Functional annotation of orthologs in metagenomes: a case study of genes for the transformation of oceanic dimethylsulfoniopropionate

José M. González¹ · Laura Hernández¹ · Iris Manzano¹ · Carlos Pedrós-Alió²

Received: 25 July 2018 / Revised: 22 November 2018 / Accepted: 25 December 2018 / Published online: 14 January 2019
© International Society for Microbial Ecology 2019

Abstract

Dimethylsulfoniopropionate (DMSP) is produced mainly by phytoplankton and bacteria. It is relatively abundant and ubiquitous in the marine environment, where bacterioplankton make use of it readily as both carbon and sulfur sources. In one transformation pathway, part of the molecule becomes dimethylsulfide (DMS), which escapes into the atmosphere and plays an important role in the sulfur exchange between oceans and atmosphere. Through its other dominant catabolic pathway, bacteria are able to use it as sulfur source. During the past few years, a number of genes involved in its transformation have been characterized. Identifying genes in taxonomic groups not amenable to conventional methods of cultivation is challenging. Indeed, functional annotation of genes in environmental studies is not straightforward, considering that particular taxa are not well represented in the available sequence databases. Furthermore, many genes belong to families of paralogs with similar sequences but perhaps different functions. In this study, we develop *in silico* approaches to infer protein function of an environmentally important gene (*dmdA*) that carries out the first step in the sulfur assimilation from DMSP. The method combines a set of tools to annotate a targeted gene in genome databases and metagenome assemblies. The method will be useful to identify genes that carry out key biochemical processes in the environment.

Introduction

Dimethylsulfoniopropionate (DMSP) and its products after transformation are key molecules in the sulfur cycle on the planet. DMSP is produced by phytoplankton and bacteria in the oceans at a rate estimated to be of 10^9 tons per year [1, 2]. The degradative metabolism of DMSP has been studied especially because it is the major biogenic precursor of marine dimethylsulfide (DMS). This volatile sulfur compound is a main player in the global atmospheric sulfur cycle and possibly has a role in climate regulation [3–5]. Around 3×10^7 tons of S in the form of DMS enter the

atmosphere from the oceans every year [2]. DMS is oxidized in the atmosphere by photochemical reactions into compounds that act as condensation nuclei for water droplets. Released by the metabolism of the microbial food web of the oceans, this simple compound alone affects global climate promoting cloud formation and, consequently, backscatter of solar radiation [4, 6, 7].

DMSP is primarily metabolized by marine bacteria following two competing pathways (Supplementary Fig. 1). Through the demethylation pathway, DMSP is converted to methylmercaptopropionate, which is then demethylated to release methanethiol at the end of the pathway [8–12]. This last compound provides the bacterioplankton with a reduced form of sulfur, as opposed to sulfate in seawater, whose direct assimilation is energetically costly in an environment scarce in nutrients [13]. In the cleavage pathway, most known DMSP lyases generate DMS and acrylate, or alternatively, 3-hydroxypropionate [14, 15].

The first gene product that was identified in the latter pathway was a DMSP-dependent demethylase A (*dmdA*) involved in the initial step to transform DMSP. The gene has been cloned and described in model bacteria in the SAR11 group and *Ruegeria pomeroyi* in the *Rhodobacteraceae* [16], both of which account for a significant fraction

Supplementary information The online version of this article (<https://doi.org/10.1038/s41396-019-0347-6>) contains supplementary material, which is available to authorized users.

✉ José M. González
jmglez@ull.es

¹ Department of Microbiology, University of La Laguna, La Laguna, Spain

² Systems Biology Program, Centro Nacional de Biotecnología, Consejo Superior de Investigaciones Científicas, Madrid, Spain

of bacterioplankton in the oceans. The crystal structure of DmdA from *Pelagibacter ubique* was determined later on [17]. DmdA is an enzyme that belongs to the glycine-cleavage system T-protein family of proteins (the GcvT family), which is diverse and also includes dimethylglycine oxidase and sarcosine oxidase among those that have been characterized. DmdA requires the cofactor tetrahydrofolate (THF) involved in a methyl transfer reaction to generate methylmercaptopropionate [18]. Other gene products demethylate methylmercaptopropionate further down the pathway to methanethiol, which is readily taken up by marine bacteria and incorporated into proteins (ref. [11]; Supplementary Fig. 1). Thus, DMSP-related genes are present in the marine environment in bacterioplankton taxa that dominate the microbial food web. Their expression is involved in a key step in the sulfur cycle in the oceans. In particular, *dmdA* is present in about one third of the bacterioplankton cells [16]. Considering its widespread presence, it would be highly unlikely to find an environmentally significant alternative enzyme that would carry out the same reaction. The main purpose of the present paper was to analyze the diversity of this gene in the marine environment, taking advantage of the large database of metagenomic sequences from Tara Oceans (www.embl.de/tara-oceans). This purpose, however, is fraught with difficulties in the proper identification of the gene. Thus, we developed a pipeline to safely annotate *dmdA* sequences. This pipeline will be useful for retrieval of any other gene from environmental samples that belongs to a gene family, which is the most common case.

Methods to detect protein coding genes, or CDSs, have improved considerably during the past few years, for prokaryotes in particular [19–21]. Even at this point annotation errors can be found in sequences, as different methods can produce different results. For example, the starting site often calls for manual curation to correct its position. Functional annotation errors are even more frequent than those associated with gene finding methods. Annotation errors are aggravated once the new annotation reaches the public databases. There, the error is propagated to new sequences, since the new annotations are based most commonly on amino-acid similarities in a pairwise alignment with previously annotated sequences. Even in the relatively few cases when there is experimental evidence for the first gene in the public databases, the functional correspondence of new sequences showing similarity is not guaranteed. Over-annotation of function (or over-predictions) is frequent [22]. Peptides that are modular in a number of conserved domains are especially troublesome as the new annotations are most likely based on a single conserved domain, instead of the entire peptide, due to sequence similarity with only part of the protein that spans a given conserved domain [22]. The more precise the functional annotation tries to be,

the larger the number of errors it will produce [23]. Considering all the above, the percentage of wrongly annotated genes in a genome tends to be high.

One way to solve mis-annotations is by relying on a curated database of sequences as a standard for annotation. Such is the case of the protein database UniProtKB/Swiss-Prot [24]. However, these databases are incomplete for particular genes or taxonomic groups that are not well represented in the databases. The same holds true for manually annotated genomes, which are still considerably less than the number of automatically annotated genomes. Consequently, researchers must carry out careful corrections of functional annotation. This is a time-consuming process still prone to subjective decisions. At any rate, new annotations always involve sequence similarities. Annotation platforms aid in the decision-making process, in facilitating manual curation, and in reducing subjectivity (see for example refs. [25, 26]). It is then up to the researcher to make decisions to correct the automatic annotation of the gene for which there is no experimental evidence, which is the case in the great majority of sequences in the public databases.

A special case is the functional annotation of genes whose products belong to protein families [27]. Over-annotations are more common in such cases than in gene products that do not belong to protein families. Groups of peptides that show sequence similarities with each other are considered to belong to the same protein family, although their members may have different functions. This is the case of the *dmdA* gene, for example. Protein families arise from a gene duplication event followed by divergence before they gain a new function in the same organism [28] or from horizontal gene transfer. Both processes are strong means of gene innovation. Due to their evolutionary history, there are no clear-cut protocols to determine the peptides with the same or different annotations within a larger family, since protein families can show great diversity. As a result, genes that belong to families are more likely to contain annotation errors due to over-annotation than single copy genes.

For environmental studies, the limitations of gene annotation are even greater. The repositories such as the National Center for Biotechnology Information (NCBI) Reference Sequence Database (RefSeq; ref. [29]) are crowded with genes and genomes from microorganisms isolated on solid media. There is a preponderance of faster-growing organisms and those that depend less on interactions with other organisms or are better adapted to a wide range of conditions. Although growth on solid media allows for the characterization of a given gene in model organisms, experimental evidence is available for relatively few genes. And, moreover, isolates only represent a minor share of the diversity in the environment. Therefore, especially for environmental studies, there is a demand for genomes from

microorganisms that require a bypass of cultivation techniques prior to genome sequencing.

A clear boundary between orthologs and paralogs is critical for the construction of a robust evolutionary classification of genes and reliable functional annotation of new sequences. In this study, we focus on an individual gene of environmental importance that belongs to a family of paralogs as a case study. We start from organisms for which this gene has been characterized experimentally and gather their genome information. The total peptide sequence space derived from metagenomes serves to gain insight into the diversity of the same enzyme in the natural environment. With the combination of amino-acid sequences from enzymes for which there is experimental evidence and peptides from assembled sequences from environmental studies, we construct a database of reliable orthologs. We have designed a workflow that makes use of a combination of methods to discriminate peptides with a given function from a pool of related genes within paralog families. The classification method captures the sequence diversity of DmdAs in the oceanic environment, thus retrieving novel *dmdA* genes in organisms previously not known to have them. It also helps predict other genes involved in the same pathway and, with modification, it would be a guide to predict other genes of environmental relevance.

Materials and methods

Databases of genomes and metagenomes

DmdA peptides were searched for in two sets of sequence databases from marine bacteria, the MAR database [30] and peptides predicted from assembled *Tara* Oceans sequences. The MAR database had 4326 genomes, including genomes from isolates, assembled genomes from metagenomes (MAGs), and single-cell genomes (SAGs). The taxonomic spectrum of the MAR database covered the major groups that dominate the bacterioplankton, *Alphaproteobacteria*, *Gammaproteobacteria*, *Bacteroidetes*, *Archaea*, *Cyanobacteria*, etc., although there is a relatively small representation of the SAR11 group (Supplementary Fig. 2). We translated the genomes with the Prokka package [31]. The *Tara* Oceans database included the metagenome assemblies from raw *Tara* Oceans contigs longer than 1000 nucleotides [32]. The assemblies were translated with Prodigal software in “anonymous” mode [19]. Truncated peptides, i.e., sequences predicted not to cover the beginning or end of the gene according to the Prodigal results, were discarded. We modified the labels of the fasta files to include information on the Open Reading Frames (ORF) coordinates and the original name for the contig. This information was later

used in analyzing the gene arrangements around the *dmdA* gene.

DmdAs were predicted in the peptides derived from the MAR database and metagenome sequences, based on a search using a hidden Markov model (HMM) as described below. In the case of the *Tara* Oceans peptide sequences, a lower sequence quality could be expected. Thus, each of the DmdA candidates were further processed before other analyses to remove spurious sequences as follows. From the sequences that had two or more copies, only one was kept with no further processing. For the rest of sequences, since the elimination of the singletons might have discarded true diversity, they were not excluded altogether but they were further filtered before also including them in the analysis. From these, those with an anomalous amino-acid composition were ignored according to a MUSCLE alignment and a composition chi-square test (p value smaller than 5%), as implemented in IQ-TREE [33]. Peptides with more conserved domains than expected were also removed, as some of them lacked stop codons due to sequencing mistakes or mis-assembly. This method allowed working with only reliable sequences to build a consistent tree or sequence similarity network (SSN) analysis as described below.

DmdA homologs

dmdA was first described in *R. pomeroyi* and SAR11, and later predicted in other taxa, such as *Gammaproteobacteria* and SAR116 [34–36]. Considering that *dmdA* homologs might be found in new taxa, we made use of a number of methods to predict DmdA sequences.

A profile method was selected instead of a pairwise alignment tool. Profile methods incorporate position-specific information since they quantify variations at each position of an alignment [37, 38]. Therefore, this approach seemed better suited to find orthologs in new taxa. An HMM profile was built based on DmdA homologs for which there is experimental evidence in *R. pomeroyi* DSS-3 (accession number AAV95190) and *P. ubique* HTCC1062 (AAZ21068). DmdA orthologs in *Rhodobacteraceae* or SAR11 derived from reciprocal best hits (RBH) were also employed in the profile. For each group of genomes separately, DmdA homologs were identified when there was a RBH for the peptide with accession number AAV95190, in the case of the *Rhodobacteraceae* genomes, or AAZ21068 in the case of SAR11. Especially since the DmdA belongs to a greater family, to avoid false hits for non-DmdAs in genomes that lack a true DmdA, a phylogenetic tree was constructed to identify these false positives as long branches in a maximum likelihood tree. These false positives were eliminated before building an alignment profile with HMMER3 [39].

To set an *e*-value cutoff to identify new DmdAs in other sets of sequences, non-DmdA homologs were first identified in a maximum likelihood tree using GcvTs from the same organisms that contained DmdAs. These peptides were predicted based on the HMM hit to both aminomethyl-transferase folate-binding domain (PF01571) and glycine-cleavage T-protein C-terminal barrel domain (PF08669). Peptides with additional conserved domains were ignored. In this way, the limit to sort out DmdAs from non-DmdAs was set to a maximum *e*-value of e^{-130} . This *e*-value was further confirmed when sequences from metagenomes were analyzed, which included a greater diversity than the genomes downloaded from the databases. For this purpose, a new tree was constructed with metagenome sequences, as well as genomes from the databases that had been obtained using the HMM designed for DmdA orthologs, using a greater maximum *e*-value (e^{-50}). Clusters of GcvT peptides (other than DmdA) closest in the phylogenetic tree to DmdA orthologs were identified. Once the *e*-value was set, RefSeq sequences (Release 87, March 2018) that passed the e^{-130} *e*-value threshold were also added to the trees making 69 new DmdA sequences.

Treeing methods

Peptide sequences were aligned with MUSCLE [40] and poorly aligned positions in the alignment were removed with TrimAl [41]. The parameters were set to a minimum overlap of 0.55 and a minimum percent of “good positions” to 60. A maximum likelihood phylogenetic tree and corresponding bootstrap support values (100 replicates) were calculated using RAxML v7.2.6 [42]. The maximum likelihood tree was selected from 20 heuristic tree searches initiated from randomized parsimony starting trees. The most appropriate best-fit evolutionary model (implemented as PROTGAMMALG for DmdA) was predicted with Prottest [43].

Sequence similarity networks

The clusters of predicted DmdA homologs, including sequences from the metagenomes and the single genomes, and closest GcvT homologs were analyzed with SSNs. The set of peptides included in the analysis were those obtained using the DmdA HMM and a relaxed *e*-value (e^{-50}) and contained the two conserved domains in DmdA (PF01571 and PF08669) and only them. The online tool EFI-EST (Enzyme Function Initiative-Enzyme Similarity Tool; ref. [44]) was applied to construct SSNs with default parameters. The methodological process involves two steps. First, it executes an all-by-all BLAST to provide sequence similarities (edges) for all pairs of sequences (nodes). In the second step, a parameter is set to visualize the clusters of

peptides predicted to have the same function. The node-edge pairs are filtered with an alignment score of 105 to generate the SSN as an XGMML file that can be imported into Cytoscape v. 3.5.1 for subsequent visualization, manipulation, and analysis [45].

Structural modeling

The three-dimensional (3D) structures of selected DmdA homologs were predicted based on *P. ubique* strain HTCC1062 DmdA crystal structure [17] using the Iterative Threading ASSEmbly Refinement (I-TASSER) method [46–48]. First, it uses a local meta-threading-server (LOMETS; ref. [49]) to identify templates for the query sequence in the non-redundant Protein Data Bank (PDB; ref. [50]; www.rcsb.org) structure library. Then, the top-ranked template hits obtained are selected for the 3D model simulations. To evaluate positively the global accuracy of the predicted model, a C-score should return between -5 and 2 . At the end, the top 10 structural analogs of the predicted model close to the target in the PDB are generated using TM-align [51]. The TM score value, which scales the structural similarity between two proteins, should return the value 1 if a perfect match between two structures is found or higher than 0.5 when they belong to the same fold family.

Gene arrangement analysis

Syntenic patterns around the DMSP genes were detected by RBH to determine the largest number of genes that were syntenic to representative genomes. Insertion of genes or deletions of up to 10 genes were allowed to find the longest clusters around DMSP genes. Syntenic genes were annotated by running HMMs to determine their ORF membership in families and superfamilies in the Protein Families Database (Pfam) v. 31.0 or TIGRFAM v. 15.0 databases. A hit was considered valid if its score was equal to, or higher than, the recommended “gathering score” for the model.

For some of the genomes in the MAR database, their taxonomic classification did not match the taxonomy of their DmdA peptides based on the phylogenetic trees. For example, five sequences that clustered with the SAR11 were from *Prochlorococcus* genomes. Their contigs that contained the predicted *dmdA* gene were translated with Prodigal and the closest homologs to every peptide on the contig were searched for in RefSeq by blastp. The results were then imported into MEGAN6 (with default parameters; ref. [52]). MEGAN6 places each peptide onto one of the taxa (or “nodes”) of the NCBI Taxonomy, based on the blastp matches, using the lowest common ancestor (LCA) algorithm. This method confirmed that there were contaminating SAR11 contigs in some *Prochlorococcus* genomes that had been assembled from metagenomes. In

other cases, gene arrangements around *dmdA* were different from most other organisms in the same taxon. To confirm lateral gene transfer (LGT) in this case, a similar approach was carried out with the peptides on the same contig. Analysis was assisted with customized Python scripts.

Results

Construction of an HMM profile for identification of DmdA homologs

An HMM profile was built to find homologs of DmdA in organisms across the taxonomic spectrum and in environmental samples (Fig. 1a). The HMM profile was based on the two groups of bacteria for which there is experimental evidence of a DMSP demethylase: *P. ubique* HTCC1062 and *R. pomeroyi*. In total, 216 *Rhodobacteraceae* genomes out of the 386 available and 35 SAR11 genomes out of 61 available in MAR database (Supplementary Fig. 2) contained predicted DmdA sequences based on RBH. These were aligned to make a maximum likelihood tree to remove any false positives. Thus, for example, for two different *Phaeobacter gallaeciensis* strains, a maximum likelihood tree showed two long branches for each of their peptides, away from the other *Rhodobacteraceae* DmdA homologs. These two were considered false positives. This was later confirmed since, in addition to the treeing method, the other methods did not predict any DmdA homolog in these strains. These false positives were removed. With this set of clean sequences, we built the HMM profile (available upon request from the corresponding author).

Since DmdA belongs to the GcvT family of genes, the HMM profile might potentially retrieve other members of the family when using a relaxed cutoff. On the other hand, a very stringent cutoff could miss many of the real DmdA sequences, especially in the environment. With the objective to determine the correct profile cutoff, we ran it with all the genomes of both *Rhodobacteraceae* and SAR11 genomes and built a tree with the sequences (Fig. 1a). A relaxed *e*-value of $e-50$ resulted in a tree that separated the DmdA homologs from other GcvT family genes (Supplementary Fig. 3). With this high *e*-value we retrieved more than one peptide from some genomes known to have only one. For example, two peptides from both *R. pomeroyi* DSS-3 and *P. ubique* HTCC1062 passed this method. It is known from mutagenesis studies that *R. pomeroyi* DSS-3 contains only one *dmdA* homolog in its genome [16]. A trial and error recursive process resulted in an optimal cutoff *e*-value of $e-130$. False positives based on RBH were confirmed in two *P. gallaeciensis* strains. Their peptides clustered further from true DmdAs than the closest GcvT peptides from organisms known not to release methanethiol from DMSP,

Fig. 1 Flowchart of the annotation method. **a** The construction of the profile hidden Markov model (HMM) and parameter selection are described. AAZ21068 and AAV95190 are the GenBank accession numbers for the peptides in *R. pomeroyi* and *P. ubique* HTCC1062 for which there is experimental evidence of their activity as DmdA. The phylogenetic analysis and taxonomic distribution of predicted DmdA homologs is described in **(b)**. DmdA homologs were searched for in the MAR peptide database and *Tara* Oceans peptides. Another approach includes phylogenetic reconstruction of the predicted DmdA peptides, as well as sequence similarity network (SSN) analysis and three-dimensional (3D) modeling as depicted in **(c)**. Syntenic analysis of the neighbor genes around *dmdA* is in **(d)**

for example, *Sulfitobacter* strains. They were also further from the closest GcvT peptide other than DmdA in *R. pomeroyi* DSS-3. For two other *P. gallaeciensis* strains, on the contrary, both RBH and the profile HMM predicted DmdAs. Lastly, we compared our results with 10 strains for which there is experimental evidence of methanethiol release from DMSP and incorporation into cell biomass. In every case, the prediction matched experimental evidence (Table 1; Supplementary Fig. 3).

Sequence similarity network analysis

This method is useful to visualize the relationship among the peptides within the same family (Fig. 1c). The combination of DmdA homologs in the MAR database and *Tara* Oceans peptides clustered together (Fig. 2) and also the MAR DmdA peptides alone (Supplementary Fig. 4). In both cases, they were all more related to each other than to any proteins in the other clusters. These other clusters can be considered to be paralogs. Figure 2 and Supplementary Fig. 4 show the results for the GcvT peptides in the MAR and *Tara* Oceans databases that passed an *e*-value of $e-50$ or lower.

Structural modeling

Since this method is computationally intensive, we selected a few predicted homologs to represent the diversity of DmdAs: *P. ubique* HTCC7211 (first copy, accession number EDZ60447; second copy, EDZ61098), *Rhodobacterales* HTCC2255, *Thioalkalivibrio* HK1, *Gammaproteobacteria* HTCC2080, and *Thioglobus singularis* 662 (Fig. 1c). A non-DmdA peptide from *P. ubique* HTCC1062 (AAZ22069) was also included in the analysis as it was expected to match the structure of an enzyme other than DmdA. This method builds a 3D model of the protein and

then compares its properties with those in the PDB. Results for *P. ubique* HTCC7211 are shown in Table 2. In each case, the top 10 independently identified threading templates had a normalized Z-score greater than one, which indicated a good alignment. The top-scoring templates in PDB with normalized Z scores were DmdA peptides. The final 3D model predicted for all of them resulted in a statistically significant template modeling score greater than 0.91, representing the best structural match. The confidence score greater than 1.61 was within the limits of the threshold set for statistical significance [46]. The 10 best PDB analogs of the model were identified with TM-align [51]. The highest-ranking structural analogs had a TM score greater than 0.89, meaning that the structures belonged to the same fold family. As expected, the three key amino-acid residues known to bind THF in the active site were identified in the alignment of predicted DmdAs. The second residue at position 197 in *P. ubique* HTCC1062 DmdA is required for both THF binding and also to accommodate DMSP (Supplementary Fig. 5; ref. [17]). For some of the sequences in the figure, including those from organisms for which there is experimental evidence of methanethiol release from DMSP, another aromatic amino acid, Tyr, takes its place. In summary, we confirmed that the structure for *P. ubique* apoenzyme DmdA was the closest analog to our predicted models. The closest structural hits for the non-DmdA homolog in the same strain were GcvT other than DmdAs (data not shown).

Number and taxonomy of *dmdA* genes in marine bacteria

DmdA in the genomes of marine bacteria

We applied the HMM profile to a set of peptides recovered from the marine bacterial genomes database (MAR,

Table 1 Prediction of *dmdA* and experimental evidence

Genome	DmdA predicted	Methanethiol release from DMSP	Reference
<i>Ruegeria pomeroyi</i> DSS-3	+	+	[16–18, 65, 66]
<i>Pelagibacter ubique</i> HTCC1062	+	+	[16–18]
<i>Roseovarius nubinhibens</i> ISM	+	+	[65, 66]
<i>Ruegeria lacuscaerulensis</i> ITI-1157	+	+	[66]
<i>Ruegeria atlantica</i> CECT 4292	+	+	[66]
<i>Roseovarius tolerans</i> DSM 11457	+	+	[66]
<i>Sulfitobacter pontiacus</i> ChLG 10	–	–	[65]
<i>Sulfitobacter</i> EE-36	–	–	[65]
<i>Sagittula stellata</i> E-37	–	–	[65]

DmdA was predicted using the profile hidden Markov model (HMM) and supported by other methods. Direct evidence for *dmdA* exists for the first two organisms

Fig. 2 Sequence similarity networks of DmdA homologs in the genomes of marine bacteria from the MAR database and *Tara* Oceans peptides. The peptides for the analysis were obtained after running the profile hidden Markov model (HMM) with a relaxed *e*-value ($e=50$).

DmdA homologs are indicated, and include the DmdA homologs for which there is experimental evidence of their activity as DmdA. The remaining clusters are considered paralogs

Table 2 Results of the three-dimensional (3D) model prediction

Rank	Class	Gene name	Organism	PDB ID	Iden1 (%) ^a	Iden2 (%) ^b	Cov ^c	Norm. Z-score ^d
1	AMT	<i>dmdA</i>	<i>P. ubique</i> (HTCC1062)	3tfhA	70	70	1.00	4.95
2	AMT	<i>dmdA</i>	<i>P. ubique</i> (HTCC1062)	3tfhA	70	70	1.00	5.61
3	AMT	<i>dmdA</i>	<i>P. ubique</i> (HTCC1062)	3tfhA	70	70	1.00	5.78
4	AMT	<i>dmdA</i>	<i>P. ubique</i> (HTCC1062)	3tfh	69	70	0.98	3.33
5	Oxido-reductase	<i>dmgdH</i> ^e	<i>Rattus norvegicus</i>	4p9sA	20	21	0.98	2.41
6	AMT	<i>dmdA</i>	<i>P. ubique</i> (HTCC1062)	3tfhA	70	70	1.00	6.17
7	AMT	<i>dmdA</i>	<i>P. ubique</i> (HTCC1062)	3tfh	66	70	0.99	3.65
8	AMT	<i>gcvT</i>	<i>Bacillus subtilis</i> (168)	1yx2A	26	26	0.96	6.47
9	AMT	<i>dmdA</i>	<i>P. ubique</i> (HTCC1062)	3tfhA	70	70	1.00	7.98
10	AMT	<i>dmdA</i>	<i>P. ubique</i> (HTCC1062)	3tfhA	70	70	1.00	7.00

Shown are the top 10 template hits to *Pelagibacter* strain HTCC7211 DmdA (first copy of the gene) identified by the package LOMETS, which make use of threading alignments

AMT aminomethyltransferase, PDB Protein Data Bank, Cov. coverage, Norm. Z-score normalized Z-score

^aIden1 is the percentage sequence identity of the templates in the threading aligned region with the query sequence

^bIden2 is the percentage sequence identity of the whole template chains with query sequence

^cCoverage of the threading alignment (calculated as the number of aligned residues divided by the length of the query protein)

^dNormalized Z-score of the threading alignments. Alignment with a normalized Z-score > 1 means a good alignment and vice versa

^eDimethylglycine dehydrogenase complexed with tetrahydrofolate

Fig. 1b). Out of 61 SAR11 genomes there were 35 with at least one *dmdA*. However, only 14 genomes were from cultures, while 41 were SAGs and six were MAGs. Therefore, the real number of SAR11 with *dmdA* gene is likely greater than this since most of the genomes were incomplete. There were two main groups of SAR11 DmdAs, based on the maximum likelihood phylogenetic reconstruction (Fig. 3) and supported by a Bayesian method as the two trees were topologically identical (Supplementary Fig. 6). Eight of the SAR11 genomes that contained the first copy of the gene also had a second copy. For example, the strain HTCC1062 contained one copy of the gene, and HTCC7211 contained two. But no complete genome contained the *dmdA* second copy by itself. In the case of *Rhodobacteraceae*, 213 strains had at least one copy of predicted *dmdA* out of 386 genomes. All except one of the 60 *Ruegeria* genomes contained the *dmdA* gene. *Leisingera* ANG-Vp stood out as it contained three copies of the gene. Two of them clustered with the rest of the *Rhodobacteraceae* DmdAs and a third clustered with other *Rhodobacteraceae* and *Kiloniella spongiae* (labeled as such in

Fig. 3). Six other *Rhodobacteraceae* genomes contained two copies of the gene.

Other groups of bacteria also contain the *dmdA* gene. As previously described [53], SAR116 genomes contained *dmdA* homologs, represented by a cluster with four sequences (*Puniceispirillum marinum* IMCC1322, READ_SEA-S10_B10N8, casp-alpha9, and casp-alpha10) and another cluster with the sequence from strain HIMB100. This last sequence grouped with the sequence of *Rhodobacterales* strain HTCC2255 and the *Rhodobacteraceae* *Amylibacter* sp. 4G11. SAR324 in the *Deltaproteobacteria* formed two clusters. DmdA homologs were also found in *Gammaproteobacteria* represented by *Thioglobus* (SUP05 clade), HTCC2080 (OM60/NOR5 clade), and *Thioalkalivibrio* (*Ectothiorhodospiraceae*). *Thioalkalivibrio* is a chemolithoautotrophic organism able to obtain reducing equivalents from inorganic sulfur among other compounds [54]. *Thioglobus* is also a versatile organism involved in the transformation of sulfur in suboxic marine environments [55].

DmdA in environmental sequences

The HMM profile was next used with a new dataset containing both the MAR genomes and the Tara Oceans metagenomes (Fig. 1b). The clusters for single genomes found in the previous step appeared again with additional sequences from the environment (Fig. 4). In addition, completely new clusters and subclusters were found, indicating that, as could be expected, the available genomes in databases underestimate real diversity. Most of the sequences belonged to SAR11, represented by four subclusters each with either no or only a few sequences from single genomes from the MAR database. Thus, SAR11 subclusters 3 and 4 did not contain any representative sequence in the MAR database. The bacterial group with the second largest fraction of *dmdA* genes was the *Rhodobacteraceae*. The *Rhodobacteraceae* DmdAs were dominated by sequences from isolates, most of which grouped together, farther from assembled sequences from metagenomes (Fig. 5). A subcluster within the *Rhodobacteraceae* contained *Alphaproteobacteria*, other than *Rhodobacteraceae*, and *Gammaproteobacteria* sequences from the MAR database (labeled as TMED51/TMED61/*Chromatiales*). This suggests LGT between *Rhodobacteraceae* and *Gammaproteobacteria*.

Figure 4 also shows other instances of likely LGT between distantly related taxa. This was likely the case of strain HTCC2255 (distantly related to *Ruegeria* and other *Rhodobacteraceae*), SAR116 bacteria, and *Amylibacter* strain 4G11 (*Rhodobacteraceae*), whose sequences grouped together. The other strain of *Amylibacter* (*Amylibacter cionae*), however, clustered with *R. pomeroyi*. The distribution of SAR324 (*Deltaproteobacteria*) and *Gammaproteobacteria* also

Fig. 3 Maximum likelihood tree of the DmdA homologs found in the genomes of marine bacteria (MAR database). Accession numbers are shown between parentheses for individual sequences. Major clusters are collapsed. The sequence of *Oceanospirillaceae* ASP10-02a is predicted to encode two DmdA peptides whose genes are in tandem. The sequence of *P. ubique* HTCC1062 with accession number AAZ22069 (a non-DmdA GcvT) served as the outgroup (not shown). Circles at nodes are bootstrap values greater than 70 (100 replicates) with a diameter proportional to their value. The scale bar indicates substitutions per site

Fig. 4 Maximum likelihood tree of the *DmdA* homologs found in both the single genomes (MAR database) and *Tara* Oceans sequences. Orange and blue circles are the proportion of *DmdA* homologs in MAR genomes (blue) vs. *Tara* Oceans sequences (orange). The sizes of the circles are proportional to the number of sequences in each

cluster and subcluster. Green circles at nodes are bootstrap values greater than 70 (100 replicates) with a diameter proportional to their value. The sequence of *P. ubique* HTCC1062 with accession number AAZ22069 (a non-*DmdA*, GcvT) served as the outgroup (not shown). The scale bar indicates substitutions per site

suggested LGT, since their peptides cluster together in two different groups. One of these groups, Cluster “A” on Fig. 4, also contained peptides from *Flavobacteriia*. To rule out mis-assembly, we examined the *Tara* Oceans contigs containing these genes and found that a few on either side of the *dmdA* gene were affiliated with *Gammaproteobacteria*, but all the other genes in such contigs belonged to *Flavobacteriia* or *Deltaproteobacteria*, based on the closest hits against RefSeq.

We were concerned that the sequences assembled from metagenomes would be of less quality and might distort the topologies of the phylogenetic trees. Then again, removing *Tara* Oceans sequences might miss some of the diversity in the natural environment. We assumed that it would be highly improbable that two erroneous peptides assembled from the metagenome would have the exact same sequence. Therefore, we built phylogenetic trees of the *DmdA* from the MAR database of genomes and sequences that had at least two copies in the *Tara* Oceans (non-singletons; Supplementary Figures 7–9 for all sequences, SAR11 or *Rhodobacteraceae* sequences, respectively). The topologies did not change, suggesting that no artifacts were introduced in the analyses. In turn, the same main groups of sequences were found when singleton sequences were included. In addition, eliminating singleton sequences did not have

much effect on the overall sequence diversity. Except for subcluster 4, the rest of SAR11 subclusters were also present (Supplementary Fig. 8) but most of the environmental sequences in the *Rhodobacteraceae* were removed (Supplementary Fig. 9) since the great majority of its members were isolated bacteria and there were only two *Tara* Oceans sequences that were the exact same sequence. Thus, we believe the methods and analysis were robust.

Quantification of *dmdA* reads in environmental samples yielded higher counts than in previous studies. On average, we found 3.4 times more counts than in a previous study for the same samples (GOS study; ref. [34]; Supplementary Fig. 10 and Supplementary Materials and methods). With our approach, for some of the samples, the number of hits to *dmdA* genes was higher than to *recA* genes. However, SAR11 members usually have two copies per genome and other bacteria might also carry more than one copy. The latter are not as easy to identify as in SAR11, whose second copies of *DmdA* group together and away from the main copy. Leaving out the second copy of the gene in SAR11, we estimate that 78% of the genomes in the bacterioplankton contained at least one copy of the *dmdA* gene on average. This percentage is higher than that of rhodopsin genes that we estimated to be 46%. However, we have to

Fig. 5 Maximum likelihood tree of the DmdA homologs in *Rhodobacteraceae* genomes and *Tara* Oceans sequences. Blue branches represent sequences from either *Tara* Oceans or assembled sequences from metagenomes in the MAR database. Branches in red are *Ruegeria* sequences. *Labrenzia* and *Jannaschia* are indicated for reference, since their sequences each cluster together further from most other *Rhodobacteraceae* sequences in genomes. The positions of *R.*

pomeroyi and *Roseobacter denitrificans* are also shown as reference. Green circles at nodes are bootstrap values greater than 70 (100 replicates) with a diameter proportional to their value. The sequence of *P. ubique* HTCC1062 with accession number AAZ22069 served as the outgroup (a non-DmdA; not shown). The scale bar indicates substitutions per site

consider it a maximal value, given that the number of other bacteria with two or more copies is unknown.

If we consider only hits to the main copy of SAR11 DmdA and to *Rhodobacteraceae* sequences, then the average percent of bacteria with *dmdA* gene is 36%, which is similar to the percentage estimated by Howard et al. [16]. Thus, the novelty of our approach consisted in retrieving the second copy of SAR11 and genes from groups previously not known to have *dmdA*, such as *Gammaproteobacteria* (in particular *Thioglobus*) and *Alphaproteobacteria* (different from SAR116 and *Rhodobacteraceae*).

Gene arrangements

It is common to find synteny within taxonomic groups. Genes that remain together can be regulated in concert or they may have arrived by LGT. Therefore, examining the neighborhood of the target gene may support the previous methods. This may be particularly useful when looking at environmental sequences, frequently of less quality than those from cultures and certainly shorter. We analyzed this possibility looking at the genes in the neighborhood of the identified *dmdA* sequences of the clusters most abundant in

the *Tara* Oceans metagenomes. Figure 6 shows the main patterns of synteny in SAR11 and *Rhodobacteraceae*, along with their distribution in the genomes and contigs analyzed.

In the case of the SAR11 cluster we found three different neighborhoods. First, subclusters 1 and 2 (Fig. 4) showed a robust synteny. Most SAR11 contained *dmdA* next to an alpha/beta hydrolase family protein gene of unknown function, followed by *dmdB* and *dmdC*. The last two encode enzymes in the two following steps in the DMSP demethylation pathway (see metabolic pathway in Supplementary Fig. 1 and gene arrangement in *P. ubique* HTCC1062 in Fig. 6A). In a few cases, the alpha/beta hydrolase family protein gene was missing from the genome by RBH, but *dmdB* and *dmdC* were present. The remaining genomes were not complete and lacked the region around *dmdA*, as they were genomes in the MAR database that were assembled from metagenomes.

Five genomes in these two subclusters were complete for the same region and did not have the *dmdA* gene. In these cases the whole set of four genes was absent (see *P. ubique* HTCC9063 in Fig. 6a). Moreover, when present, this set of genes was always within a collection of housekeeping genes, including cytochrome *c*, ATPase, ribosomal proteins, DNA

Fig. 6 Gene alignments in representative sequences of the SAR11 group and the *Rhodobacteraceae*. **a** The position of the first copy of *dmdA* is shown. IMCC9063 does not contain any *dmdA* gene in its genome. **b** Represented are two contigs in the SAR11 subcluster 3, one with and another without *dmdA*. The second copy of *dmdA* in SAR11 is shown in (c). HTCC1062 does not contain the second copy of *dmdA*. **d** The most common arrangement around *dmdA* in *Rhodobacteraceae*. For this family, the region shows a cluster of genes that

are shared between *R. pomeroyi* DSS-3, *R. denitrificans* Och 114, and *Sulfitobacter pontiacus* DSM 10014. The latter does not contain the *dmdA* gene but it does contain *acul*. White genes are ATPase synthase genes, Krebs cycle in yellow, amino-acid metabolism in red, ribosomal protein genes in orange, transporter genes in black, DNA replication in maroon, in gray genes that encode a transporter for DMSP and related molecules, and other genes in blue. The scale bar represents kilobase pairs

polymerase, and glycolysis and Krebs cycle genes between approximately positions 194,000 and 288,000 using the *P. ubique* HTCC1062 genome as reference (Fig. 6a). Within this neighborhood, the *dmdA* gene was always at the same position in the genome, starting at nucleotide 248,657 in *P. ubique* HTCC1062. The *dmdA* gene neighborhood that is characteristic of the SAR11 subclusters 1 and 2 was identical in the *Tara* Oceans contigs as far as the length of each of contig could span.

Second, in SAR11 subclusters 3 and 4 (Fig. 4), the *dmdA* gene was approximately at position 985,000 in *P. ubique* HTCC1062. Figure 6b highlights the syntenic pattern around this region. The same alpha/beta hydrolase family protein gene as in subclusters 1 and 2 was present between *dmdA* and *dmdB*, along with a 3-hydroxyl-CoA dehydrogenase gene, both of which might also be involved in the metabolic pathway for DMSP transformation. Finally, some SAR11 genomes in MAR database had a second copy of *dmdA* that formed a very distant cluster together with environmental sequences from *Tara* Oceans (cluster

SAR11 second copy in Figs. 3 and 4). In this case the gene was surrounded by 12 genes in a conserved region. Most of these genes were transporter genes (see *P. ubique* HTCC7211 in Fig. 6c). These genes were absent in SAR11 genomes without the second copy of *dmdA* (see *P. ubique* HTCC1062 in Fig. 6c). Some degree of synteny was observed in the region of the second copy of *dmdA* in SAR11, SAR116, and *Alphaproteobacteria* clusters (Supplementary Fig. 11).

We also analyzed synteny for the different *Rhodobacteraceae* members. Whenever *dmdA* was present as single copy, including *R. pomeroyi*, in most cases it was next to, or one gene away from, *acul* (Fig. 6d). There were a few exceptions as detailed in Supplementary Table 1. These include the sequences in the subclusters NAT102 (a *Rhodobacteraceae* assembled from *Tara* Oceans metagenomes; accession number PACC00000000; ref. [56]) and those sequences related to *Proteobacteria* (TMED51/TMED61/*Chromatiales*) in Fig. 5. In NAT102 and related sequences in the *Tara* Oceans dataset, next to *dmdA* there were 10

genes that are part of the SOX system to oxidize inorganic sulfur compounds (Supplementary Fig. 12). The TMED51/TMED61/*Chromatiales* subcluster seems to be another case of LGT since TMED51 and TMED61 are not classified as *Rhodobacteraceae*. For every sequence, including those from the Tara Oceans set of sequences, this was confirmed by doing BLAST against RefSeq. Therefore, there is a representation of *Gammaproteobacteria*, most closely related to *Chromatiales*, within the *Rhodobacteraceae*. In addition, sequences classified as *Rhodobacteraceae*, but whose DmdA peptides clustered instead with other *Alphaproteobacteria* (labeled as such in Fig. 4), did not contain *acul* next to *dmdA*.

The *dmdA*–*acul* pair, when present, was in a region of variable length with a fair degree of synteny that span an average of 12 genes, despite the relatively long phylogenetic distance between members of the different subclusters (Fig. 6d). The gene product AcuI is an enzyme involved in the degradation of acrylate, a product of the competing pathway for the transformation of DMSP into DMS (Supplementary Fig. 1; ref. [57]). AcuI is NADPH-dependent and reduces acryloyl-CoA to propionyl-CoA, which is further transformed through several steps into succinyl-CoA before it enters the Krebs cycle [58]. The tandem *dmdA*–*acul* was also found in one more microorganism outside the *Rhodobacteraceae*, *Kiloniella* P1-1 (*Alphaproteobacteria*, *Kiloniellales*). The tandem *dmdA*–*acul* was also present in the Tara Oceans peptides classified in the *Rhodobacteraceae* cluster (Fig. 4).

As an example of the usefulness of gene neighborhood to annotate genes, we looked at *dmdB* and *dmdC*. DmdB has two conserved domains to bind adenosine monophosphate (AMP). We searched the MAR database for their corresponding PFAMs, PF00501 and PF13193. From the obtained sequences we then selected those that were within five gene distance from *dmdA*, assuming these would likely correspond to *dmdB*. The tree built with these sequences was congruent with the *dmdA* tree, thus giving confidence to the annotation (Supplementary Fig. 13). A similar approach for DmdC rendered similar results as shown in Supplementary Fig. 14. The gene for this last enzyme was retrieved using one of the conserved domains (middle domain of the acyl-CoA dehydrogenase; PF02770). The presence of the alpha/beta hydrolase family protein gene is intriguing since it is found next to *dmdA* in all SAR11 subclusters. It was missing in some SAR11 contigs assembled from metagenomes, but at least in one case in the SAR11 genomes in MAR, the gene was somewhere else in the genome by RBH. The four strains of *Thioglobus* did not contain *dmdA* next to *dmdC* and *dmdB* but somewhere else in the genome. Predicted sequences in *Thioglobus* also grouped with the rest of DmdC and DmdB peptides (Supplementary Figures 13 and 14). The tandem *dmdC* and

dmdB cluster with a 3-hydroxyacyl-CoA dehydrogenase gene followed by the alpha/beta hydrolase family gene and *metY* (Supplementary Fig. 15). *metY* encodes *O*-acetylhomoserine aminocarboxypropyltransferase, involved in the incorporation of methanethiol or sulfide into Met [13]. The same cluster of genes was found in four *Gammaproteobacteria* SAGs likely to be *Thioglobus* strains. The *dmdA* gene was not predicted in *Thioglobus* strain EF1 and, in this case, *metY* was found, but not the other four genes downstream of *metY*. The alpha/beta hydrolase family gene was also present next to *dmdA* and *metY* in the genomes of HTCC2255 and *Amylibacter* sp. 4G11 (Supplementary Fig. 15). This suggests an unknown role of the alpha/beta hydrolase family gene in the pathway for the transformation of DMSP.

Discussion

Genes involved in DMSP transformation play a pivotal role in the sulfur cycle of the oceans [16]. The first reaction in the breakdown of this molecule decides the route to follow at the bifurcation of the pathways [13]. Given its central role in a key biogeochemical process, there is great interest in the diversity and expression of the genes involved in DMSP transformation. Indeed, focusing on the expression of marker genes is a means to study the fate of DMSP. However, like most other genes in any genome, DmdA belongs to a family of peptides with a wide range of substrates, and paralogs are a major source of errors in gene annotation. The usual quantification of a gene like *dmdA* is most likely an overestimate, since the methods that rely on sequence similarities tend to over-annotate. For example, as a case in point, the automatic annotation of the MAR database predicted twice as many DmdA peptides as the profile HMM method developed in this study.

Certainly, one of the most basic computational approaches for function prediction has been pairwise sequence alignment to identify proteins with high sequence similarity and known functions. The function is assigned based on the best matching protein in a database, after complying with some alignment statistics, and is then considered to be an ortholog. Other methods that rely on sequence alignments are more comprehensive and include RBH [59, 60] and analyses based on similarity networks [45, 61]. The advantage of such methods is that they rely less on arbitrary parameters set a priori by the researcher. Profile methods, on the other hand, are based on the alignment of a problem peptide against an alignment of a curated set of peptides, belonging to a family or subfamily. Since orthologs follow different evolutionary paths from paralogs in the same protein family, then we assume that enzymes with the same function cluster together in a phylogenetic reconstruction.

The alignment of GcvTs that cluster with known DmdA peptides is the basis to build a profile HMM. The profile contains information on the diversity of each of the amino-acid positions along the alignment, which provides richer information than pairwise sequence comparisons. This is advantageous, especially in the case of taxa that are poorly represented in databases or those distant from the closest described homologs [37]. Lastly, considering that protein tertiary structure is expected to influence its function, the annotation method may take into consideration the 3D structure of the peptide. However, prediction of protein structure remains challenging, besides being computationally intensive.

In this study, we have developed a procedure that improves the annotation of *dmdA* and other genes down the DMSP demethylation pathway integrating several methods. The pipeline, with modifications, can be applied to other genes of environmental significance. The goal is to develop a method that relies less on the subjectivity of the researcher, although it requires a priori detailed knowledge of the molecular diversity of the enzyme in the environment. To this end, the annotation method starts off from the two DmdA homologs for which there is experimental evidence and builds from there. A profile HMM offers the possibility of finding new DmdA homologs in taxonomic groups where they have not been found before. The gene neighborhood and/or position of the gene along the genome help confirm the annotation of the gene. However, the application of syntenic patterns is not universal, since it depends on the taxonomic group.

We found most DmdA homologs in bacterial groups already known to harbor them, namely SAR11, *Rhodobacteraceae*, *Gammaproteobacteria*, SAR324, and SAR116. The HMM profile, however, also retrieved new DmdA homologs from other taxonomic groups, mostly in the *Alphaproteobacteria*. The SAR11 DmdA peptides clustered into two main groups. Most of the sequences were found in the main cluster, which could be divided in several sub-clusters. The main cluster contained most of the SAR11 sequences with no representatives from other groups, while the second copy was closely related to other *Alphaproteobacteria*, including SAR116 and some *Rhodobacteraceae* with second or third copies of their *dmdA* genes, suggesting an early LGT event.

Gene neighborhood makes a distinction between the main *Rhodobacteraceae* group of sequences and the position of other *Alphaproteobacteria*, which include the second copy in SAR11, *Rhodobacteraceae* other than those in the main cluster, SAR116, and other *Alphaproteobacteria*. Considering the close association of these groups of sequences, one needs to be careful in drawing conclusions when quantifying, for example, the presence or expression of the *dmdA* gene in taxonomic groups that share a recent common ancestor. The close proximity of sequences in

these groups might give rise to errors in the taxonomic assignments when searching in a database. These errors would be less expected in assignment to the SAR11 group of sequences, or its subclusters, whose members are clearly separated from the rest.

Presence of a conserved unknown neighbor gene suggests that it may participate in the same transformation pathway. This is a widely recognized property of prokaryotic genomes, i.e., genes that are maintained adjacent through evolutionary time tend to be associated functionally [62, 63]. The association of these genes responds to an organization level above operon structures. This information is useful to characterize genes involved in the same pathway. For example, a number of transporter genes are associated with the presence of the second copy of *dmdA* in SAR11 bacteria, indicating good candidates for novel genes involved in DMSP uptake. Conservation of the position of *dmdA* in SAR11 genomes is particularly striking, considering the breadth of diversity within this group, for example, based on ribosomal RNA sequences. One can deduce that no LGT would be functional unless a given gene occupies the right position, not just next to the “right” neighbor genes but also the “right” location along the genome. Regarding this as an annotation method, the localization of the gene in the genome and gene neighbors is a telltale sign of its function, consistent with the other approaches described in this study.

For environmental studies, such as quantification of a gene and its taxonomic description, we propose assigning identifications to the clusters of sequences based on the combination of both taxonomy and gene neighborhood. For example, the main set of SAR11 DmdA peptides could be divided into four groups. However, the *Rhodobacteraceae* require some more consideration. They could be divided into two main groups based on the gene neighborhoods: NAT102 and closely related sequences would be one, and the remaining *Rhodobacteraceae* sequences would be another. In addition, some *Rhodobacteraceae* DmdAs cluster with the *Alphaproteobacteria* other than those classified in *Rhodobacteraceae* and some *Gammaproteobacteria* DmdAs cluster with the *Rhodobacteraceae*.

The method developed here has been successful in retrieving a wide diversity of *dmdA* genes (Fig. 3). The gene is found in SAR11, *Rhodobacteraceae*, SAR116, and *Gammaproteobacteria* as previously reported. A closer look at the taxonomic classification of the *Gammaproteobacteria*, for example, shows that it is represented in six different clusters, by themselves, or with *Alphaproteobacteria*, *Deltaproteobacteria*, and even with *Flavobacteriia* that presumably acquired the gene through LGT. SAR324 homologs belong to two clusters and SAR116-related sequences to three. Additionally, a substantial fraction of the predicted *dmdA* homologs have been found in

sequences of unknown organisms, with no representatives described yet. However, what is remarkable is that most other groups do not have the gene. Aside from the few *Flavobacteriia* just mentioned, most *Bacteroidetes* and all *Cyanobacteria* seem to lack the gene. These are the two most abundant bacterial phyla in the oceans after the *Proteobacteria*. One would expect that the occurrence of *dmdA* in the genomes of bacterioplankton cells would be advantageous, allowing them to incorporate reduced sulfur directly to their proteins [13]. In fact, the cyanobacterial genera *Synechococcus* and *Prochlorococcus* were shown to assimilate DMSP [64]. However, none of the *dmdA* sequences belonged to these groups. Unless there is an alternative pathway to obtain S from DMSP, nothing is known as to how other marine groups metabolize sulfur from an organic compound such as DMSP [65, 66].

In summary, the *dmdA* gene is found, out of the many different phyla in the oceans, in relatively few bacterial taxa. Yet, these few dominate the bacterioplankton. SAR11 bacteria, in particular, have evolved to accommodate even a second *dmdA* gene in their reduced genomes. It remains to be elucidated how each of the copies functions in the physiology of the cell. The method described in this paper starts from whatever little information the experimental data provide, and makes use of the deluge of sequence information from metagenomic studies. This procedure guides the researcher along the right path to discover genes involved in given metabolic pathways, and could be applied to any gene of environmental importance.

Acknowledgements This research was supported by grant CTM2016-80095-C2 from the Spanish Ministry of Economy and Competitiveness.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

- Malin G, Kirst GO. Algal production of dimethyl sulfide and its atmospheric role. *J Phycol.* 1997;33:889–96.
- Kettle AJ, Andreae MO. Flux of dimethylsulfide from the oceans: a comparison of updated data seas and flux models. *J Geophys Res Atmos.* 2000;105:26793–808.
- Bates TS, Charlson RJ, Gammon RH. Evidence for the climatic role of marine biogenic sulphur. *Nature.* 1987;329:319–21.
- Charlson RJ, Lovelock JE, Andreae MO, Warren SG. Oceanic phytoplankton, atmospheric sulfur, cloud albedo and climate. *Nature.* 1987;326:655–61.
- Simó R. Production of atmospheric sulfur by oceanic plankton: biogeochemical, ecological and evolutionary links. *Trends Ecol Evol.* 2001;16:287–94.
- Andreae MO. Ocean-atmosphere interactions in the global biogeochemical sulfur cycle. *Mar Chem.* 1990;30:1–29.
- Malin G, Turner SM, Liss PS. Sulfur: the plankton/climate connection. *J Phycol.* 1992;28:590–7.
- Taylor BF, Gilchrist DC. New routes for aerobic biodegradation of dimethylsulfoniopropionate. *Appl Environ Microbiol.* 1991;57:3581–4.
- Visscher PT, Diaz MR, Taylor BF. Enumeration of bacteria which cleave or demethylate dimethylsulfoniopropionate in the Caribbean Sea. *Mar Ecol Prog Ser.* 1992;89:293–6.
- Taylor BF. Bacterial transformations of organic sulfur compounds in marine environments. In: Oremland RS, editor. *Biogeochemistry of global change*. Boston, MA: Springer; 1993. p.745–81.
- Kiene RP. Production of methanethiol from dimethylsulfoniopropionate in marine surface waters. *Mar Chem.* 1996;54:69–83.
- Kiene RP. Turnover of dissolved DMSP in estuarine and shelf waters of the northern Gulf of Mexico. In: Kiene RP, Visscher PT, Keller MD, Kirst GO, editors. *Biological and environmental chemistry of DMSP and related sulfonium compounds*. Boston, MA: Springer; 1996. p. 337–49.
- Kiene RP, Linn LJ, González J, Moran MA, Bruton JA. Dimethylsulfoniopropionate and methanethiol are important precursors of methionine and protein-sulfur in marine bacterioplankton. *Appl Environ Microbiol.* 1999;65:4549–58.
- Curson ARJ, Todd JD, Sullivan MJ, Johnston AWB. Catabolism of dimethylsulphoniopropionate: microorganisms, enzymes and genes. *Nat Rev Microbiol.* 2011;9:849–59.
- Reisch CR, Moran MA, Whitman WB. Bacterial catabolism of dimethylsulfoniopropionate (DMSP). *Front Microbiol.* 2011;2:172.
- Howard EC, Henriksen JR, Buchan A, Reisch CR, Bürgmann H, Welsh R, et al. Bacterial taxa that limit sulfur flux from the ocean. *Science.* 2006;314:649–52.
- Schuller DJ, Reisch CR, Moran MA, Whitman WB, Lanzilotta WN. Structures of dimethylsulfoniopropionate-dependent demethylase from the marine organism *Pelagibacter ubique*. *Protein Sci.* 2012;21:289–98.
- Reisch CR, Moran MA, Whitman WB. Dimethylsulfoniopropionate-dependent demethylase (DmdA) from *Pelagibacter ubique* and *Silicibacter pomeroyi*. *J Bacteriol.* 2008;190:8018–24.
- Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinforma.* 2010;11:119.
- Overbeek R, Bartels D, Vonstein V, Meyer F. Annotation of bacterial and archaeal genomes: improving accuracy and consistency. *Chem Rev.* 2007;107:3431–47.
- Overbeek R, Olson R, Pusch GD, Olsen GJ, Davis JJ, Disz T, et al. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Res.* 2014;42:D206–14.
- Galperin MY, Koonin EV. Sources of systematic error in functional annotation of genomes: domain rearrangement, non-orthologous gene displacement and operon disruption. In *Silico Biol.* 1998;1:55–67.
- Devos D, Valencia A. Intrinsic errors in genome annotation. *Trends Genet.* 2001;17:429–31.
- Poux S, Magrane M, Arighi CN, Bridge A, O'Donovan C, Laiho K, et al. Expert curation in UniProtKB: a case study on dealing with conflicting and erroneous data. *Database (Oxford).* 2014;2014:bau016.
- Bryson K, Loux V, Bossy R, Nicolas P, Chaillou S, van de Guchte M, et al. AGMIAL: implementing an annotation strategy for prokaryote genomes as a distributed system. *Nucleic Acids Res.* 2006;34:3533–45.
- Meyer F, Goesmann A, McHardy AC, Bartels D, Bekel T, Clausen J, et al. GenDB – an open source genome annotation

- system for prokaryote genomes. *Nucleic Acids Res.* 2003;31:2187–95.
27. Schnoes AM, Brown SD, Dodevski I, Babbitt PC. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS Comput Biol.* 2009;5:e1000605.
 28. Fitch WM. Distinguishing homologous from analogous proteins. *Syst Zool.* 1970;19:99–113.
 29. O’Leary NA, Wright MW, Brister JR, Ciufu S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* 2016;44:D733–45.
 30. Klemetsen T, Raknes IA, Fu J, Agafonov A, Balasundaram SV, Tartari G, et al. The MAR databases: development and implementation of databases specific for marine metagenomics. *Nucleic Acids Res.* 2018;46:D692–9.
 31. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics.* 2014;30:2068–9.
 32. Eren AM. <https://doi.org/10.6084/m9.figshare.4902920.v1> (2017). Accessed on August 2017.
 33. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol.* 2015;32:268–74.
 34. Howard EC, Sun S, Biers EJ, Moran MA. Abundant and diverse bacteria involved in DMSP degradation in marine surface waters. *Environ Microbiol.* 2008;10:2397–410.
 35. Howard EC, Sun S, Reisch CR, del Valle DA, Bürgmann H, Kiene RP, et al. Changes in dimethylsulfoniopropionate demethylase gene assemblages in response to an induced phytoplankton bloom. *Appl Environ Microbiol.* 2011;77:524–31.
 36. Varaljay VA, Howard EC, Sun S, Moran MA. Deep sequencing of a dimethylsulfoniopropionate-degrading gene (*dmdA*) by using PCR primer pairs designed on the basis of marine metagenomic data. *Appl Environ Microbiol.* 2010;76:609–17.
 37. Eddy SR. Profile hidden Markov models. *Bioinformatics.* 1998;14:755–63.
 38. Madera M, Gough J. A comparison of profile hidden Markov model procedures for remote homology detection. *Nucleic Acids Res.* 2002;30:4321–8.
 39. Eddy SR. A probabilistic model of local sequence alignment that simplifies statistical significance estimation. *PLoS Comput Biol.* 2008;4:e1000069.
 40. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 2004;32:1792–7.
 41. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics.* 2009;25:1972–3.
 42. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics.* 2014;30:1312–3.
 43. Darriba D, Taboada GL, Doallo R, Posada D. ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics.* 2011;27:1164–5.
 44. Gerlt JA, Bouvier JT, Davidson DB, Imker HJ, Sadkhin B, Slater DR, et al. Enzyme function initiative-enzyme similarity tool (EFI-EST): a web tool for generating protein sequence similarity networks. *Biochim Biophys Acta.* 2015;1854:1019–37.
 45. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 2003;13:2498–504.
 46. Zhang Y. I-TASSER server for protein 3D structure prediction. *BMC Bioinforma.* 2008;9:40.
 47. Roy A, Kucukural A, Zhang Y. I-TASSER: a unified platform for automated protein structure and function prediction. *Nat Protoc.* 2010;5:725–38.
 48. Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y. The I-TASSER suite: protein structure and function prediction. *Nat Methods.* 2015;12:7–8.
 49. Wu S, Zhang Y. LOMETS: a local meta-threading-server for protein structure prediction. *Nucleic Acids Res.* 2007;35:3375–82.
 50. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The protein data bank. *Nucleic Acids Res.* 2000;28:235–42.
 51. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* 2005;33:2302–9.
 52. Huson DH, Beier S, Flade I, Gorska A, El-Hadidi M, Mitra S, et al. MEGAN community edition - interactive exploration and analysis of large-scale microbiome sequencing data. *PLoS Comput Biol.* 2016;12:e1004957.
 53. Oh H-M, Kwon KK, Kang I, Kang SG, Lee J-H, Kim S-J, et al. Complete genome sequence of “*Candidatus Puniceispirillum marinum*” IMCC1322, a representative of the SAR116 clade in the Alphaproteobacteria. *J Bacteriol.* 2010;192:3240–1.
 54. Ahn A-C, Meier-Kolthoff JP, Overmars L, Richter M, Woyke T, Sorokin DY, et al. Genomic diversity within the haloalkaliphilic genus *Thioalkalivibrio*. *PLoS ONE.* 2017;12:e0173517.
 55. Mattes TE, Brook L, Nunn BL, Marshall KT, Proskurowski G, Kelley DS, et al. Sulfur oxidizers dominate carbon fixation at a biogeochemical hot spot in the dark ocean. *ISME J.* 2013;7:2349–60.
 56. Graham ED, Heidelberg JF, Tully BJ. Potential for primary productivity in a globally-distributed bacterial phototroph. *ISME J.* 2018;12:1861–6.
 57. Sullivan MJ, Curson ARJ, Shearer N, Todd JD, Green RT, Johnston AWB. Unusual regulation of a leaderless operon involved in the catabolism of dimethylsulfoniopropionate in *Rhodospirillum rubrum*. *PLoS ONE.* 2011;6:e15972.
 58. Schneider K, Asao M, Carter MS, Albera BE. *Rhodospirillum rubrum* uses a reductive route via propionyl coenzyme A to assimilate 3-hydroxypropionate. *J Bacteriol.* 2012;194:225–32.
 59. Tatusov RL, Koonin EV, Lipman DJ. A genomic perspective on protein families. *Science.* 1997;278:631–7.
 60. Bork P, Dandekar T, Diaz-Lazcoz Y, Eisenhaber F, Huynen M, Yuan Y. Predicting function: from genes to genomes and back. *J Mol Biol.* 1998;283:707–25.
 61. Atkinson HJ, Morris JH, Ferrin TE, Babbitt PC. Using sequence similarity networks for visualization of relationships across diverse protein superfamilies. *PLoS ONE.* 2009;4:e4345.
 62. Yin Y, Zhang H, Olman V, Xu Y. Genomic arrangement of bacterial operons is constrained by biological pathways encoded in the genome. *Proc Natl Acad Sci USA.* 2010;107:6310–5.
 63. Zhao S, Kumar R, Sakai A, Vetting MW, Wood BM, Brown S, et al. Discovery of new enzymes and metabolic pathways by using structure and genome context. *Nature.* 2013;502:698–702.
 64. Vila-Costa M, Simó R, Harada H, Gasol JM, Slezak D, Kiene RP. Dimethylsulfoniopropionate uptake by marine phytoplankton. *Science.* 2006;314:652–4.
 65. Kiene RP, Linn LJ, González JM, Moran MA. Dimethylsulfoniopropionate and methanethiol are important precursors of methionine and protein-sulfur in marine bacterioplankton. *Appl Environ Microbiol.* 1999;65:4549–58.
 66. González JM, Covert JS, Whitman WB, Henriksen JR, Mayer F, Scharf B, et al. *Silicibacter pomeroyi* sp. nov. and *Roseovarius nubinhibens* sp. nov., dimethylsulfoniopropionate-demethylating bacteria from marine environments. *Int J Syst Evol Microbiol.* 2003;53:1261–9.