Additional files 9, 10, 11, and 12.

GC content

The GC content of the sequences was calculated and compared to the GC content of the whole genome. The formula used for the calculations was that described by Karlin *et al*., 2001 [48].

Molecular modeling

Protein sequences were 3D modeled with Geno3D [49] and molecular representation were performed by Pymol [50].

Additional files

Additional file 1: Distribution of SDH gene among bacteria. The table indicates the bacterial species that encode a SDH gene, its taxonomy, ecology and niche.

Additional file 2: GC content differences between SDH genes and genome. The table indicates the differences in GC content between the SDH gene and the core genome of the bacterium, where it is encoded.

Additional file 3: Structure of the SDH clusters among bacterial groups. SDH gene was placed in the middle to facilitate cluster visualization. Variants of the cluster among taxonomic groups are represented by numbers (1–21).

Additional file 4: Multiple sequence alignments (MSAs) obtained with GUIDANCE and their corresponding Guidance scores [25,26]. The file shows the result obtained for the MSA of the 6 lineages, 3 sublineages and 16S rRNA. In all cases, an overall quality assessment (Guidance score) of above 0.97 was obtained with the MSA algorithm PRANK [27].

Additional file 5: Phylogenetic trees of bacteria containing SDH gene. The file shows two phylogenetic trees of SDH gene using Bayesian and Maximun Likelihood, as tree building methods.

Additional file 6: Phylogenetic trees (16S rRNA) of bacteria containing the SDH gene. The file contained two phylogenetic trees of 16S rRNA of bacteria containing SDH gene, using Bayesian and Maximum Likelihood, as tree building methods.

Additional file 7: Molecular modeling of SDH enzymes from the six lineages described in bacterial SDH. The proteins were modeled with Geno3D [49] and rendered by PyMol [50]. Selected proteins were Uniprot codes: lineage 1, A3MHB9; lineage 2, A1WMN; lineage 3 E1VCL7; lineage 4, A5FVQ; lineage 5, A9CES4; lineage 6.1, O68112; lineage 6.2, A3K129; lineage 6.3, A3PKH5.

Additional file 8: Multiple sequence alignment of bacterial SDH proteins. ESPript [37] output was obtained with the aligned sequences from Additional file 1. Strictly conserved residues have a solid background. Symbols above sequences represent the secondary structure. Conserved structural blocks (I-XI) are shown below the sequences. Circles represent the divergent amino acids according to the following color code: Grey; conserved amino acids in Lineage I and divergent in Lineage V, Red; conserved amino acids in Lineage I and divergent in Lineage VI, Orange; conserved amino acids in Lineage II and divergent in Lineage I, Black; conserved amino acids in Lineage II and divergent in Lineage VI, Pink; conserved amino acids in Lineage V and divergent in Lineage I, Purple; conserved amino acids in Lineage V and divergent in Lineage VI, Green; conserved amino acids in Lineage VI and divergent in Lineage I, Blue; conserved amino acids in Lineage VI and divergent in Lineage II.

Additional file 9: Predicted critical amino acid sites responsible for functional divergence. The table represents the values of Q_k of all pair analyses. Values with a $Q_k > 0.75$ are indicated with a red background.

Additional file 10: Candidates for amino acid sites related with functional divergence. The numbering used corresponds with that of the alignment implemented by DIVERGE. Cat. I refers to Category I and Cat. II refers to Category II. A: Cat. II; conserved tandem in Lineage II and variable in Linage I. B: Cat. I; Conserved tandem in Lineage I and variable in Lineage V and Cat. II; conserved tandem in Lineage V and variable in Lineage I. C: Cat. I; conserved tandem in Lineage V and variable in Lineage VI. D: Cat. I; conserved tandem in Lineage I and variable in Lineage VI and Cat. II; conserved tandem in Lineage VI and variable in Lineage I. E: Cat. I; conserved tandem in Lineage II and variable in Lineage VI and Cat. II; conserved tandem in Lineage VI and variable in Lineage II. Critical amino acids responsible for functional divergence are shown in the *R. sphaeroides* SDH crystal structure (pdb:1k2w). Divergent residues of category I are depicted in red and those of category II are depicted in blue.

Additional file 11: Alignment of SDH family in ClustalW format (.aln) used in the DIVERGE analysis.

Additional file 12: Phylogenetic tree in PhylM (.ph) used in the DIVERGE analysis.

Abbreviations

SDR: Short-chain dehydrogenase/reductase(s); SDH: Sorbitol dehydrogenase(s); ABC: ATP-binding cassette; MDR: Medium-chain dehydrogenase/reductase(s); AKR: Aldo-keto reductase(s); HAD: Haloacid dehalogenase(s); MDH: Mannitol dehydrogenase(s); ML: Maximum likelihood; BY: Bayesian analysis; NJ: Neighbor joining.

Competing interests

The authors declare that they have no competing interest.

Authors' contributions

ASF and FGC designed research. MIGG did sequence retrieval and alignments with ESPript. ASC carried out phylogenetic analysis by using related computer programs, and together with ASF, drafting of the manuscript. All authors read and approved the final manuscript.

Acknowledgements

This study was partially supported by MINECO-FEDER (BIO2010-22225-C02-01) and Programa de Ayuda a Grupos de Excelencia de la Región de Murcia de la Fundación Séneca (04541/GERM/06, Plan Regional de Ciencia y Tecnología 2007–2010). A.S.C and M.I.G.G are holders of predoctoral research grants from Grupos de Excelencia de la Región de Murcia, Spain.

Received: 2 December 2011 Accepted: 8 August 2012
Published: 16 August 2012

References

1. Kavanagh KL, Jornvall H, Persson B, Oppermann U: **Medium- and short-chain dehydrogenase/reductase gene and protein families: the SDR superfamily: functional and structural diversity within a family of metabolic and regulatory enzymes.** *Cell Mol Life Sci* 2008, **65**(24):3895–3906.
2. Schwartz MF, Jornvall H: **Structural analyses of mutant and wild-type alcohol dehydrogenases from drosophila melanogaster.** *Eur J Biochem* 1976, **68**(1):159–168.
3. Thatcher DR: **The complete amino acid sequence of three alcohol dehydrogenase alleloenzymes (AdhN-11, AdhS and AdhUF) from the fruitfly *Drosophila melanogaster*.** *Biochem J* 1980, **187**(3):875–883.
4. Jornvall H, Persson B, Krook M, Atrian S, Gonzalez-Duarte R, Jeffery J, Ghosh D: **Short-chain dehydrogenases/reductases (SDR).** *Biochemistry* 1995, **34**(18):6003–6013.
5. Krook M, Marekov L, Jornvall H: **Purification and structural characterization of placental NAD(+)-linked 15-hydroxyprostaglandin dehydrogenase. The primary structure reveals the enzyme to belong to the short-chain alcohol dehydrogenase family.** *Biochemistry* 1990, **29**(3):738–743.
6. Oppermann UC, Maser E, Hermans JJ, Koolman J, Netter KJ: **Homologies between enzymes involved in steroid and xenobiotic carbonyl reduction in vertebrates, invertebrates and procaryonts.** *J Steroid Biochem Mol Biol* 1992, **43**(7):665–675.
7. Persson B, Krook M, Jornvall H: **Characteristics of short-chain alcohol dehydrogenases and related enzymes.** *Eur J Biochem* 1991, **200**(2):537–543.
8. Kallberg Y, Oppermann U, Persson B: **Classification of the short-chain dehydrogenase/reductase superfamily using hidden Markov models.** *FEBS J* 2010, **277**(10):2375–2386.
9. Kallberg Y, Persson B: **Prediction of coenzyme specificity in dehydrogenases/reductases. A hidden Markov model-based method and its application on complete genomes.** *FEBS J* 2006, **273**(6):1177–1184.
10. Pearl F, Todd A, Sillitoe I, Dibley M, Redfern O, Lewis T, Bennett C, Marsden R, Grant A, Lee D, *et al*: **he CATH Domain Structure Database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis.** *Nucleic Acids Res* 2005, **33**(Database issue):D247–D251.
11. Keller B, Volkmann A, Wilckens T, Moeller G, Adamski J: **Bioinformatic identification and characterization of new members of short-chain dehydrogenase/reductase superfamily.** *Mol Cell Endocrinol* 2006, **248**(1–2):56–60.
12. Kallberg Y, Oppermann U, Jornvall H, Persson B: **Short-chain dehydrogenases/reductases (SDRs).** *Eur J Biochem* 2002, **269**(18):4409–4417.
13. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, *et al*: **Environmental genome shotgun sequencing of the Sargasso Sea.** *Science* 2004, **304**(5667):66–74.

14. Filee J, Forterre P, Laurent J: **The role played by viruses in the evolution of their hosts: a view based on informational protein phylogenies.** *Res Microbiol* 2003, **154**(4):237–243.
15. Persson B, Kallberg Y, Bray JE, Bruford E, Dellaporta SL, Favia AD, Duarte RG, Jornvall H, Kavanagh KL, Kedishvili N, *et al*: **The SDR (short-chain dehydrogenase/reductase and related enzymes) nomenclature initiative.** *Chem Biol Interact* 2009, **178**(1–3):94–98.
16. Schneider K-H, Giffhorn F: **Sorbitol dehydrogenase from Pseudomonas sp.: purification, characterization and application to quantitative determination of sorbitol.** *Enzyme Microb Technol* 1991, **13**(4):332–337.
17. Jörnvall H, Hedlund J, Bergman T, Oppermann U, Persson B: **Superfamilies SDR and MDR: from early ancestry to present forms. Emergence of three lines, a Zn-metalloenzyme, and distinct variabilities.** *Biochem Biophys Res Commun* 2010, **396**(1):125–130.
18. Apisarnthanarak A, Kiratisin P, Mundy LM: **Evaluation of *Ochrobactrum intermedium* bacteremia in a patient with bladder cancer.** *Diagn Microbiol Infect Dis* 2005, **53**(2):153–155.
19. Arora U, Kaur S, Devi P: ***Ochrobactrum anthropi* septicaemia.** *Indian J Med Microbiol* 2008, **26**(1):81–83.
20. Bilgin AA, Silverstein J, Hernandez M: **Effects of soluble ferri-hydroxide complexes on microbial neutralization of acid mine drainage.** *Environ Sci Technol* 2005, **39**(20):7826–7832.
21. Schauder S, Schneider KH, Giffhorn F: **Polyol metabolism of *Rhodobacter sphaeroides*: biochemical characterization of a short-chain sorbitol dehydrogenase.** *Microbiology* 1995, **141**(Pt 8):1857–1863.
22. de Vrankrijker AM, Wolfs TF, van der Ent CK: **Challenging and emerging pathogens in cystic fibrosis.** *Paediatr Respir Rev* 2010, **11**(4):246–254.
23. Rodriguez-Moreno L, Jimenez AJ, Ramos C: **Endopathogenic lifestyle of Pseudomonas savastanoi pv. savastanoi in olive knots.** *Microb Biotechnol* 2009, **2**(4):476–488.
24. Stein MA, Schafer A, Giffhorn F: **Cloning, nucleotide sequence, and overexpression of smoS, a component of a novel operon encoding an ABC transporter and polyol dehydrogenases of *Rhodobacter sphaeroides* Si4.** *J Bacteriol* 1997, **179**(20):6335–6340.
25. Penn O, Privman E, Ashkenazy H, Landan G, Graur D, Pupko T: **GUIDANCE: a web server for assessing alignment confidence scores.** *Nucleic Acids Res* 2010, **38**(suppl 2):W23–W28.
26. Penn O, Privman E, Landan G, Graur D, Pupko T: **An alignment confidence score capturing robustness to guide tree uncertainty.** *Mol Biol Evol* 2010, **27**(8):1759–1767.
27. Jordan G, Goldman N: **The effects of alignment error and alignment filtering on the sitewise detection of positive selection.** *Mol Biol Evol* 2012, **29**(4):1125–1139.
28. Woese C: **The universal ancestor.** *Proc Natl Acad Sci* 1998, **95**(12):6854–6859.
29. Filling C, Berndt KD, Benach J, Knapp S, Prozorovski T, Nordling E, Ladenstein R, Jornvall H, Oppermann U: **Critical residues for structure and catalysis in short-chain dehydrogenases/reductases.** *J Biol Chem* 2002, **277**(28):25677–25684.
30. Jornvall H, Hoog JO, Persson B: **SDR and MDR: completed genome sequences show these protein families to be large, of old origin, and of complex nature.** *FEBS Lett* 1999, **445**(2–3):261–264.
31. Jornvall H, Persson M, Jeffery J: **Alcohol and polyol dehydrogenases are both divided into two protein types, and structural properties cross-relate the different enzyme activities within each type.** *Proc Natl Acad Sci USA* 1981, **78**(7):4226–4230.
32. Gu X: **Functional divergence in protein (family) sequence evolution.** *Genetica* 2003, **118**(2):133–141.
33. Gu X, Vander Velden K: **DIVERGE: phylogeny-based analysis for functional–structural divergence of a protein family.** *Bioinformatics* 2002, **18**(3):500–501.
34. Gu J, Wang Y, Gu X: **Evolutionary analysis for functional divergence of jak protein kinase domains and tissue-specific genes.** *J Mol Evol* 2002, **54**(6):725–733.
35. Philippsen A, Schirmer T, Stein MA, Giffhorn F, Stetefeld J: **Structure of zinc-independent sorbitol dehydrogenase from Rhodobacter sphaeroides at 2.4 A resolution.** *Acta Crystallogr D Biol Crystallogr* 2005, **61**(Pt 4):374–379.
36. Nicholas KB, Nicholas HB, Deerfield DW: **{GeneDoc: analysis and visualization of genetic variation}.** *EMBNEW NEWS* 1997, **4**:14.
37. Gouet P, Courcelle E, Stuart DI, Metoz F: **ESPript: analysis of multiple sequence alignments in PostScript.** *Bioinformatics* 1999, **15**(4):305–308.
38. Posada D, Crandall KA: **MODELTEST: testing the model of DNA substitution.** *Bioinformatics* 1998, **14**(9):817–818.
39. Goldman N, Whelan S: **Statistical tests of gamma-distributed rate heterogeneity in models of sequence evolution in phylogenetics.** *Mol Biol Evol* 2000, **17**(6):975–978.
40. Lanave C, Preparata G, Saccone C, Serio G: **A new method for calculating evolutionary substitution rates.** *J Mol Evol* 1984, **20**(1):86–93.
41. Felsenstein J: **Mathematics vs. Evolution: Mathematical Evolutionary Theory.** *Science* 1989, **246**(4932):941–942.
42. Guindon S, Gascuel O: **A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood.** *Syst Biol* 2003, **52**(5):696–704.
43. Huelsenbeck JP, Ronquist F: **MRBAYES: Bayesian inference of phylogenetic trees.** *Bioinformatics* 2001, **17**(8):754–755.
44. Huson DH, Bryant D: **Application of phylogenetic networks in evolutionary studies.** *Mol Biol Evol* 2006, **23**(2):254–267.
45. Tamura K, Dudley J, Nei M, Kumar S: **MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0.** *Mol Biol Evol* 2007, **24**(8):1596–1599.
46. Abascal F, Zardoya R, Posada D: **ProtTest: selection of best-fit models of protein evolution.** *Bioinformatics*, **21**(9):2104–2105.
47. Benítez-Páez A: *Cárdenas-Brito S*. Gutiérrez AJ: A practical guide for the computational selection of residues to be experimentally characterized in protein families. Briefings in Bioinformatics; 2011.
48. Karlin S: **Detecting anomalous gene clusters and pathogenicity islands in diverse bacterial genomes.** *Trends Microbiol* 2001, **9**(7):335–343.
49. Combet C, Jambon M, Deleage G, Geourjon C: **Geno3D: automatic comparative molecular modelling of protein.** *Bioinformatics* 2002, **18**(1):213–214.
50. Schrödinger: *The PyMOL Molecular Graphics System, Version 1.3. In.: LLC.*

doi:10.1186/1471-2148-12-147
Cite this article as: Sola-Carvajal *et al.*: **Insights into the evolution of sorbitol metabolism: phylogenetic analysis of SDR196C family.** *BMC Evolutionary Biology* 2012 **12**:147.