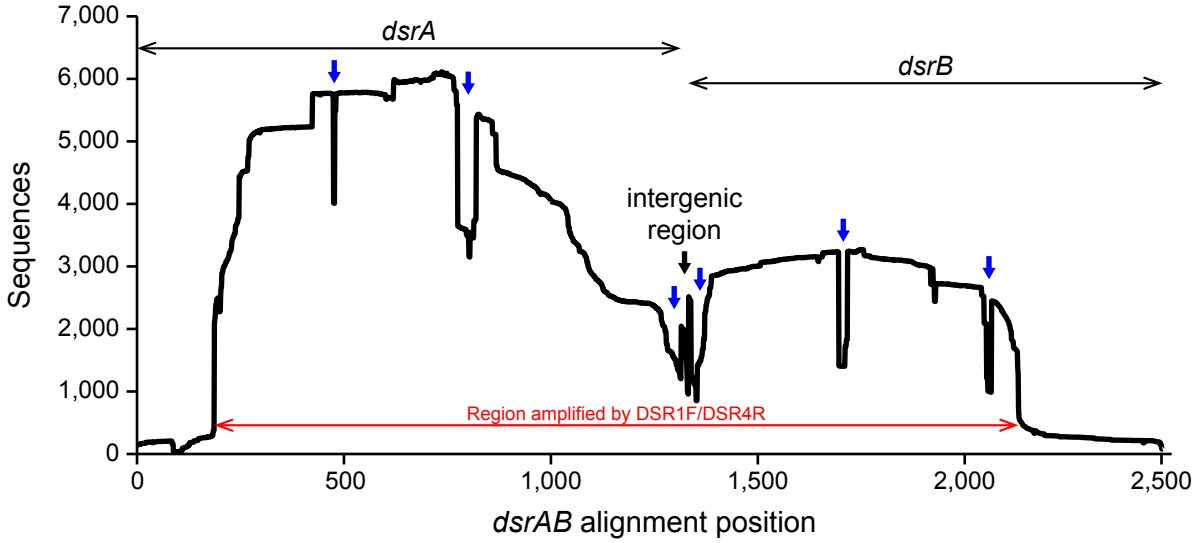
dsrAB/DsrAB ARB database description

The *dsrAB*/DsrAB ARB database contains 7,695 publicly available partial and full-length *dsrAB* nucleotide sequences along with inferred corresponding amino acid sequences obtained by 530 amplicon sequencing, metagenome or genome studies (status August 2013). All sequences are at least 300 nucleotides long and contain less than 1% ambiguities. Sequences are subdivided into a core dataset of 1,292 sequences that completely cover the region amplified by the most commonly used sequencing primers DSR1F/DSR4R and that can be used for reliable phylogenetic inferences and into 6,403 shorter, partial sequences.

Following Figure shows an overview of the sequence distribution in the database. Alignment positions correspond to the *dsrAB* sequence of *Desulfovibrio vulgaris* (NC_002937, 449880..452365). Blue arrows indicate common insertion/deletion regions.



Operational classification system

All sequences are classified (database field <classification>) according to DsrAB type (reductive bacterial or archaeal and oxidative bacterial type), supercluster (for reductive bacterial type DsrAB), phylum and class (based on the taxonomy of known representatives), as well as DsrAB lineage.

Ecological categories

All sequences are assigned to broad environmental categories (database field <environment>) based on the qualitative description submitted with the sequence and/or the information in the corresponding publication. Categories based on the environmental origin of the sample (marine, estuarine, freshwater, soil, and industrial) are complemented by categories denoting special microbial lifestyles (thermophilic, alkali-/halophilic, and symbiotic). Sequences are not assigned to multiple categories, lifestyle categories are given precedence in case of sequences fitting in two or more categories (e.g. a sequence from a marine thermophile is classified as thermophilic and not as marine).

Additional associated information

Sequences in the database are identified by an eight digit identifier (database field <name>) and are accompanied by information such as a short description (database field <full name>), accession number (database field <acc>), clone or strain (database fields <clone> and <strain>), corresponding publication (database fields <author>, <title>, and <journal>), and sequence length (database fields <nuc> and <aa>).

Sequence filters

Filters omitting insertions/deletions from the alignment (indel filters) are available in the database, one for each type of *dsrAB*:

- ***indel_SRP_aa*** (and corresponding nucleotide filter ***indel_SRP_nuc***) covering 530 amino acid positions (318 in DsrA, 212 in DsrB) in the region amplified by the most commonly used primers DSR1F/DSR4R used for phylogenetic inference of reductive bacterial type DsrAB sequences.
- ***indel_SOP_aa*** (and corresponding nucleotide filter ***indel_SOP_nuc***) covering 552 amino acid positions (323 in DsrA, 229 in DsrB) in the region amplified by the most commonly used primers rDSR1F/rDSR4R used for phylogenetic inference of oxidative bacterial type DsrAB sequences.
- ***indel_ARC_aa*** (and corresponding nucleotide filter ***indel_ARC_nuc***) covering 629 amino acid positions (362 in DsrA, 267 in DsrB) used for phylogenetic inference of reductive archaeal type DsrAB sequences.

The filter section of the amino acid alignment contains also a filter named *Features* that indicates the regions of the DsrA and DsrB subunits, the intergenic spacer, as well as the conserved siroheme binding motifs.

Phylogenetic trees

The database also contains a collection of DsrAB reference trees. All trees are consensus trees reconstructed from trees calculated using maximum parsimony, maximum likelihood and neighbor joining algorithms:

- ***tree_DsrAB_consensus_phylogeny***, a consensus tree of 1292 DsrAB sequences of the core dataset that completely cover the region amplified by the most commonly used sequencing primers DSR1F/4R.
- ***tree_oxidative_bacterial_type_DsrAB***, a consensus tree of oxidative bacterial type DsrAB sequences of the core dataset.
- ***tree_reductive_archaeal_type_DsrAB***, a consensus tree of reductive archaeal type DsrAB sequences of the core dataset.
- ***tree_reductive_DsrAB_cultures_genomes***, a consensus tree of reductive type DsrAB sequences from pure cultures and genomes.
- ***tree_oxidative_DsrAB_cultures_genomes***, a consensus tree ofoxidative type DsrAB sequences from pure cultures and genomes.
- ***tree_DsrAB_overview_all_sequences***, the consensus phylogeny tree additionally containing all partial sequences that were added to the consensus tree using RAxML-HPC.

dsrAB/DsrAB-database as FASTA files

Aligned nucleotide and inferred amino acid sequences are also available as FASTA files. FASTA headers contain information pertaining to operational classification and environmental categories using following format (tab-delimited):

>8digitID description accession number environmental category operational classification