

Evolutionary history of dimethylsulfoniopropionate (DMSP) demethylation enzyme DmdA in marine bacteria (#44271)

1

First submission

Guidance from your Editor

Please submit by **21 Apr 2020** for the benefit of the authors (and your \$200 publishing discount) .



Structure and Criteria

Please read the 'Structure and Criteria' page for general guidance.



Raw data check

Review the raw data.



Image check

Check that figures and images have not been inappropriately manipulated.

Privacy reminder: If uploading an annotated PDF, remove identifiable information to remain anonymous.

Files

Download and review all files from the [materials page](#).

33 Figure file(s)

11 Table file(s)

1 Other file(s)



Structure and Criteria

Structure your review

The review form is divided into 5 sections. Please consider these when composing your review:

1. BASIC REPORTING
2. EXPERIMENTAL DESIGN
3. VALIDITY OF THE FINDINGS
4. General comments
5. Confidential notes to the editor

 You can also annotate this PDF and upload it as part of your review

When ready [submit online](#).

Editorial Criteria

Use these criteria points to structure your review. The full detailed editorial criteria is on your [guidance page](#).

BASIC REPORTING

-  Clear, unambiguous, professional English language used throughout.
-  Intro & background to show context. Literature well referenced & relevant.
-  Structure conforms to [Peerj standards](#), discipline norm, or improved for clarity.
-  Figures are relevant, high quality, well labelled & described.
-  Raw data supplied (see [Peerj policy](#)).

EXPERIMENTAL DESIGN

-  Original primary research within [Scope of the journal](#).
-  Research question well defined, relevant & meaningful. It is stated how the research fills an identified knowledge gap.
-  Rigorous investigation performed to a high technical & ethical standard.
-  Methods described with sufficient detail & information to replicate.

VALIDITY OF THE FINDINGS

-  Impact and novelty not assessed. Negative/inconclusive results accepted. *Meaningful* replication encouraged where rationale & benefit to literature is clearly stated.
-  All underlying data have been provided; they are robust, statistically sound, & controlled.
-  Speculation is welcome, but should be identified as such.
-  Conclusions are well stated, linked to original research question & limited to supporting results.

Standout reviewing tips

3



The best reviewers use these techniques

Tip

Support criticisms with evidence from the text or from other sources

Example

Smith et al (J of Methodology, 2005, V3, pp 123) have shown that the analysis you use in Lines 241-250 is not the most appropriate for this situation. Please explain why you used this method.

Give specific suggestions on how to improve the manuscript

Your introduction needs more detail. I suggest that you improve the description at lines 57- 86 to provide more justification for your study (specifically, you should expand upon the knowledge gap being filled).

Comment on language and grammar issues

The English language should be improved to ensure that an international audience can clearly understand your text. Some examples where the language could be improved include lines 23, 77, 121, 128 – the current phrasing makes comprehension difficult.

Organize by importance of the issues, and number your points

1. Your most important issue
2. The next most important item
3. ...
4. The least important points

Please provide constructive criticism, and avoid personal opinions

I thank you for providing the raw data, however your supplemental files need more descriptive metadata identifiers to be useful to future readers. Although your results are compelling, the data analysis should be improved in the following ways: AA, BB, CC

Comment on strengths (as well as weaknesses) of the manuscript

I commend the authors for their extensive data set, compiled over many years of detailed fieldwork. In addition, the manuscript is clearly written in professional, unambiguous language. If there is a weakness, it is in the statistical analysis (as I have noted above) which should be improved upon before Acceptance.

Evolutionary history of dimethylsulfoniopropionate (DMSP) demethylation enzyme DmdA in marine bacteria

Laura Hernández^{Corresp., 1}, Alberto Vicens², Luis E Eguiarte^{Corresp., 3}, Valeria Souza³, Valerie De Anda⁴, José M González¹

¹ Facultad de Ciencias, Departamento de Microbiología, Universidad de La Laguna, La Laguna, Spain

² Departamento de Bioquímica, Genética e Inmunología, Universidad de Vigo, Vigo, Spain

³ Instituto de Ecología, Departamento de Ecología Evolutiva, Universidad Nacional Autónoma de México, Mexico D.F., Mexico

⁴ Marine Science Institution, Department of Marine Sciences, University of Texas at Austin, Texas, United States

Corresponding Authors: Laura Hernández, Luis E Eguiarte
Email address: dellia222@yahoo.es, luiseguiarte@gmail.com

Dimethylsulfoniopropionate (DMSP), an osmolyte produced by oceanic phytoplankton, is predominantly degraded by bacteria belonging to the *Roseobacter* lineage and other marine *Alphaproteobacteria* via DMSP-dependent demethylase A protein (DmdA). To date, the evolutionary history of DmdA gene family is unclear. Some studies indicate a common ancestry between DmdA and GcvT gene families and a co-evolution between *Roseobacter* and the DMSP-producing-phytoplankton around 250 million years ago (Mya). In this work, we analyzed the evolution of DmdA under three possible evolutionary scenarios: 1) a recent common ancestor of DmdA and GcvT, 2) a coevolution between *Roseobacter* and the DMSP-producing-phytoplankton, and 3) pre-adapted enzymes to DMSP prior to *Roseobacter* origin. Our analyses indicate that DmdA is a new gene family originated from GcvT genes by duplication and functional divergence driven by positive selection before a coevolution between *Roseobacter* and phytoplankton. Our data suggest that *Roseobacter* acquired *dmdA* by horizontal gene transfer prior to exposition to an environment with higher DMSP. Here, we propose that the ancestor that carried the DMSP demethylation pathway genes evolved in the Archean, and was exposed to a higher concentration of DMSP in a sulfur-rich atmosphere and anoxic ocean, compared to recent *Roseobacter* ecoparalogs (copies performing the same function under different conditions), which should be adapted to lower concentrations of DMSP.

Evolutionary history of dimethylsulfoniopropionate (DMSP) demethylation enzyme DmdA in marine bacteria

Laura Hernández^{1*}, Alberto Vicens², Luis E Eguiarte^{3*}, Valeria Souza³,
Valerie De Anda⁴, José M. González¹

¹Departamento de Microbiología, Universidad de La Laguna, La Laguna, Spain

²Departamento de Bioquímica, Genética e Inmunología, Universidad de Vigo, Vigo, Spain

³Departamento de Ecología Evolutiva, Instituto de Ecología, Universidad Nacional Autónoma de México, D.F., Mexico

⁴University of Texas Austin, Department of Marine Sciences, Marine Science Institute, Port Aransas

Corresponding Author:

Laura Hernández

Departamento de Microbiología, Universidad de La Laguna, La Laguna, Spain

lhernanj@ull.edu.es

Luis E Eguiarte

Departamento de Ecología Evolutiva, Instituto de Ecología, Universidad Nacional Autónoma de México, D.F., Mexico

fruns@unam.mx

Abstract

Dimethylsulfoniopropionate (DMSP), an osmolyte produced by oceanic phytoplankton, is predominantly degraded by bacteria belonging to the *Roseobacter* lineage and other marine *Alphaproteobacteria* via DMSP-dependent demethylase A protein (DmdA). To date, the evolutionary history of DmdA gene family is unclear. Some studies indicate a common ancestry between DmdA and GcvT gene families and a co-evolution between *Roseobacter* and the DMSP-producing-phytoplankton around 250 million years ago (Mya). In this work, we analyzed the evolution of DmdA under three possible evolutionary scenarios: 1) a recent common ancestor of DmdA and GcvT, 2) a coevolution between *Roseobacter* and the DMSP-producing-phytoplankton, and 3) pre-adapted enzymes to DMSP prior to *Roseobacter* origin. Our analyses indicate that DmdA is a new gene family originated from GcvT genes by duplication and functional divergence driven by positive selection before a coevolution between *Roseobacter* and phytoplankton. Our data suggest that *Roseobacter* acquired *dmdA* by horizontal gene transfer prior to exposition to an environment with higher DMSP. Here, we propose that the ancestor that carried the DMSP demethylation pathway genes evolved in the Archean, and was exposed to a higher concentration of DMSP in a sulfur rich atmosphere and anoxic ocean, compared to recent *Roseobacter* ecoparalogs (copies performing the same function under different conditions), which should be adapted to lower concentrations of DMSP.

Introduction

Dimethylsulfoniopropionate (DMSP) is an osmolyte synthesized by oceanic phytoplankton (Galinski, 1995; Yoch, 2002). This molecule became abundant in the oceans 250 million years ago (Mya), coinciding with the expansion and diversification of dinoflagellates (Bullock et al., 2017). Since then, it has played an important role in the biogeochemistry of sulfur cycle on Earth (Lovelock, 1983). DMSP is the main precursor of the climate-relevant gas dimethylsulfide (DMS; Reisch et al., 2011). In marine ecosystems, DMSP is rapidly degraded by different bacterial communities (González et al., 1999), and some strains seem to be very efficient and even become dependent on its degradation (Tripp et al., 2008). In fact, DMSP supports up to 13% of the bacterial carbon demand in surface waters, making it one of the most significant substrates for bacterioplankton (Kiene et al., 1999; González et al., 1999). *Candidatus Pelagibacter ubique* (SAR11), dominant in the bacterioplankton and especially in surface waters, can only use sulfur atoms derived organic molecules, such as DMSP (Tripp et al., 2008). In the case of *Ruegeria pomeroyi* DSS-3, a model organism for DMSP studies, the turnover rate of DMSP transformation depends on salinity conditions (Salgado et al., 2014).

The first step in the degradation of DMSP involves two competing pathways, cleavage and demethylation. The DMSP cleavage pathway metabolizes DMSP with the release of DMS (Kiene et al., 1999), a step catalyzed by a number of enzymes (Curson et al., 2011). In the alternative pathway, DMSP is first demethylated by a DMSP-dependent demethylase A protein (DmdA; Howard et al., 2006). Compared to the DMS-releasing pathway, *dmdA* is the most frequent gene in the genomes of oceanic bacteria (Newton et al., 2010). The DmdA enzyme was originally annotated as a glycine cleavage T-protein (GcvT) in the model bacteria *R. pomeroyi* (Reisch et al., 2011a), although it forms a separate clade from the known GcvTs (*gcvT*, *gcvH*, *gcvP* and *gcvT-C*) (Bullock et al., 2017). Despite their structural similarity which might indicate a common ancestry, DmdA and GcvT are mechanistically distinct (Schuller et al., 2012). DmdA produces 5-methyl-THF from DMSP as the result of a redox-neutral methyl transfer while GcvT converts glycine to 5,10-methylene-THF (Reisch et al., 2008).

Nearly all known DMSP-catabolizing bacteria belong to the phylum *Proteobacteria* with DmdA orthologs found in most of the sequenced members of the *Rhodobacteraceae* family, as well as strains of SAR11, SAR324, SAR116 and in marine *Gammaproteobacteria* (González et al., 1999; González, 2003; Howard et al., 2006; Bürgmann et al., 2007; Reisch et al., 2008; González et al., 2019). This phylogenetic distribution suggests an expansion of *dmdA* through horizontal gene transfer events (HGT) between different lineages of bacteria, presumably through viruses (Raina et al., 2010). Since the genome expansion of *Roseobacter* coincides with the diversification of the dinoflagellates and coccolithophores around 250 Mya (Luo et al., 2013; Luo & Moran, 2014; Bullock et al., 2017) it has been suggested a co-evolutionary event between *Roseobacter* and the DMSP-producing-phytoplankton (González et al., 1999; Zubkov et al., 2001; Moran et al., 2007; Bullock et al., 2017). Under this scenario, the enzymes of the DMSP demethylation pathway could have evolved within the last 250 Mya, as phytoplankton responded to the marine catastrophe at the end of the Permian with the diversification of dinoflagellates that produce DMSP and *Roseobacter* clade expanding by using DMSP as its main sulfur source. Despite this hypothesis, there is a lack of knowledge about the main evolutionary events that lead the DMSP adaptation in *Roseobacter*.

In terms of production, the biosynthesis of DMSP has been reported in marine heterotrophic bacteria, such as the *Alphaproteobacteria*, i.e. *Labrenzia aggregata* (Curson et al., 2017). Since a common ancestor within the *Roseobacter* originated in the Archean, more than 2 billion years

ago (Kumar et al., 2017), the *Roseobacter* and other *Alphaproteobacteria* might have been exposed to this DMSP early (Reisch et al. 2011a,b). According to this hypothesis, the DMSP demethylation and the cleavage pathways arose by the evolution of enzymes that were already present in bacterial genomes and adapted in response to the wide availability of DMSP. As mentioned earlier, *Alphaproteobacteria* in the SAR11 seems to thrive at the expense of organic sulfur compounds, such as DMSP and has a common ancestor that lived ca. 826 Mya, at the end of the Precambrian (Luo et al., 2013). We would then expect a common ancestor of the DmdA gene family during the early Proterozoic Mya and that the functional divergence between DmdA and GcvT gene families was driven by both functional constraints and widespread HGT. Probably in the Huronian snowball earth, a period of planetary crisis where the greatest microbial diversity took refuge in the shallow seas close to the equator (Tang, Thomas, & Xia, n.d.).

Here, we analyzed the evolutionary history of the DmdA gene family in marine *Proteobacteria* by considering three evolutionary scenarios: 1) a recent common ancestry of DmdA and GcvT, 2) a coevolution between *Roseobacter* and the DMSP-producing-phytoplankton, and 3) pre-adapted enzymes to DMSP prior to *Roseobacter* origin. We first analyzed if convergent, independent or HGT-based evolution can explain the presence of *dmdA* genes in different bacterial lineages of SAR11, SAR116 and *Rhodobacteraceae*. Then, we inferred the most recent common ancestor (MRCA) of the DmdA gene family, the timing of its origin and any duplication events. We also reconstructed the ancestral forms of DmdA enzymes to infer the most likely ecological conditions where DmdA thrive. We provide insights into their function by analyzing DmdA structural evolution. Finally, we examined how natural selection could have driven the divergence of the DmdA gene family. Our results indicate that *dmdA* appeared before the origin of *Roseobacter* clade and the conditions of the late Permian created by eukaryotic phytoplankton. Therefore, DmdA is an adapted version of enzyme that evolved in response to the availability of DMSP.

Materials & Methods

Data mining

DmdA orthologs and *dmdA* homologs were collected from a set of 771 genomes manually curated and hosted in the MarRef database (Klemetsen et al., 2018). The sequences were obtained as described by González et al. (2019). The DmdA homologs included were obtained using a HMM designed for DmdA orthologs (González et al., 2019), with a relaxed maximum e-value (e-50). A total of 204 sequences from 184 genomes were used to infer the evolutionary history of DmdA gene family (Supplementary Table 1).

Phylogenetic tree reconstruction and topology tests

The phylogenetic tree of the DmdA protein sequences included DmdA orthologs and DmdA homologs (called non-DmdA). The sequences were aligned using MUSCLE (Edgar, 2004). Regions poorly aligned or with gaps were removed using TrimAl (Capella-Gutiérrez et al., 2009) with parameters set to a minimum overlap of 0.55 and a percent of good positions to 60. Best-fit evolutionary model was selected based on the results of the package ProtTest 3 (Darriba et al., 2011) to determine the best-fit model for maximum likelihood (ML) and Bayesian inference (BI).

For the maximum likelihood analysis, PhyML v3.0 (Guindon et al., 2010) or RaxML v7.2.6 (Stamatakis, 2006) were used to generate 100 ML bootstrap trees, using the Le Gascuel (LG) model with a discrete gamma distribution (+G) with four rate categories, as this was the model with the lowest Akaike information criterion and Bayesian information criterion score. For the

Bayesian analysis, trees were constructed using the PhyloBayes program (Lartillot & Philippe, 2004, 2006; Lartillot et al., 2007) with the CAT model that integrates heterogeneity of amino acid composition across sites of a protein alignment. In this case, two chains were run in parallel and checked for convergence using the tracecomp and bpcomp scripts provided in PhyloBayes. As an alternative, we computed a phylogenetic tree using a Bayesian inference implemented in BEAST2 program which was run with relaxed clock model and Birth Death tree prior (Bouckaert et al., 2014). Finally, we used R v3.6.1 (R Core Team, 2017) with phangorn v2.5.5 (Schliep, 2011) to perform consensus unrooted tree.

We ran several topology tests to establish whether the trees generated using the ML and BI methods provided an equivalent explanation for the two main groups, i.e., the non-DmdA and DmdA clades. For this analysis, the topologies were compared with the TOPD/FMTS software v4.6 (Puigbo et al., 2007). A random average split distance of 100 trees was also created to check if the differences observed were more likely to have been generated by chance.

Horizontal gene transfer (HGT) test and GC content analysis

Two approaches were used to detect HGT. First, a phylogenetic incongruence analysis (Ravenhall, Škunca, Lassalle, & Dessimoz, 2015) through three topology tests, the Kishino-Hasegawa (KH) (Kishino & Hasegawa, 1989), the Shimodaira-Hasegawa (SH) (Shimodaira & Hasegawa, 1999) and the approximately unbiased (AU) (Shimodaira, 2002), implemented in the IQ-TREE software v1.5.5 (Nguyen et al., 2015). Two topologies were tested, the ML topology obtained for the species tree of the genomes here analyzed, and the ML phylogeny of DmdA. To construct the species tree, ribosomal protein 16 small subunit (RPS16) sequences were collected from the MarRef database (Klemetsen et al., 2018), one for each genome (Supplementary Table 1).

The GC content variation was studied to identify genes that have a different percentage of GC content at the third position of codons with respect to the neighboring genomic regions. The EPIC-CoGe browser (Nelson et al., 2018) was used to visualize the genomes and sequences and look for genes that use different codons with respect to the rest of the genomic dataset (data are available under permission as “ULL-microevolution” on <https://genomevolution.org/>).

Molecular dating

We first tested for heterogeneities in the substitution rates of the genes using a likelihood ratio test (LRT) (Felsenstein, 1981) with the ML-inferred tree. Likelihoods' values were estimated using baseml in PAML v4.8 (Yang, 2007) under rate constant and rate variable models and used to compute the likelihood ratio test (LRT) statistic according to the following equation:

$$LRT = -2(\log L_1 - \log L_0)$$

where L_1 is the unconstrained (nonclock) likelihood value, and L_0 is the likelihood value obtained under the rate constancy assumption. LRT is distributed approximately as a chi-square random variable with $(m-2)$ degrees of freedom (df), m being the number of branches/parameters.

To conduct a molecular dating analysis with BEAST 2 (Bouckaert et al., 2014), two independent MCMC tree searches were run for 50 million generations, with a sampling frequency of 1000 generations over codon alignment obtained, as we explain in the next section. The GTR substitution model with a gamma shape parameter and a proportion of invariants (GTR + G + I), was selected with PartitionFinder software v2.1.1 (Lanfear et al., 2016) based on the Bayesian

Information Criterion (Darriba et al., 2012), applied with a Birth Death tree prior (Gernhard, 2008) and an uncorrelated relaxed clock log-normal. The molecular clock was calibrated using information from the TimeTree database (Hedges et al., 2006, 2015; Kumar et al., 2017). We used the **dates** of the most recent common ancestor of (1) the *Alpha-* and *Gammaproteobacteria* (2480 Mya), (2) the *Halobacteriales* (455 Mya) (Supplementary Fig 1-3), and (3) the SAR11 (826 Mya) (Luo et al., 2013). A log-normal prior distribution on the calibrated nodes centered at the values mentioned above was specified with 20 standard deviations and constrained to be monophyletic. Convergence of the stationary distribution was checked by visual inspection of plotted posterior estimates in Tracer v1.6 (Rambaut, & Drummond, 2013) to ensure effective sample sizes (ESSs) of parameters were $\gg 200$, as recommended by the authors. After discarding the first 15% trees as burn-in, the samples were summarized in the maximum clade credibility tree using TreeAnnotator v1.6.1 (Rambaut, & Drummond, 2002) with a PP limit of 0.5 and summarizing mean node heights. Means and 95 % higher posterior densities (HPDs) of age estimates are obtained from the combined outputs using Tracer v1.6. The results were visualized using FigTree v.1.4.3 (Rambaut, 2009).

Maximum likelihood tests of positive selection

To measure the strength and mode of natural selection during the evolution of **DmdA gene** family, the ratio of non-synonymous (dN) to synonymous substitutions (dS) ($\omega = dN/dS$) was calculated in CodeML implemented in the suite Phylogenetic Analysis by Maximum Likelihood (PAML package v4.8) (Yang, 2007).

CodeML requires an alignment of coding sequences, and a phylogenetic tree. DNA alignment was achieved by MUSCLE (Edgar, 2004) implemented in MEGA-CC v7.0.26 (Kumar et al., 2016) and poorly aligned segments were eliminated with Gblocks under defaults parameters (Castresana, 2000). The phylogenetic tree was built using ML with PhyML v3.0 (Guindon et al., 2010) as described above and a nucleotide substitution model selected by jModelTest (Darriba et al., 2012). DAMBE (Xia, 2001) was also used to check for saturation of nucleotide substitutions using a plot of the number of transitions and transversions for each pairwise comparison against the genetic distance calculated with the F84 model of nucleotide substitution (Huelsenbeck & Rannala, 1997), which allows different equilibrium nucleotide frequencies and a transition rate-transversion rate bias. Multiple sequence alignments with similar characteristics (i.e., showing saturation of nucleotide substitutions) were then analyzed with CodeML (Yang, 2007).

Three sets of models were used (site-specific, branch-specific and branch-site models) to detect pervasive and episodic selection during the evolution of *dmdA* orthologs. Likelihood-ratio tests (LRTs) were used to compare models, and significant results ($p\text{-value} < 0.05$) were determined contrasting with a chi-square distribution (chisq) (Anisimova et al., 2001).

In the site-specific analysis, we tested for variability of selection (type and magnitude) across the codons of the gene using three pairs of nested models. The first pair includes M0 (just one dN/dS ratio) and M3 (“K” discrete categories of dN/dS) and has four degrees of freedom (df). The second pair of models considers M1a (just two classes of sites, purifying [$dN/dS < 1$] and neutral selection [$dN/dS = 1$]) and M2a (the same as M1a adding a third class of sites dedicated to positive selection [$dN/dS > 1$]), this has two df. Finally, the third pair of models comprised M7 (a beta distribution that allows dN/dS to vary among the interval [0,1]) and M8 (adds an extra discrete category to M7 with $dN/dS > 1$), with two df. Whereas M0 vs M3 test for evidence of dN/dS

variation across sites, M1a vs M2a and M7 vs M8 test for the presence of sites under positive selection ($dN/dS > 1$).

Using three branch models (Yang, 1998), we tested for variation of selection over evolutionary time. The null model (M0) assumes that all branches evolve at the same rate, therefore, there is only one value of dN/dS for all the branches of the tree. The two-ratio model allows two dN/dS values, one value for all *Roseobacter* lineages (we called this group A) and another for the rest of branches (named group B). The free-ratio model, allows one dN/dS value for each branch. Null and two-ratio model are compared by LRT with one df but null and free-ratio model are compared with 36 df.

For the last set of models, we identified sites that have been under positive selection at a particular point of evolution using branch-site models, in which dN/dS can vary among sites and among branches (Zhang, 2005). We computed two models: a null model, in which the “foreground branch” may have different proportions of sites under neutral selection to the “background branches”, and an alternative model in which the “foreground branch” may have a proportion of sites under positive selection. We compare these models for each terminal branch with a LRT of one df. For each branch-site analysis, we applied the Bonferroni correction for multiple testing.

In site and branch-site tests, we identified sites under positive selection as those with Bayes Empirical Bayes (BEB) posterior probability above the 0.95 (Yang, 2005). We also checked for convergence of the parameter estimates in PAML by carrying out at least two runs for each tree and starting the analysis with different ω (0.2, 1, 1.2 and 2). In addition, to test for convergent selection in several lineages, we ran at Branch-site analysis selecting as “foreground branches” all those under positive selection in a previous analysis.

Analysis of functional divergence

Divergent selection is indicated by different ω 's values among paralogous clades. We tested whether selective pressures diverged following duplication that led to *dmdA* and non-*dmdA* genes (Bielawski & Yang, 2004). We compared the M3 model, which accounts for ω variation among sites but not among branches or clades, with a model allowing a fraction of sites to have different ω between two clades of a phylogeny (clade model D). We also tested M0 and M3 models and we used a posterior BEB probability above the 0.95 to identify sites evolving under divergent selective pressures. We checked for convergence of the parameter estimates in PAML by carrying out at least two runs for the tree and starting the analysis with different ω (0.1, 0.25, 2, 3 and 4).

Finally, we applied two branch-site models (as described above) to test dN/dS differences on the branches representing the ancestral lineages of the *DmdA* and non-*DmdA* clades (see results) (Supplementary Fig 25). We considered the ancestral sequences from *DmdA* and non-*DmdA* clades as foreground branches in two different models.

Reconstruction of ancestral *DmdA* sequence

To reconstruct the ancient conditions where *dmdA* gene prospered, we inferred the ancestral sequences of the *DmdA* node using the FastML web server (Ashkenazy et al., 2012) and then computed estimated physico-chemical properties on predecessor sequence using Compute

ProtParam tool from ExPASy – SIB Bioinformatics Resource Portal (Gasteiger et al., 2005). Moreover, we also reconstructed the ancestral sequence of the non-DmdA node, as well as the ancestral sequence of both the DmdA, and the non-DmdA families. FastML was run considering the alignment of proteins and the ML phylogenetic tree for those DmdA orthologs or homologs inferred as we explained above. Posterior amino acid probabilities at each site were calculated using the Le Gascuel (LG) matrix (Le & Gascuel, 2008) and Gamma distribution. Both marginal and joint probability reconstructions were performed. Protein sequences resulting from marginal reconstructions were used to predict tertiary structure (see below) as well as to identify family domains using Pfam v32 (Finn et al., 2010).

Protein tertiary structure analysis

Predicted three-dimensional structures of protein sequences were examined by Iterative Threading ASSEMBLY Refinement (I-TASSER) (Roy et al., 2010; Yang et al., 2015). First, I-TASSER uses local meta-threading-server (LOMETS) (Wu & Zhang, 2007) to identify templates for the query sequence in a non-redundant Protein Data Bank (PDB) 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, top 10 structural analogs of the predicted model close to the target in the PDB (Berman et al., 2000) are generated using TM-align (Zhang, 2005). The TM-score value scales the structural similarity between two proteins, and should return 1 if a perfect match between two structures is found. A TM-score value higher than 0.5 suggests that the proteins belong to the same fold family.

We used PyMol v1.7.4 (DeLano, 2002) to visualize the 3D structure of the proteins and to map the positively selected sites onto the 3D structure of DmdA (pdb: 3tfh).

Results

Phylogenetic tree for DmdA family

We identify a total of 204 DmdA protein sequences out of 150 curated genomes, and reconstruct their evolutionary relationships by Bayesian Inference (BI) (Fig 1) and Maximum Likelihood (ML) (Supplementary Fig 4). Unrooted trees in TOPD-FMTS indicated that split distances did not exceed 0.19, indicating that the phylogenetic reconstruction is robust, with minor variations in alignment filtering and methods for inferring topologies (Supplementary Table 2).

The BI tree (Fig 1) shows a main duplication between two lineages. The larger phylogenetic group comprises genes from *Bacteroidetes*, while the smaller group includes genes from *Alphaproteobacteria*. We focused on this smaller group as it includes the DmdA sequences (Fig 1; green color) and the closest homologs to DmdA (Fig 1; yellow color).

Using phylogenetic analyses including DmdA orthologs and DmdA homologs close to those (the limit to select closer homologs was set to a maximum e-value of e-80) we resolve the position of the first DmdA sequences isolated from two marine bacterial species, *R. pomeroyi* (AAV95190.1) and *Ca. P. ubique* (AAZ21068.1). In addition, the inclusion of DmdA homologs allowed to resolve a robust phylogenetic relationship of DmdA gene family (Fig 2). We detected a clear separation between DmdA and putative non-DmdA families. Indeed, the four DmdA family trees constructed using different methods compared in TOPD-FMTS using split distances

(Supplementary Table 3) and unrooted trees (Supplementary Fig 5) agreed with this result. The average split distance was 0.60, indicating that the trees were neither identical (split difference=0) nor completely different (1). A random split distance was calculated to analyze whether the split distances were significantly different. Because the random split distance resulted in a value close to 1 (0.988), our observations are unlikely to be given by chance.

To identify HGT and duplication events, we constructed a proxy for the species tree of the genomes considered here by using a set of small subunit ribosomal protein (see Material and Methods). Given this (proxy) species tree (Supplementary Fig 6), the positions of many sequences on the DmdA tree are better explained as cases of HGT (Supplementary Fig 6; Fig 3) with high statistical support. We then tested whether the topology for a common set of taxa within the DmdA family (Supplementary Fig 7) were similar to that of the species tree (Supplementary Fig 8). We found significant differences (at an alpha of 0.01) between the topology of DmdA group and that of the proxy species tree (Table 1); this incongruence between phylogenies is true irrespective of the test used (Kishino-Hasegawa, Shimodaira-Hasegawa and unbiased tests). From these results we conclude that the phylogenetic relationships within each DmdA group are different to those of the species tree, strongly supporting a HGT-based evolution of DmdA family (Supplementary Fig 9).

Moreover, we found many genes that use different codons than the neighboring genomic regions. These genes are inferred as having been horizontally transferred given their (G+C) wobble content (Supplementary Table 1), supporting an HGT-based evolution of DmdA family (Supplementary Fig 9).

Structural modeling

The structure for DmdA orthologs inferred on the protein sequences by Iterative Threading ASSEmbly Refinement (I-TASSER) were threaded onto the known structure of DMSP-dependent demethylase A protein (PDB accession: 3tfhA) with a C-score ≤ 2 (Table 2). However, the predicted models for DmdA homologs were threaded onto two types of known structure; DmdA orthologs, and the structure of the mature form of rat dimethylglycine dehydrogenase (DmgdH) (PDB accession, 4ps9sA) with a C-score < 2 except for the sequence with accession number AEM59334.1, which shows a C-score > 2 (Supplementary Fig 10a, Supplementary Data 1).

We clustered sequences with a putative DmgdH structure in a separate group using principal component analysis (Supplementary Fig 11). There is a clear 3D-structure coincidence between DmdA clade (red color in Supplementary Fig 10a) and the majority of lineages from non-DmdA clade (orange color in Supplementary Fig 10a) as well as a conserved folate-binding domain (Supplementary Fig 10b: 99S, 178E and 180Y). However, in the alignment we found a pattern of conserved residues coherent with phylogeny results (Supplementary Fig 10a, Supplementary Fig 10b), where non-DmdA clade is formed by three subclades, one of them with DmgdH tertiary structure. Indeed, key residue for DMSP specific interaction is shown in clades with DmdA tertiary structure (Supplementary Fig 10b: W171) but not in a clade with DmgdH tertiary structure (Supplementary Fig 10b: F171).

Molecular dating

The log likelihood test (LRT) detected heterogeneity in the substitution rates of *dmdA* orthologs and *dmdA* homologs genes (Fig 2) ($\log L_0 = -29,827.108$; $\log L_1 = -29,546.053$; degrees of freedom = 46; $\chi^2 = 562.11$; $P < 0.001$), thus rejecting the hypothesis of a strict molecular clock. This finding validates the use of relaxed molecular clock approach to estimate the node ages throughout Bayesian analysis (see Methods for details). We observed that the marginal densities for each run of the divergence time estimate analysis were nearly identical, pointing that the runs converged on the same stationary distributions. In all runs, the marginal densities for the standard deviation hyperparameter of the uncorrelated log-normal relaxed clock model were quite different from the prior, with no significant density at zero and with a coefficient of variation around 0.2. Analyses using three different calibrated prior dates showed not discrepancies in the final divergence time estimates (Table 3).

The time estimates for the MRCA of each gene family (Table 3 and Fig 4) indicate that the most recent common ancestor of DmdA gene family occurred in the late Archean, around 2,400 Mya, after a gene duplication event. Also, a duplication within the DmdA lineage generated a separated SAR11 and *Roseobacter* DmdA lineage in the early Precambrian ca. 1,894 Mya (Fig 4: red arrow). *Ca. P. ubique* HTCC1062 within the first cluster and *R. pomeroyi* DSS-3 within the second cluster, resulted from a duplication around 300 Mya (Fig 4: blue arrow). However, a higher number of duplication events took place in the second cluster. Thus the number of paralogous genes comprising the *Roseobacter* DmdA family is larger than in SAR11 (Fig 4).

We detected two duplication events within the putative non-DmdA clade (Fig 4; orange color); showing that the gene families ~~were originated~~ through old duplication events. One duplication involving the DmgdH family (Fig 4 dark yellow color; Table 2) occurred 1,480 Mya and another duplication 1,000 Mya (Fig 4: green arrow), with tertiary structure similar to the DmdA from *Ca. P. ubique*. The other event of duplication took place during the Huronian glaciation, around 2100 Mya (Fig 4: violet arrow).

Reconstruction of ancestral DmdA sequence

Our analysis was focused on the reconstruction of the ancestral sequences of the DmdA clade, the non-DmdA clade as well as the ancestral sequence of both the DmdA and non-DmdA clades. FastML inferred the 100 most likely ancestral sequences of the DmdA family. We observed that the same sequences were always inferred. Indeed, the difference in log-likelihood between the most likely ancestral sequence at this node (N1; Supplementary Fig 12) and the 100th most likely sequence was only 0.105, indicating that both sequences are almost as likely to reflect the “true” ancestral sequence. That ancestral protein contains both PF01571 (GCV_T) and PF08669 (GCV_T_C) domains, found in the DmdA orthologs and it is nearly identical to *Ca. P. ubique* HTCC1062 DmdA sequence. Moreover, PSI-BLAST search confirmed that the ancestral sequence in node 1 close to DmdA genes hosted in EMBL-EBI databases (Supplementary Fig 13) and the structure for *Ca. P. ubique* apoenzyme DmdA was the closest analog to our predicted models (Table 2; Supplementary Data 1). Inferred physico-chemical properties are identical between *Ca. P. ubique* and the DmdA ancestral sequence (Supplementary Table 4).

On the other hand, the ancestral sequence inferred for non-DmdA family (N1; Supplementary Fig 14) and the ancestral sequence previous to functional divergence (N1; Supplementary Fig 15) contains only the PF01571 domain. That domain was located onto the known structure of T-protein of the Glycine Cleavage System (PDB accession: 1wooA) with a C-score= 1.25 (Table 2; Supplementary Data 1) in the case of the ancestral DmdA and non-DmdA sequence. However,

the ancestral sequence for non-DmdA was better threaded onto the known structure of mature form of rat DmgdH (PDB accession: 4p9sA) with a C-score= 0.76 (Table 2; Supplementary Data 1).

Detection of positive selection on *dmdA* sequences

To infer how natural selection has influenced the evolutionary history of *DmdA* gene family, we used an alignment of the 20 sequences clustered as *dmdA* orthologs. The phylogenetic tree for these sequences was constructed by ML using the symmetrical model (SYM) with a discrete gamma distribution.

The average dN/dS value for the *dmdA* gene was 0.085, suggesting that this gene evolved under strong negative (purifying) selection. Then, we analyzed dN/dS variation across the codons in the gene, comparing M0 and M3 models through a LRT. The M3 model had better fit to the data than the M0 model (chisq= 775.387, p-value< 0.01). All codons in the gene are under strong purifying selection with dN/dS <1 (Fig 5), suggesting the importance of this sulfur pathway for the cells. In accordance with this, the LTRs designed to detect codons under positive selection were not significant (M1 vs M2, chisq= 0 and p-value = 1, and M7 vs M8, chisq = 1.459 and p-value = 0.482). Hence, we did not detect sites in *dmdA* subjected to positive selection (Supplementary Fig 17).

We tested the variation in the intensity of selection over evolutionary time. A two-ratio model comparing the *Roseobacter* with the rest of lineages (Supplementary Fig 18) fits better the data, as the LRT was 23.777 and p-value < 0.01 (Table 4). dN/dS value in *Roseobacter* (ω_1 : 0.0767) was significantly lower than in the remaining branches (ω_2 : 0.1494), suggesting stronger purifying selection on *dmdA* in *Roseobacter*. When we tested the intensity of selection over evolutionary time using the free-ratio model (Table 4), we found changes in the selection pressure from the branches which defines the separation of SAR11 and *Roseobacter* *DmdA* gene families (Supplementary Fig 19: branches from nodes 21 to 23). In particular, we observed a dN/dS value > 1 in the branch connecting nodes 21-23. We also identified some more recent branches (connecting nodes 25-26 and 28-29) for which dN/dS >> 1 was estimated (Supplementary Fig 19).

Finally, we applied the two branch-site models to test for sites under selection on the individual lineages associated with *dmdA* (Supplementary Fig 20). Four sequences (WP_047029467, AHM05061.1, ABV94056.1, AFS48343.1) had a significant LRT after correcting for multiple testing (Table 5), suggesting episodic positive selection on these lineages (Supplementary Fig 20). It should be highlighted that three selected sites are shared by at least two lineages (Table 5; Fig 6). One shared site is located next to the GcvT domain (152 K; Supplementary Fig 21), and two shared sites are closed to conserved positions (17E; 87Y; Supplementary Fig 21). The residue 87Y is adjacent to the conserved interaction site with THF (88Y; Supplementary Fig 21). Interestingly, since the selected lineages are separated in the tree, the adaptive mutations seem to have occurred through three parallel independent changes (Supplementary Fig 22).

Functional divergence during the molecular evolution of *DmdA* sequences

We tested whether *DmdA* and non-*DmdA* gene families were subjected to different functional constraints after gene duplication (Supplementary Fig 5). We estimated the one-ratio model (M0) that yielded a value ω = 0.053 (Table 6), indicating that purifying selection dominated the

evolution of these proteins. The discrete model (M3) was applied to these sequences (Table 6) and the LRTs comparing M0 and M3 indicated significant variation in selective pressure among sites (Table 6; Supplementary Fig 23).

The M3 model was compared with Model D, which accommodates both heterogeneity among sites and divergent selective pressures. The LRT was significant and supported the model D (Table 6), implying statistical evidence of functional divergence between DmdA and non-DmdA. Parameter estimates under Model D with $k=3$ site classes suggested that 23.6% of sites were evolving under strong purifying selection ($\omega = 0.006$), while 26.7% of sites were evolving under much weaker selective pressure ($\omega = 0.04$). Interestingly, a large set of sites (49.6%) were evolving under divergent selective pressures, with weaker purifying selection in the DmdA-clade ($\omega = 0.169$) than non-DmdA-clade ($\omega = 0.100$). We identified 77 sites evolving under divergent selective pressures between DmdA and non-DmdA (Table 6). Nineteen sites were located within the alpha helix (red tube in Supplementary Fig 24) of the secondary structure prediction and sixteen were located in the beta sheet (green arrows in Supplementary Fig 24). According to the global dN/dS estimates, for all divergent positions *dmdA* sequences seem to be more conserved than non-*dmdA* sequences. Moreover, this data is only compatible with recombination breaking linkage disequilibrium within the gene set that we observed with the HGT analysis.

Finally, we are interested in knowing if adaptive evolution has occurred in the lineages immediately following the main duplication event (Supplementary Fig 25). We applied two branch-site models to test for sites under selection on the ancestor associated with the DmdA and non-DmdA clades (Table 5). The LRT was significant for both ancestral branches (LRT > 7 and $p\text{-value} < 0.05$). Nonetheless, the foreground ω for class 2 sites tended to infinite ($\omega=999$) in both cases, indicating lack of synonymous substitutions ($dS=0$) in these sites. We also performed two-ratio models to estimate global ω on these branches, but both estimates tended to infinite (Supplementary Table 5), suggesting lack of synonymous substitution in the divergence of DmdA and non-DmdA ancestors. Therefore, although the fixation of only non-synonymous substitutions following gene duplication might indicate strong positive selection driving functional divergence of DmdA and non-DmdA families, we cannot confirm it with the applied tests.

Discussion

In this study we evaluated three scenarios for the evolutionary history of the DmdA gene family in marine bacteria. The results for each one are discussed separately.

First scenario: a recent common ancestry between DmdA and GcvT

In relation to the first scenario, we found that contrary to our initial expectations, DmdA and GcvT have not a recent common ancestry, but they share an old common ancestor. However, the clear separation between DmdA and putative non-DmdA gene families that originated in the Archean ca. 2,400 Mya after a gene duplication, supports a common recent ancestry for DmdA and non-DmdA (Fig. 7; down and up). Our tertiary structure analyses indicate that they share a putative GcvT protein (EC 2.1.2.10) as their ancestor sequence. Indeed, our results agree with other studies in the case of DmdA (Reisch et al., 2008). Then, this clade seems to have originally been a GcvT (Fig. 7) as Bullock et al. (2017) suggested.

The DmdA clade is a member of aminomethyltransferase (AMT/GCV_T) family with DMSP-dependent demethylase tertiary structure while non-DmdA clade includes an ancestor with a tertiary structure that better matches the dimethylglycine dehydrogenase oxidoreductase

(DmgdH, EC 1.5.99.2) (Fig. 7) and members with DmdA tertiary structure. To establish structural convergence as the reason of this DmdA structure coincidence between DmdA and non-DmdA members, we used a phylogenetic approach based on reconstructing ancestral sequences of the two clades, and then to model the ancestral proteins. We determined different structural features between ancestral sequence reconstructed from DmdA and non-DmdA families. In the first case, the ancestral sequence reconstructed coincides with a DmdA tertiary structure, as well as with a DmdA sequence with physico-chemical properties inferred in this study and agree with previous ones (Reisch et al., 2008). However, the non-DmdA ancestral sequence reconstructed is a DmgdH that seems to be kept in the clade called DmgdH (Fig. 7: yellow color) as well as in some members of DmdA clades (within non-DmdA clade) where the majority of sequence gained DmdA structure (Fig. 7). Therefore, DmdA structural features seem to have emerged independently in both clades: DmdA and non-DmdA. This finding is extremely interesting, since known cases of structural convergence of proteins are rare (Zakon, 2002). Experimental assays expressing and screening the activity of the ancestral proteins at different conditions will be required to corroborate the structural convergence.

Since GcvT does not share the most recent common ancestry with DmdA, we examined the functional divergence between DmdA and non-DmdA clades to explain how natural selection could have driven the divergence of the DmdA gene family. We found 77 codon sites evolving under divergent selective pressures between DmdA and non-DmdA gene families. Structural divergence seemed to be imposed on the protein during sequence divergence, since nineteen sites were located within the alpha helix of 2D structure and sixteen in the beta sheet. Nonetheless, essential regions of the enzymes as active sites seem to be under strong purifying selection, suggesting preservation of the ancestral function. The observation that DmdA sequences have less conserved divergent sites than non-DmdA sequences, suggests that non-DmdA conserves the ancestral function, whereas DmdA evolved to acquire new functions in different environments, probably as a response to the Huronia ice ball Earth (Zhang, 2003).

Second scenario: coevolution between *Roseobacter* and DMSP-producing-phytoplankton

In the second scenario, our data does not support the hypothesis of a co-evolution scenario between *Roseobacter* and DMSP-producing-phytoplankton (Luo et al., 2013). On the contrary, we found an ancestor sequence of DmdA cluster similar to DmdA from a strain of *Ca. P. ubiquus* that diverged after a more recent duplication event, before the dinoflagellate radiation in the late Permian. This finding indicates that the enzyme activity has not changed in the course of DmdA evolution. Indeed, we found that most of the codons in DmdA clade are under purifying selection probably due to the importance of this pathway for sulfur acquisition. Nonetheless, we also detected episodic positive selection in four sequences affecting a few sites, suggesting that adaptive evolution fine-tuned the function of DmdA in *Roseobacter*. Furthermore, positively selected residues were located around the GcvT domain and close to the residue involved in conserved interaction with THF, reinforcing the idea of adaptive evolution in response to the external environment.

During the study of this scenario, we suspected that *dmdA* was acquired by HGT in *Roseobacter* and SAR11. This agrees with Luo et al., (2013) and Tang et al. (2010) which found that the expansion of *dmdA* was by HGT. Moreover, our study evidence that DmdA ancestral sequence in our phylogeny comes from a marine heterotrophic bacteria adapted to presence of DMSP in the Archean, after a HGT event from this bacteria to another lineage that acquired the *dmdA* ancestral sequence. However, after the HGT events, some *dmdA* sequences have acquired similar residue

changes by independent (parallel) evolution, reinforcing the idea of functional/ecological constraints. Therefore, *Rhodobacteraceae* can live in an environment where DMSP is the main source of sulfur because they acquired the DmdA ancestor sequence by HGT, prior to have been exposed to the environment in which this protein proved useful, as Luo & Moran (2014) suggested. We did not find any signal of positive selection in *Roseobacter* group, but in contrast we found episodic evolution between SAR11 sequences. Yet, as we already mentioned DMSP is part of an ancient pathway in *Alphaproteobacteria* (Bullock et al., 2017) and it could explain the ancient origin of DmdA.

On the other hand, *Roseobacter* paralogs analyzed in this study were functionally annotated as DmdA function (González et al., 2019), as they perform the same function as the original gene (DmdA ancestor). However, we found differences in predicted isoelectric point values (pI), which were inferred in this study. Then, these paralogs could be considered as ecoparalogs as Sánchez-Pérez et al (2008) proposed for their study. Isoelectric point of a protein provides an indication of its acidic nature (Oren et al., 2005) and in this case, differences in pI suggest that the proteins differ in halophilicity. We observed proteins with the highest pI values in the DmdA ancestor sequence, as well as *Ca. P. ubique* sequence and this last one has a pI similar to the first (DmdA ancestor) (Fig. 7). Therefore, we deduced that DmdA ancestor was adapted to a higher concentration of salinity, which could have modulated the selection of the DMSP enzymatic degradation routes as in bacteria such as the model organism *R. pomeroyi* DSS-3 (Salgado et al., 2014). Interestingly, *R. pomeroyi* degrades more DMSP by the demethylation pathway under high salinity conditions, and then produces a high amount of MeSH (Howard et al., 2008; Magalhães et al., 2012; Salgado et al., 2014).

Given our data, we propose that the ancestor of the pathway that evolved in the Archean, was exposed to a higher concentration of DMSP in a sulfur-rich atmosphere and in an anoxic ocean, compared to recent ecoparalogs which should adapt to lower concentration of DMSP (Fig 7). Indeed, the ancestral ecoparalog from which recent ecoparalogs derived (*Ca. Puniceispirillum marinum* IMCC1322 or ADE38317.1 and the *Roseobacter* clade) could have undergone episodes of adaptation (the branch showed positive selection in branch-models) which would explain the change in protein stability (Pál et al., 2006). As consequence, the protein could have experimented slight reductions or loss of function.

Third scenario: pre-adapted enzymes to DMSP prior to *Roseobacter* origin

In this evolutionary scenario, *Roseobacter* clade was pre-adapted to the conditions created by eukaryotic phytoplankton at the late Permian, including dinoflagellates that released vast amounts of DMSP (Bullock et al., 2017; Luo & Moran, 2014). Our analyses indicate that the *Roseobacter* ancestor has already adapted to a high DMSP before *Roseobacter* clade arose (Luo et al., 2013). Therefore, we support Reisch et al. (2011 a,b) hypothesis where DMSP demethylation pathway enzymes are adapted versions of enzymes that were already in bacterial genomes, and evolved in response to the availability of DMSP. Since the first step in DMSP demethylation is a reaction catalyzed by DMSP demethylase encoded by *dmdA* gene (Dickschat et al., 2015), DMSP adaptation could have been evolved in this gene that originated in the Archean, a time where several lineages of bacteria produced DMSP as an osmolyte or antioxidant in the presence of the early cyanobacteria, or as a cryoprotectant in the Huronian glaciation. In bacteria, a methyltransferase gene, *dysB*, is up-regulated during increased salinity, nitrogen limitation, and at low temperatures (Curson et al., 2017), conditions already predicted to stimulate DMSP production in phytoplankton and algae (Bullock, et al., 2017; Ito, et al., 2011). Afterward, those

roles may have helped to drive the fine adaptation of existing enzymes for DMSP metabolism, and those adaptations came handy in the late Precambrian glaciations that allowed the radiation of algae and animals.

Conclusions

In conclusion, we found that *Roseobacter* adaptation to DMSP occurred via functional diversification after duplication events of the DmdA gene and adaptations to environmental variations via ecoparalogs of intermediate divergence. Our findings suggest that salinity could have been a trigger for the adaptation to DMSP metabolism.

Acknowledgements

We would like to thank to Dr. Romain Studer from BenevolentAI for his critical role in the 3D visualization of protein and mapping sites onto the 3D structure and Dr. Buckley Iglesias from Universidad Autónoma de Madrid for his introduction to molecular dating analysis with BEAST 2. This research was supported by grant CTM2016-80095-C2 from the Spanish Ministry of Economy and Competitiveness.

References

- Anisimova, M., Bielawski, J. P., & Yang, Z. (2001). Accuracy and Power of the Likelihood Ratio Test in Detecting Adaptive Molecular Evolution. *Molecular Biology and Evolution*, 18(8), 1585–1592. <https://doi.org/10.1093/oxfordjournals.molbev.a003945>
- Ashkenazy, H., Penn, O., Doron-Faigenboim, A., Cohen, O., Cannarozzi, G., Zomer, O., & Pupko, T. (2012). FastML: a web server for probabilistic reconstruction of ancestral sequences. *Nucleic Acids Research*, 40(W1), W580–W584. <https://doi.org/10.1093/nar/gks498>
- Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research*, 28(1), 235–242. <https://doi.org/10.1093/nar/28.1.235>
- Bielawski, J. P., & Yang, Z. (2004). A Maximum Likelihood Method for Detecting Functional Divergence at Individual Codon Sites, with Application to Gene Family Evolution. *Journal of Molecular Evolution*, 59(1). <https://doi.org/10.1007/s00239-004-2597-8>
- Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.-H., Xie, D., ... Drummond, A. J. (2014). BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. *PLoS Computational Biology*, 10(4), e1003537. <https://doi.org/10.1371/journal.pcbi.1003537>
- Bullock, H. A., Luo, H., & Whitman, W. B. (2017). Evolution of Dimethylsulfoniopropionate Metabolism in Marine Phytoplankton and Bacteria. *Frontiers in Microbiology*, 8. <https://doi.org/10.3389/fmicb.2017.00637>
- Bürgmann, H., Howard, E. C., Ye, W., Sun, F., Sun, S., Napierala, S., & Moran, M. A. (2007). Transcriptional response of *Silicibacter pomeroyi* DSS-3 to dimethylsulfoniopropionate (DMSP).

- 682 *Environmental Microbiology*, 9(11), 2742–2755. <https://doi.org/10.1111/j.1462->
- 683 2920.2007.01386.x
- 684
- 685 Capella-Gutiérrez, S., Silla-Martínez, J. M., & Gabaldon, T. (2009). trimAl: a tool for automated
- 686 alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, 25(15), 1972–1973.
- 687 <https://doi.org/10.1093/bioinformatics/btp348>
- 688
- 689 Castresana, J. (2000). Selection of Conserved Blocks from Multiple Alignments for Their Use in
- 690 Phylogenetic Analysis. *Molecular Biology and Evolution*, 17(4), 540–552.
- 691 <https://doi.org/10.1093/oxfordjournals.molbev.a026334>
- 692
- 693 Clamp, M., Cuff, J., Searle, S.M., Barton, G.J. (2004). The Jalview Java alignment editor.
- 694 *Bioinformatics*, 20(3), 426–427. <https://doi.org/10.1093/bioinformatics/btg430>
- 695
- 696 Curson, A. R. J., Todd, J. D., Sullivan, M. J., & Johnston, A. W. B. (2011). Catabolism of
- 697 dimethylsulphonio propionate: microorganisms, enzymes and genes. *Nature Reviews*
- 698 *Microbiology*, 9, 849. <https://doi.org/10.1038/nrmicro2653>
- 699
- 700 Curson, A. R. J., Liu, J., Martínez, A. B., Green, R. T., Chan, Y., Carrión, O., ... & Todd, J. D.
- 701 (2017). Dimethylsulfonylpropionate biosynthesis in marine bacteria and identification of the key
- 702 gene Dimethylsulphonylpropionate biosynthesis in marine bacteria and this process. *Nature*
- 703 *Microbiology*, 2(17009). <https://doi.org/10.1038/nmicrobiol2017.9>
- 704
- 705 Darriba, D., Taboada, G. L., Doallo, R., & Posada, D. (2011). ProtTest 3: fast selection of best-fit
- 706 models of protein evolution. *Bioinformatics*, 27(8), 1164–1165.
- 707 <https://doi.org/10.1093/bioinformatics/btr088>
- 708
- 709 Darriba, D., Taboada, G. L., Doallo, R., & Posada, D. (2012). jModelTest 2: more models, new
- 710 heuristics and parallel computing. *Nature Methods*, 9(8), 772–772.
- 711 <https://doi.org/10.1038/nmeth.2109>
- 712
- 713 DeLano, W. L. (2002). Pymol: An open-source molecular graphics tool. *CCP4 Newsletter On*
- 714 *Protein Crystallography*, 40(1), 82–92
- 715
- 716 Dickschat, J. S., Rabe, P., & Citron, C. A. (2015). The chemical biology of
- 717 dimethylsulfonylpropionate. *Organic & Biomolecular Chemistry*, 13(7), 1954–1968.
- 718 <https://doi.org/10.1039/C4OB02407A>
- 719
- 720 Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high
- 721 throughput. *Nucleic Acids Research*, 32(5), 1792–1797. <https://doi.org/10.1093/nar/gkh340>
- 722
- 723 Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood
- 724 approach. *Journal of Molecular Evolution*, 17(6), 368–376. <https://doi.org/10.1007/BF01734359>
- 725
- 726 Finn, R. D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J. E., ... Bateman, A. (2010).
- 727 The Pfam protein families database. *Nucleic Acids Research*, 38(suppl_1), D211–D222.
- 728 <https://doi.org/10.1093/nar/gkp985>
- 729

- Galinski, E. A. (1995). Osmoadaptation in Bacteria. In *Advances in Microbial Physiology* (Vol. 37, pp. 273–328). [https://doi.org/10.1016/S0065-2911\(08\)60148-4](https://doi.org/10.1016/S0065-2911(08)60148-4)
- Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M. R., Appel, R. D., & Bairoch, A. (2005). Protein Identification and Analysis Tools on the ExPASy Server. In J. M. Walker (Ed.), *The Proteomics Protocols Handbook* (pp. 571–607). <https://doi.org/10.1385/1-59259-890-0:571>
- Gernhard, T. (2008). The conditioned reconstructed process. *Journal of Theoretical Biology*, 253(4), 769–778. <https://doi.org/10.1016/j.jtbi.2008.04.005>
- González, J. M. G., Kiene, R. P., & Moran, M. A. (1999). Transformation of Sulfur Compounds by an Abundant Lineage of Marine Bacteria in the Subclass of the Class Proteobacteria. *APPL. Environ. Microbiol.*, 65, 10.
- González, J. M. (2003). *Silicibacter pomeroyi* sp. nov. and *Roseovarius nubinhibens* sp. nov., dimethylsulfoniopropionate-demethylating bacteria from marine environments. *International Journal of Systematic and Evolutionary Microbiology*, 53(5), 1261–1269. <https://doi.org/10.1099/ijs.0.02491-0>
- González, J. M., Hernández, L., Manzano, I., & Pedrós-Alió, C. (2019). Functional annotation of orthologs in metagenomes: a case study of genes for the transformation of oceanic dimethylsulfoniopropionate. *The ISME Journal*, 13(5), 1183–1197. <https://doi.org/10.1038/s41396-019-0347-6>
- Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., & Gascuel, O. (2010). New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Systematic Biology*, 59(3), 307–321. <https://doi.org/10.1093/sysbio/syq010>
- Hedges, S. B., Dudley, J., & Kumar, S. (2006). TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics*, 22(23), 2971–2972. <https://doi.org/10.1093/bioinformatics/btl505>
- Hedges, S. Blair, Marin, J., Suleski, M., Paymer, M., & Kumar, S. (2015). Tree of Life Reveals Clock-Like Speciation and Diversification. *Molecular Biology and Evolution*, 32(4), 835–845. <https://doi.org/10.1093/molbev/msv037>
- Howard, E. C., Henriksen, J. R., Buchan, A., Reisch, C. R., Burgmann, H., Welsh, R., ... Moran, M. A. (2006). Bacterial Taxa That Limit Sulfur Flux from the Ocean. *Science*, 314(5799), 649–652. <https://doi.org/10.1126/science.1130657>
- Howard, E. C., Sun, S., Biers, E. J., & Moran, M. A. (2008). Abundant and diverse bacteria involved in DMSP degradation in marine surface waters. *Environmental Microbiology*, 10(9), 2397–2410. <https://doi.org/10.1111/j.1462-2920.2008.01665.x>
- Huelsenbeck, J. P., & Rannala, B. (1997). Phylogenetic Methods Come of Age: Testing Hypotheses in an Evolutionary Context. *Science*, 276(5310), 227. <https://doi.org/10.1126/science.276.5310.227>

- Hug, L. A., Baker, B. J., Anantharaman, K., Brown, C. T., Probst, A. J., Castelle, C. J., ... Banfield, J. F. (2016). A new view of the tree of life. *Nature Microbiology*, 1, 16048. <https://doi.org/10.1038/nmicrobiol.2016.48>
- Ito, T., Asano, Y., Tanaka, Y., Takabe, T. (2011). Regulation of biosynthesis of dimethylsulfoniopropionate and its uptake in sterile mutant of *Ulva pertusa* (Chlorophyta). *Journal of Phycology*, 47(3), 517-523. <https://doi.org/10.1111/j.1529-8817.2011.00977.x>
- Kiene, R. P., Linn, L. J., González, J. G., Moran, M. A., & Bruton, J. A. (1999). Dimethylsulfoniopropionate and Methanethiol Are Important Precursors of Methionine and Protein-Sulfur in Marine Bacterioplankton. *Appl. Environ. Microbiol.*, 65(10), 4549-4558.
- Kinoshita, K., & Nakamura, H. (2003). Protein informatics towards function identification. *Current Opinion in Structural Biology*, 13, 296-400. [https://doi.org/10.1016/s0959-440x\(03\)00074-5](https://doi.org/10.1016/s0959-440x(03)00074-5)
- Kishino, H., & Hasegawa, M. (1989). Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. *Journal of Molecular Evolution*, 29(2), 170–179. <https://doi.org/10.1007/BF02100115>
- Klemetsen, T., Raknes, I. A., Fu, J., Agafonov, A., Balasundaram, S. V., Tartari, G., ... Willassen, N. P. (2018). The MAR databases: development and implementation of databases specific for marine metagenomics. *Nucleic Acids Research*, 46(D1), D692–D699. <https://doi.org/10.1093/nar/gkx1036>
- Kumar, S., Stecher, G., & Tamura, K. (2016). MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Molecular Biology and Evolution*, 33(7), 1870–1874. <https://doi.org/10.1093/molbev/msw054>
- Kumar, S., Stecher, G., Suleski, M., & Hedges, S. B. (2017). TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. *Molecular Biology and Evolution*, 34(7), 1812–1819. <https://doi.org/10.1093/molbev/msx116>
- Lanfear, R., Frandsen, P. B., Wright, A. M., Senfeld, T., & Calcott, B. (2016). PartitionFinder 2: New Methods for Selecting Partitioned Models of Evolution for Molecular and Morphological Phylogenetic Analyses. *Molecular Biology and Evolution*, msw260. <https://doi.org/10.1093/molbev/msw260>
- Lartillot, N., & Philippe, H. (2004). A Bayesian Mixture Model for Across-Site Heterogeneities in the Amino-Acid Replacement Process. *Molecular Biology and Evolution*, 21(6), 1095–1109. <https://doi.org/10.1093/molbev/msh112>
- Lartillot, N., & Philippe, H. (2006). Computing Bayes Factors Using Thermodynamic Integration. *Systematic Biology*, 55(2), 195–207. <https://doi.org/10.1080/10635150500433722>
- Lartillot, N., Brinkmann, H., & Philippe, H. (2007). Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BMC Evolutionary Biology*, 7(Suppl 1), S4. <https://doi.org/10.1186/1471-2148-7-S1-S4>

- 828
- 829 Le, S. Q., & Gascuel, O. (2008). An Improved General Amino Acid Replacement Matrix.
- 830 *Molecular Biology and Evolution*, 25(7), 1307–1320. <https://doi.org/10.1093/molbev/msn067>
- 831
- 832 Lovelock, J. E. (1983). Gaia as Seen Through the Atmosphere. In P. Westbroek & E. W. de Jong
- 833 (Eds.), *Biomineralization and Biological Metal Accumulation* (pp. 15–25).
- 834 https://doi.org/10.1007/978-94-009-7944-4_2
- 835
- 836 Luo, H., Csúros, M., Hughes, A. L., & Moran, M. A. (2013). Evolution of Divergent Life History
- 837 Strategies in Marine Alphaproteobacteria. *MBio*, 4(4). <https://doi.org/10.1128/mBio.00373-13>
- 838
- 839 Luo, H., & Moran, M. A. (2014). Evolutionary Ecology of the Marine Roseobacter Clade.
- 840 *Microbiology and Molecular Biology Reviews*, 78(4), 573–587.
- 841 <https://doi.org/10.1128/MMBR.00020-14>
- 842
- 843 Magalhães, C., Salgado, P., Kiene, R. P., & Bordalo, A. A. (2012). Influence of salinity on
- 844 dimethyl sulfide and methanethiol formation in estuarine sediments and its side effect on nitrous
- 845 oxide emissions. *Biogeochemistry*, 110(1–3), 75–86. <https://doi.org/10.1007/s10533-011-9690-z>
- 846
- 847 Moran, M. A., Belas, R., Schell, M. A., González, J. M., Sun, F., Sun, S., ... & Buchan, A.
- 848 (2007). Ecological Genomics of Marine Roseobacters. *Applied and Environmental Microbiology*,
- 849 73(14), 4559–4569. <https://doi.org/10.1128/AEM.02580-06>
- 850
- 851 Nelson, A. D. L., Haug-Baltzell, A. K., Davey, S., Gregory, B. D., & Lyons, E. (2018). EPIC-
- 852 CoGe: managing and analyzing genomic data. *Bioinformatics*, 34(15), 2651–2653.
- 853 <https://doi.org/10.1093/bioinformatics/bty106>
- 854
- 855 Newton, R. J., Griffin, L. E., Bowles, K. M., Meile, C., Gifford, S., Givens, C. E., ... & Moran,
- 856 M. A. (2010). Genome characteristics of a generalist marine bacterial lineage. *The ISME Journal*,
- 857 4(6), 784–798. <https://doi.org/10.1038/ismej.2009.150>
- 858
- 859 Nguyen, L. T., Schmidt, H. A., von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: A Fast and
- 860 Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular*
- 861 *Biology and Evolution*, 32(1), 268–274. <https://doi.org/10.1093/molbev/msu300>
- 862
- 863 Oren, A., Larimer, F., Richardson, P., Lapidus, A., & Csonka, L. N. (2005). How to be
- 864 moderately halophilic with broad salt tolerance: clues from the genome of *Chromohalobacter*
- 865 *salexigens*. *Extremophiles*, 9(4), 275–279. <https://doi.org/10.1007/s00792-005-0442-7>
- 866
- 867 Pál, C., Papp, B., & Lercher, M. J. (2006). An integrated view of protein evolution. *Nature*
- 868 *Reviews Genetics*, 7(5), 337–348. <https://doi.org/10.1038/nrg1838>
- 869
- 870 Puigbo, P., García-Vallve, S., & McInerney, J. O. (2007). TOPD/FMTS: a new software to
- 871 compare phylogenetic trees. *Bioinformatics*, 23(12), 1556–1558.
- 872 <https://doi.org/10.1093/bioinformatics/btm135>
- 873
- 874 R Core Team. (2017). R: A language and environment for statistical computing. R Foundation for
- 875 Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>
- 876

- 877 Raina, J.-B., Dinsdale, E. A., Willis, B. L., & Bourne, D. G. (2010). Do the organic sulfur
878 compounds DMSP and DMS drive coral microbial associations? *Trends in Microbiology*, 18(3),
879 101–108. <https://doi.org/10.1016/j.tim.2009.12.002>
880
- 881 Rambaut, A., & Drummond, A.J. (2002, 2010). TreeAnnotator, v1.6.1. Available from
882 <http://beast.bio.ed.ac.uk/>.
883
- 884 Rambaut, A. (2009). FigTree, version 1.4.3. Available from
885 <http://tree.bio.ed.ac.uk/software/figtree>.
886
- 887 Rambaut, A., & Drummond, A.J. (2013). Tracer v1.6. Available from: URL
888 <http://beast.bio.ed.ac.uk/Tracer>.
889
- 890 Ravenhall, M., Škunca, N., Lassalle, F., & Dessimoz, C. (2015). Inferring Horizontal Gene
891 Transfer. *PLOS Computational Biology*, 11(5), e1004095.
892 <https://doi.org/10.1371/journal.pcbi.1004095>
893
- 894 Reisch, C. R., Moran, M. A., & Whitman, W. B. (2008). Dimethylsulfoniopropionate-Dependent
895 Demethylase (DmdA) from *Pelagibacter ubique* and *Silicibacter pomeroyi*. *Journal of*
896 *Bacteriology*, 190(24), 8018–8024. <https://doi.org/10.1128/JB.00770-08>
897
- 898 Reisch, C. R., Moran, M. A., & Whitman, W. B. (2011a). Bacterial Catabolism of
899 Dimethylsulfoniopropionate (DMSP). *Frontiers in Microbiology*, 2.
900 <https://doi.org/10.3389/fmicb.2011.00172>
901
- 902 Reisch, C. R., Stoudemayer, M.J., Varaljay, V.A., Amster, I.J., Moran, M.A., & Whitman, W.B.
903 (2011b). Novel pathway for assimilation of dimethylsulphoniopropionate widespread in marine
904 bacteria. *Nature*, 473, 208–211. <https://doi.org/10.1038/nature10078>
905
- 906 Rost, B. (2002). Enzyme function less conserved than anticipated. *Journal of Molecular Biology*,
907 318(2), 595–608. [https://doi.org/10.1016/S0022-2836\(02\)0016-5](https://doi.org/10.1016/S0022-2836(02)0016-5)
908
- 909 Roy, A., Kucukural, A., & Zhang, Y. (2010). I-TASSER: a unified platform for automated
909 protein structure and function prediction. *Nature Protocols*, 5(4), 725–738.
910 <https://doi.org/10.1038/nprot.2010.5>
911
- 912 Salgado, P., Kiene, R., Wiebe, W., & Magalhães, C. (2014). Salinity as a regulator of DMSP
913 degradation in *Ruegeria pomeroyi* DSS-3. *Journal of Microbiology*, 52(11), 948–954.
914 <https://doi.org/10.1007/s12275-014-4409-1>
915
- 916 Sánchez-Pérez, G., Mira, A., Nyirö, G., Pasić, L., & Rodríguez-Valera, F. (2008). Adapting to
917 environmental changes using specialized paralogs. *Trends in Genetics*, 24(4), 154–158.
918 <https://doi.org/10.1016/j.tig.2008.01.002>
919
- 920 Schliep, K.P. (2011). phangorn: phylogenetic analysis in R. *Bioinformatics*, 27(4), 592–593.
921 <https://doi.org/10.1093/bioinformatics/btg706>
922
- 923 Schuller, D. J., Reisch, C. R., Moran, M. A., Whitman, W. B., & Lanzilotta, W. N. (2012).
924 Structures of dimethylsulfoniopropionate-dependent demethylase from the marine organism

- 925 *Pelagabacter ubique*: Structures and Mechanism of DMDA from *Pelagabacter ubique*. *Protein*
926 *Science*, 21(2), 289–298. <https://doi.org/10.1002/pro.2015>
- 927
- 928 Shimodaira, H., & Hasegawa, M. (1999). Multiple Comparisons of Log-Likelihoods with
929 Applications to Phylogenetic Inference. *Molecular Biology and Evolution*, 16(8), 1114–1116.
930 <https://doi.org/10.1093/oxfordjournals.molbev.a026201>
- 931
- 932 Shimodaira, H., (2002). An Approximately Unbiased Test of Phylogenetic Tree Selection.
933 *Systematic Biology*, 51(3), 492–508. <https://doi.org/10.1080/10635150290069913>
- 934
- 935 Siltberg-Liberes, J., Grahnen, J.A., & Liberis, D.A. (2011). The evolution of protein structures
936 and structural ensembles under functional constraint. *Genes (Basel)*, 2(4), 748–762. <https://doi.org/10.3390/genes2040748>
- 937
- 938
- 939 Stamatakis, A. (2006). Phylogenetic models of rate heterogeneity: a high performance computing
940 perspective. *Proceedings of the 20th IEE International Parallel and Distributed Processing*
941 *Symposium*, 253. <https://doi.org/10.1109/IPDPS.2006.1639535>.
- 942
- 943 Tang, K., Huang, H., Jiao, N. & Wu, C. H. (2010). Phylogenomic Analysis of Marine
944 Roseobacters. *PLoS One*, 5(7): e11604. <https://doi.org/10.1371/journal.pone.0011604>.
- 945
- 946 Tang, H., Thomas, P., & Xia, H. (n.d.). Reconstruction of the evolutionary history of gene gains
947 and losses since the last universal common ancestor. *ArXiv:1802.06035*.
- 948
- 949 Tripp, H. J., Kitner, J. B., Schwalbach, M. S., Dacey, J. W. H., Wilhelm, L. J., & Giovannoni, S.
950 J. (2008). SAR11 marine bacteria require exogenous reduced sulphur for growth. *Nature*, 452,
951 741. <https://doi.org/10.1038/nature06776>
- 952
- 953 Wu, S., & Zhang, Y. (2007). LOMETS: A local meta-threading-server for protein structure
954 prediction. *Nucleic Acids Research*, 35(10), 3375–3382. <https://doi.org/10.1093/nar/gkm251>
- 955
- 956 Xia, X. (2001). DAMBE: Software Package for Data Analysis in Molecular Biology and
957 Evolution. *Journal of Heredity*, 92(4), 371–373. <https://doi.org/10.1093/jhered/92.4.371>
- 958
- 959 Yang, Z. (1998). Likelihood ratio tests for detecting positive selection and application to primate
960 lysozyme evolution. *Molecular Biology and Evolution*, 15(5), 568–573.
961 <https://doi.org/10.1093/oxfordjournals.molbev.a025957>
- 962
- 963 Yang, Z. (2005). Bayes Empirical Bayes Inference of Amino Acid Sites Under Positive
964 Selection. *Molecular Biology and Evolution*, 22(4), 1107–1118. <https://doi.org/10.1093/molbev/msi097>
- 965
- 966
- 967 Yang, Z. (2007). PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology*
968 *and Evolution*, 24(8), 1586–1591. <https://doi.org/10.1093/molbev/msm088>
- 969
- 970 Yang, Z., & dos Reis, M. (2011). Statistical Properties of the Branch-Site Test of Positive
971 Selection. *Molecular Biology and Evolution*, 28(3), 1217–1228. <https://doi.org/10.1093/molbev/msq303>
- 972
- 973

- 974 Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J., & Zhang, Y. (2015). The I-TASSER Suite:
975 protein structure and function prediction. *Nature Methods*, 12(1), 7–8.
976 <https://doi.org/10.1038/nmeth.3213>
977
- 978 Yoch, D. C. (2002). Dimethylsulfoniopropionate: Its Sources, Role in the Marine Food Web, and
979 Biological Degradation to Dimethylsulfide. *Applied and Environmental Microbiology*, 68(12),
980 5804–5815. <https://doi.org/10.1128/AEM.68.12.5804-5815.2002>
981
- 982 Zakon, H. H. (2002). Convergent Evolution on the Molecular Level. *Brain, Behavior and*
983 *Evolution*, 59(5–6), 250–261. <https://doi.org/10.1159/000063562>
984
- 985 Zhang, Jianzhi. (2003). Evolution by gene duplication: an update. *Trends in Ecology &*
986 *Evolution*, 18(6), 292–298. [https://doi.org/10.1016/S0169-5347\(03\)00033-8](https://doi.org/10.1016/S0169-5347(03)00033-8)
987
- 988 Zhang, J. (2005). Evaluation of an Improved Branch-Site Likelihood Method for Detecting
989 Positive Selection at the Molecular Level. *Molecular Biology and Evolution*, 22(12), 2472–2479.
990 <https://doi.org/10.1093/molbev/msi237>
991
- 992 Zhang, Y. (2005). TM-align: a protein structure alignment algorithm based on the TM-score.
993 *Nucleic Acids Research*, 33(7), 2302–2309. <https://doi.org/10.1093/nar/gki524>
994
- 995 Zubkov, M. V., Fuchs, B. M., Archer, S. D., Kiene, R. P., Amann, R., & Burkill, P. H. (2001).
996 Linking the composition of bacterioplankton to rapid turnover of dissolved
997 dimethylsulphoniopropionate in an algal bloom in the North Sea. *Environmental Microbiology*,
998 3(5), 304–311. <https://doi.org/10.1046/j.1462-2920.2001.00196.x>

Table 1(on next page)

Topology tests of DmdA phylogenetic tree with respect to species tree

*p-values under the Kishino-Hasegawa (KH) test, the Shimodaira-Hasegawa (SH) test and the approximately unbiased (AU) test, respectively.

Group	pKH*	pSH*	pAU*
DmdA family	0.0010	0.0010	0.0001

Table 2 (on next page)

Structural model predicted by I-TASSER for each sequence used in the evolutionary study of DmdA gene family and the best identified structural analogs in PDB by TM-align.

¹A confidence score for estimating the quality of predicted models ²A standard for measuring structural similarity between two structures ³The Protein Data Bank structure name ⁴DmdA DMSP-dependent demethylase ⁵Glycine cleavage system T protein ⁶Dimethylglycine dehydrogenase complexed with tetrahydrofolate

Sequence information		Predicted model		Best structural analog from PDB			
Gene name	ID	C-score ¹	TM-score ² ± dev	Gene name	Organism	PDB ID ³	TM-score
<i>dmdA</i> ⁴	AAV95190.1	1.45	0.92 ± 0.06	<i>dmdA</i>	<i>Ca. P. ubiqu</i> HTCC1062	3tfhA	0.974
<i>dmdA</i>	AHD01041.1	1.69	0.95 ± 0.05	<i>dmdA</i>	<i>Ca. P. ubiqu</i> HTCC1062	3tfhA	0.990
<i>dmdA</i>	WP_047029467.1	2	0.99 ± 0.04	<i>dmdA</i>	<i>Ca. P. ubiqu</i> HTCC1062	3tfhA	0.997
<i>dmdA</i>	WP_048536000.1	2	0.99 ± 0.04	<i>dmdA</i>	<i>Ca. P. ubiqu</i> HTCC1062	3tfhA	0.997
<i>dmdA</i>	AHM05061.1	2	0.99 ± 0.04	<i>dmdA</i>	<i>Ca. P. ubiqu</i> HTCC1062	3tfhA	0.989
<i>dmdA</i>	ABF64177.1	1.62	0.94 ± 0.05	<i>dmdA</i>	<i>Ca. P. ubiqu</i> HTCC1062	3tfiA	0.947
<i>dmdA</i>	WP_065273401.1	2	0.99 ± 0.04	<i>dmdA</i>	<i>Ca. P. ubiqu</i> HTCC1062	3tfhA	0.997
<i>dmdA</i>	WP_076627280.1	2	0.99 ± 0.04	<i>dmdA</i>	<i>Ca. P. ubiqu</i> HTCC1062	3tfhA	0.997
<i>dmdA</i>	AEI94210.1	2	0.99 ± 0.04	<i>dmdA</i>	<i>Ca. P. ubiqu</i> HTCC1062	3tfhA	0.997
<i>dmdA</i>	ABG31871.1	2	0.99 ± 0.04	<i>dmdA</i>	<i>Ca. P. ubiqu</i> HTCC1062	3tfhA	0.997
<i>dmdA</i>	ABD55296.1	2	0.99 ± 0.04	<i>dmdA</i>	<i>Ca. P. ubiqu</i> HTCC1062	3tfhA	0.997
<i>dmdA</i>	WP_049834197.1	2	0.99 ± 0.04	<i>dmdA</i>	<i>Ca. P. ubiqu</i> HTCC1062	3tfhA	0.997
<i>dmdA</i>	AGI72139.1	2	0.99 ± 0.04	<i>dmdA</i>	<i>Ca. P. ubiqu</i> HTCC1062	3tfhA	0.997
<i>dmdA</i>	ABV94056.1	2	0.99 ± 0.04	<i>dmdA</i>	<i>Ca. P. ubiqu</i> HTCC1062	3tfhA	0.998
<i>dmdA</i>	AAZ21068.1	2	0.99 ± 0.04	<i>dmdA</i>	<i>Ca. P. ubiqu</i> HTCC1062	3tfhA	0.997
<i>dmdA</i>	AFS46782.1	1.95	0.99 ± 0.04	<i>dmdA</i>	<i>Ca. P. ubiqu</i> HTCC1062	3tfhA	0.997
<i>dmdA</i>	AFS48343.1	2	0.99 ± 0.04	<i>dmdA</i>	<i>Ca. P. ubiqu</i> HTCC1062	3tfhA	0.995
<i>dmdA</i>	AGI68776.1	2	0.99 ± 0.04	<i>dmdA</i>	<i>Ca. P. ubiqu</i> HTCC1062	3tfhA	0.997
<i>dmdA</i>	ASJ73090.1	1.77	0.96 ± 0.05	<i>dmdA</i>	<i>Ca. P. ubiqu</i>	3tfhA	0.956

					HTCC1062		
<i>dmdA</i>	ADE38317.1	1.96	0.99 ± 0.04	<i>dmdA</i>	<i>Ca. P. ubiqu</i> HTCC1062	3tfhA	0.992
<i>gcvT</i> ⁵	AEM59334.1	2.53	0.42 ± 0.14	<i>dmgdh</i> ⁶	<i>Rattus</i> <i>norvegicus</i>	4p9sA	0.637
<i>gcvT</i>	WP_09638981 6.1	0.48	0.78 ± 0.10	<i>dmgdh</i>	<i>Rattus</i> <i>norvegicus</i>	4p9sA	0.885
<i>gcvT</i>	CAJ51984.2	0.23	0.68 ± 0.12	<i>dmgdh</i>	<i>Rattus</i> <i>norvegicus</i>	4p9sA	0.855
<i>gcvT</i>	CCC39909.1	-0.06	0.71 ± 0.12	<i>dmgdh</i>	<i>Rattus</i> <i>norvegicus</i>	4p9sA	0.865
<i>gcvT</i>	AFS48830.1	0.64	0.80 ± 0.09	<i>dmgdh</i>	<i>Rattus</i> <i>norvegicus</i>	4p9sA	0.894
<i>gcvT</i>	AGM40509.1	0.55	0.79 ± 0.09	<i>dmgdh</i>	<i>Rattus</i> <i>norvegicus</i>	4p9sA	0.887
<i>gcvT</i>	AHI32422.1	0.61	0.80 ± 0.09	<i>dmgdh</i>	<i>Rattus</i> <i>norvegicus</i>	4p9sA	0.896
<i>gcvT</i>	WP_05311283 5.1	0.56	0.79 ± 0.09	<i>dmgdh</i>	<i>Rattus</i> <i>norvegicus</i>	4p9sA	0.997
<i>gcvT</i>	CBV41552.1	0.68	0.81 ± 0.09	<i>dmgdh</i>	<i>Rattus</i> <i>norvegicus</i>	4p9sA	0.906
<i>gcvT</i>	WP_07194184 1.1	1.11	0.87 ± 0.07	<i>dmgdh</i>	<i>Rattus</i> <i>norvegicus</i>	4p9sA	0.997
<i>gcvT</i>	AAV94935.1	1.96	0.99 ± 0.04	<i>dmdA</i>	<i>Ca. P. ubiqu</i> HTCC1062	3tfhA	0.994
<i>gcvT</i>	AII87408.1	1.64	0.94 ± 0.05	<i>dmdA</i>	<i>Ca. P. ubiqu</i> HTCC1062	3tfhA	0.985
<i>gcvT</i>	ADE40415.1	2	0.99 ± 0.04	<i>dmdA</i>	<i>Ca. P. ubiqu</i> HTCC1062	3tfhA	0.995
<i>gcvT</i>	AHM03102.1	1.69	0.95 ± 0.05	<i>dmdA</i>	<i>Ca. P. ubiqu</i> HTCC1062	3tfhA	0.981
<i>gcvT</i>	WP_07197292 0.1	1.99	0.99 ± 0.04	<i>dmdA</i>	<i>Ca. P. ubiqu</i> HTCC1062	3tfhA	0.988
<i>gcvT</i>	BAN00949.1	1.13	0.87 ± 0.07	<i>dmg</i>	<i>Arthrobacter</i> <i>globiformis</i>	1pj6A	0.948
<i>gcvT</i>	WP_05381998 0.1	1.71	0.95 ± 0.05	<i>dmdA</i>	<i>Ca. P. ubiqu</i> HTCC1062	3tfhA	0.988
<i>gcvT</i>	ABF63906.1	1.53	0.93 ± 0.06	<i>dmdA</i>	<i>Ca. P. ubiqu</i> HTCC1062	3tfhA	0.960
<i>gcvT</i>	AGI71303.1	1.65	0.95 ± 0.05	<i>dmdA</i>	<i>Ca. P. ubiqu</i>	3tfhA	0.960

					HTCC1062		
<i>gcvT</i>	AIH85872.1	1.52	0.93 ± 0.06	<i>dmdA</i>	<i>Ca. P. ubiqu</i> HTCC1062	3tfhA	0.960
<i>gcvT</i>	WP_06754545 2.1	1.59	0.94 ± 0.05	<i>dmdA</i>	<i>Ca. P. ubiqu</i> HTCC1062	3tfhA	0.961
<i>gcvT</i>	ADE39159.1	1.50	0.92 ± 0.06	<i>dmdA</i>	<i>Ca. P. ubiqu</i> HTCC1062	3tfhA	0.950
<i>gcvT</i>	AGI71500.1	1.47	0.92 ± 0.06	<i>dmdA</i>	<i>Ca. P. ubiqu</i> HTCC1062	3tfhA	0.949
<i>gcvT</i>	AFS47213.1	1.66	0.95 ± 0.05	<i>dmdA</i>	<i>Ca. P. ubiqu</i> HTCC1062	3tfhA	0.966
<i>gcvT</i>	AFS48354.1	1.60	0.94 ± 0.05	<i>dmdA</i>	<i>Ca. P. ubiqu</i> HTCC1062	3tfhA	0.963
<i>gcvT</i>	WP_05382073 0.1	0.34	0.67 ± 0.13	<i>dmgdh</i>	<i>Rattus</i> <i>norvegicus</i>	4p9sA	0.874
<i>gcvT</i>	WP_06535384 5.1	1.56	0.93 ± 0.06	<i>dmdA</i>	<i>Ca. P. ubiqu</i> HTCC1062	3tfhA	0.961
<i>gcvT</i>	Ancestral DmdA and non-DmdA sequence	1.25	0.89 ± 0.07	<i>gcvT</i>	<i>Thermotoga</i> <i>maritima</i>	1wooA	0.960
<i>dmdA</i>	Ancestral DmdA sequence	2	0.99 ± 0.04	<i>dmdA</i>	<i>Ca. P. ubiqu</i> HTCC10626	3tfhA	0.997
<i>gcvT</i>	Ancestral non- DmdA sequence	0.76	0.82 ± 0.09	<i>dmgdh</i>	<i>Rattus</i> <i>norvegicus</i>	4p9sA	0.940

Table 3(on next page)

Divergence time estimates in Mya, and node 95% highest posterior density interval for the clades of the MRCA of *Halobacteriales*, SAR11 and *Alphaproteobacteria* from each set of calibration priors.

Taxonomic group of MRCA	Clade	Age	95% HPD
<i>Halobacteriales</i> (455)	Mrca1	438	311.1 – 572.3
SAR11 (826)	Mrca2	827.5	588.3 – 1089.8
<i>Alphaproteobacteria</i> (2480)	Mrca3	2118.6	1543 – 2717.1

Table 4(on next page)

Parameters of branch-models.

* ω values are shown in Supplementary Fig 19. ¹Log-likelihood score under the model

²Likelihood ratio test.

Model	ω_1	ω_2	$-\ln L^1$	LRT^2	P-value
One ω (one-ratio)	0.08518	NA	-14580.019867	NA	NA
Two ω (two-ratio)	0.0767	0.1494	-14568.131038	23.777658	0.0
38 ω (free-ratio)	*	*	-14428.881747	302.27624	0

Table 5 (on next page)

Parameters of PAML branch-site models.

Branch identifiers follow the nomenclature of Supplementary Fig 19 Colors show same mutation in different lineages. *Amino acids refer to the first sequence in the alignment:

AFS48343.1¹Log-likelihood score under the model under Null model²Log-likelihood score under alternative model³Likelihood ratio test⁴Uncorrected p-value: raw- p-value without correction for multiple testing⁵p-value corrected for multiple testing by Bonferroni

Branch	Ho (-lnL) ¹	Ha (-lnL) ²	LRT ³	P-value ⁴	Corrected P-value ⁵	Pos. Selected sites* (BEB>0.95)
ADE38317.1	-14465.244	-14463.099	4.290	0.038	0.767	NA
AAV95190.1	-14476.763	-14476.763	0	1	1	NA
AHD01041.1	-14476.763	-14476.763	0	1	1	NA
WP_047029467.1	-14476.763	-14437.565	78.397	0.00	0.00	7V; 17E; 47H; 65D; 68Y; 87Y; 89A; 152K; 157M; 163N; 203V; 279G; 290P; 319T; 320H
WP_048536000.1	-14476.763	-14476.763	0	1	1	NA
AHM05061.1	-14466.948	-14460.844	12.206	0.000	0.000	17E ; 152K; 178E; 285V
ABF64177.1	-14476.763	-14476.763	0	1	1	NA
WP_065273401.1	-14476.763	14476.763	0	1	1	NA
WP_076627280.1	-14476.763	14476.763	0	1	1	NA
AEI94210.1	-14476.763	-14476.763	0	1	1	NA
ABG31871.1	-14476.763	-14476.763	0	1	1	NA
ABD55296.1	-14476.764	-14476.764	0	1	1	NA
WP_049834197.1	-14476.763	-14476.763	0	1	1	NA
AGI72139.1	-14476.763	-14476.763	0	1	1	NA
AGI68776.1	-14476.763	-14476.763	0	1	1	NA
ABV94056.1	-14462.942	-14454.885	16.112	0.000	0.000	87Y; 152K; 243N; 247L; 257F
ASJ730990.1	-14463.474	-14461.176	4.595	0.032	0.641	NA
AAZ21068.1	-14465.122	-14462.171	5.902	0.015	0.302	NA
AFS46782.1	-14467.961	-14464.484	6.954	0.008	0.167	NA

AFS48343.1	-14460.566	-14425.923	31.802	0.000	0.000	4S; 5A; 9S; 35S; 38V; 70T; 83D; 84H; 85I; 91V; 94D; 95Q; 103L; 109P; 119T; 139T; 155E; 158K; 168N; 176N; 179F; 210L; 211R; 217G; 231S; 253A; 259P; 270Q; 274V; 277S; 292N; 298T; 305S; 311C; 321T
Ancestral branch to the DmdA clade	-28761.935	-28758.081	7.7084	0.005	0.010	39Q
Ancestral branch to the non-DmdA clade	-28770.533	-28766.874	7.3182	0.006	0.013	-

Table 6 (on next page)

Parameter estimates of models evaluating functional divergence of DmdA and non-DmdA after gene duplication.

*Sites with predicted functional divergence between DmdA and non-DmdA at significance

(BEB > 0.95) ¹NP: number of free parameters in the model ²Average over all sites ³Kappa

⁴Log-likelihood score under the model ⁵Likelihood ratio test

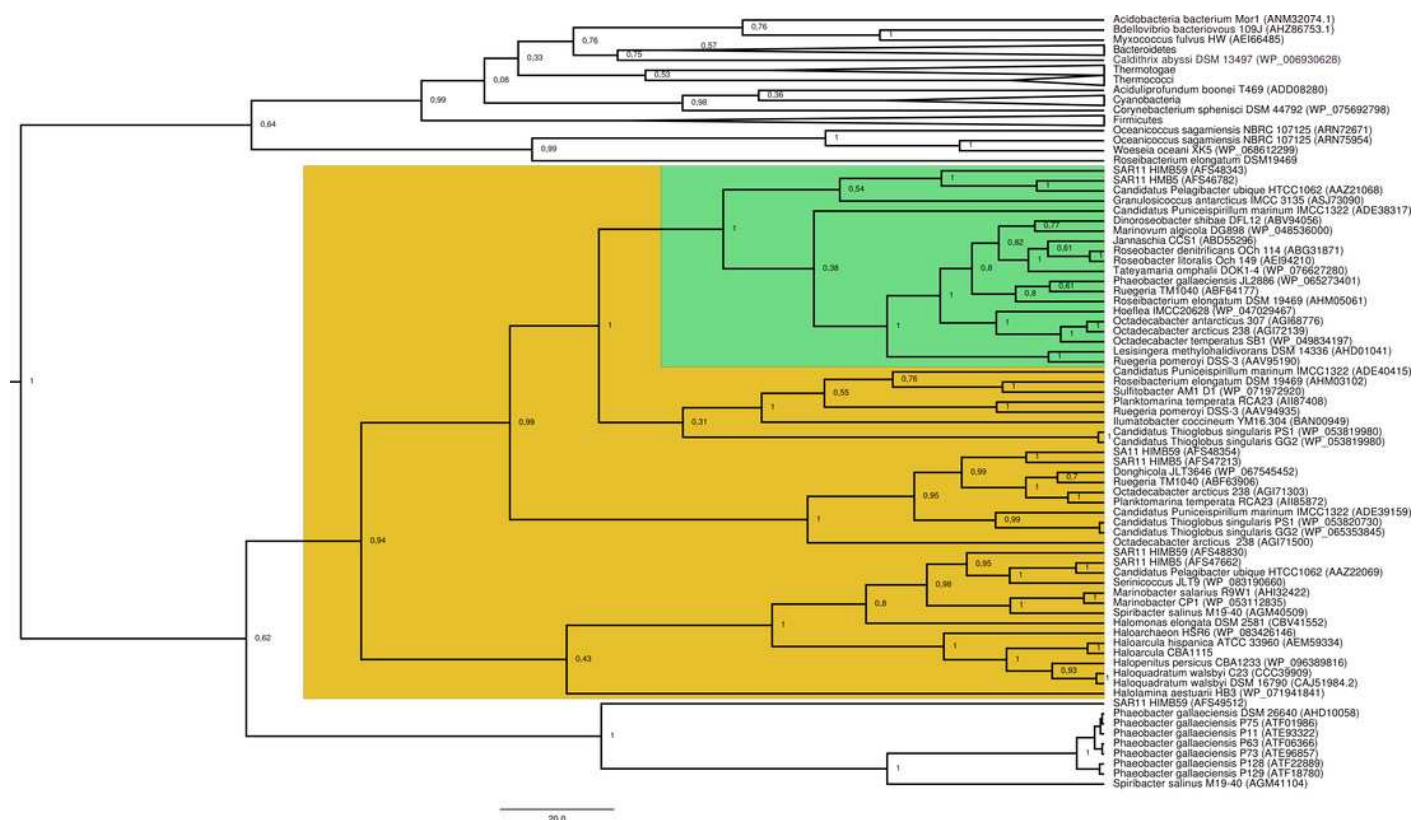
Model	NP ¹	ω^2	Site class 0		Site class 1		Site class 2		K ³	-LnL ⁴	LTR ⁵	P-value	Divergent sites*
			ω_0	p_0	ω_1	p_1	ω_2	p_2					
M0	95	0.053							1.341	-28818.866	na	na	
M3 (k=3)	99	0.058	0.006	0.238	0.045	0.506	0.132	0.255	1.342	-28079.171	1479.391	0.00	
MD (k=3)	100		0.006	0.235	0.042	0.492	ω_{2a} 0.100 ω_{2b} :0.169	0.272	1.337	-28061.808	34.725	0.00	2V, 9Q, 12E, 14Y, 16Q, 17A, 28S, 32N, 36N, 37H, 52E, 57D, 58Y, 60T, 62L, 69S, 70Q, 71A, 72K, 73D, 77Y, 85Q, 98K, 101T, 118I, 127T, 132N, 142F, 146K, 147R, 150E, 156K, 157R, 158Y, 159A, 161N, 163H, 164E, 166L, 185D, 187V, 188Q, 192Q, 194L, 198K, 199D, 211S, 218M, 226A, 229S, 230P, 240K, 241K, 242S, 244S, 247I, 248M, 250D, 253T, 254L, 258C, 259Y, 264G, 265K,

													272Q, 273L, 274D, 275Q, 276D, 277L, 278K, 280Q, 283K, 285T, 286N, 287L
--	--	--	--	--	--	--	--	--	--	--	--	--	---

Figure 1

GcvT phylogenetic tree based on 20 DmdA orthologs protein sequences and 184 DmdA homologs using **Beast** and the same parameters set for molecular dating but with 100 million generations.

DmdA sequences are indicated with green color and closer homologs for those with yellow color. Tip labels include a maximum e-value < e-50.



Phylogenetic tree of DmdA based on 20 DmdA orthologs protein sequences and 28 DmdA homologs (more information in Supplementary Table 1) using RaxML.

A non-parametric bootstrap is shown to establish the support for the clades. DmdA sequences are indicated with blue branch. Tip labels show color for first dmdA gene identified or taxonomy classification. Tip labels include a maximum e-value $< e^{-80}$.

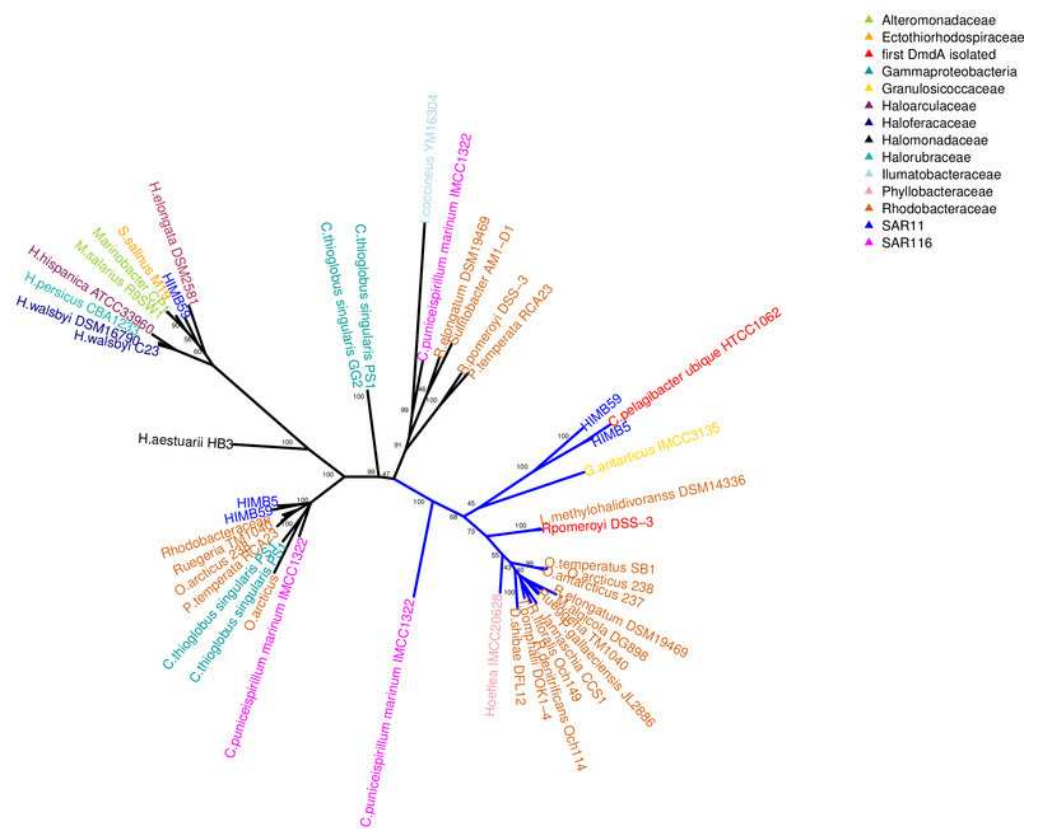


Figure 3

Phylogenetic tree of DmdA based on 20 DmdA orthologs protein sequences and 28 DmdA homologs using BEAST2.

Bayesian posterior probabilities (PP) is shown to establish the support for the clades. Red color indicates DmdA clade.

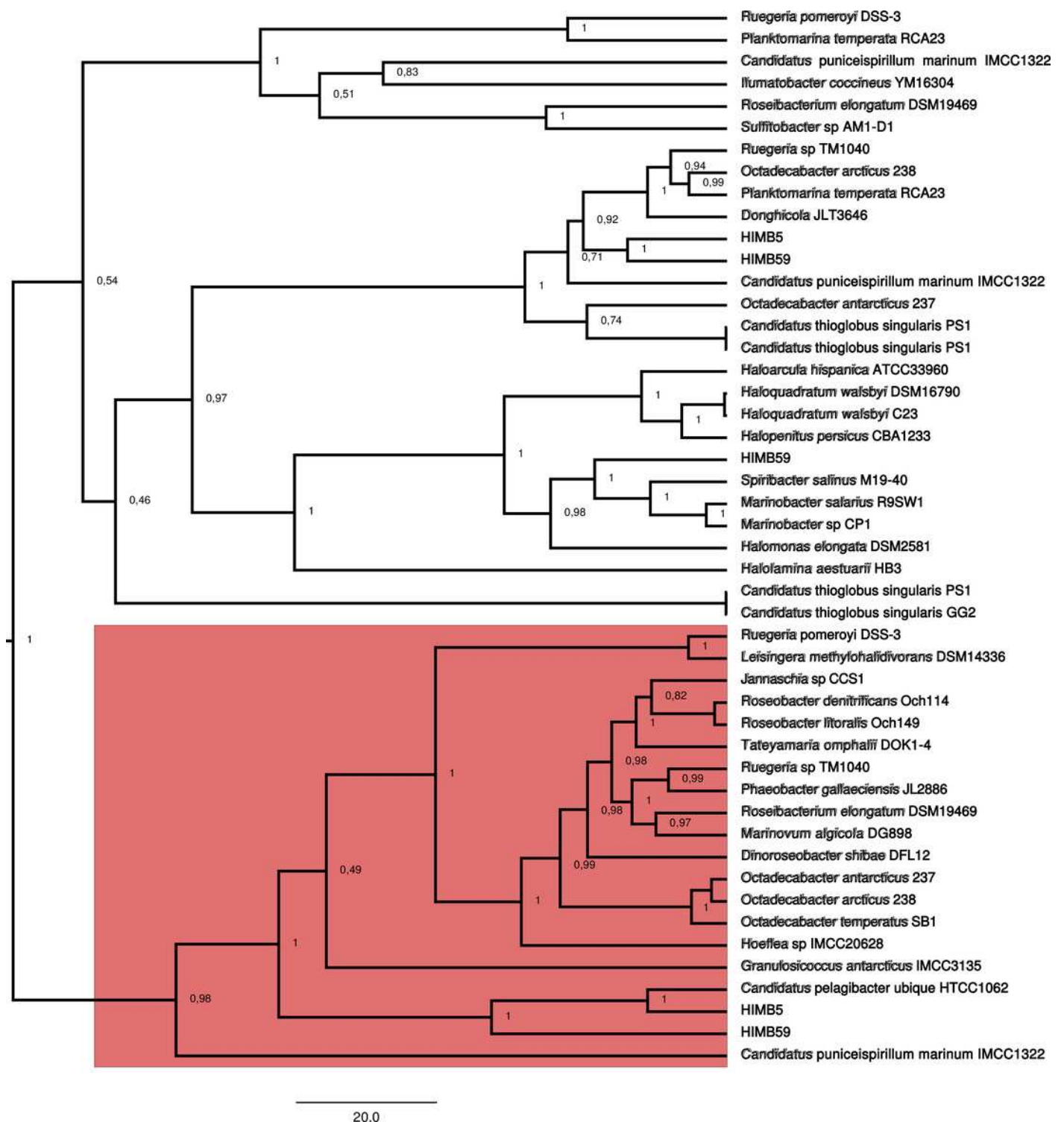


Figure 4

(Upper) BEAST divergence time estimates from *dmdA* and non-*dmdA* genes under uncorrelated relaxed clock model and Birth-death tree model. (Lower) Absolute time scale in Ma.

Nodes are at mean divergence times and gray bars represent 95% HPD of node age. Nodes used as calibrated priors in BEAST analysis are marked as mrca1, mrca2 and mrca3 as well as colored. Arrows indicate duplication events occurred 1894 Mya (red), 300 Mya (blue) and 1000 Mya (green).

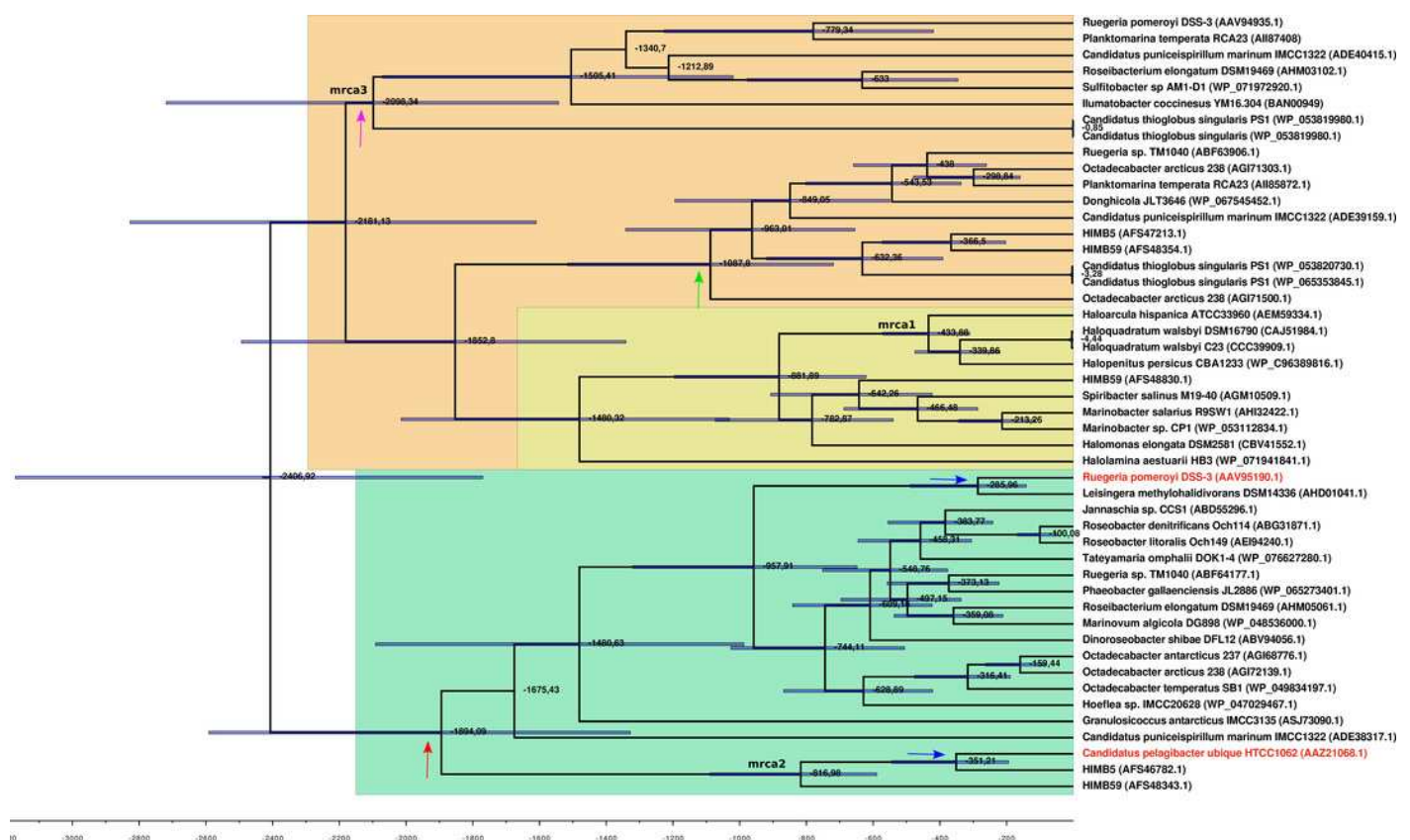


Figure 5

Posterior probabilities for dN/dS categories under the M3 model. Grey, red and blue bars depict the three dN/dS categories (values for each category are provide in the key).

Sites that are mostly grey denote codons under strong purifying selection, whereas those predominantly red show codons under weaker purifying selection. Red, blue and grey colors indicate codon sites with $\omega_2 = 0.2483$, $\omega_1 = 0.06923$ and $\omega_0 = 0.00485$, respectively.

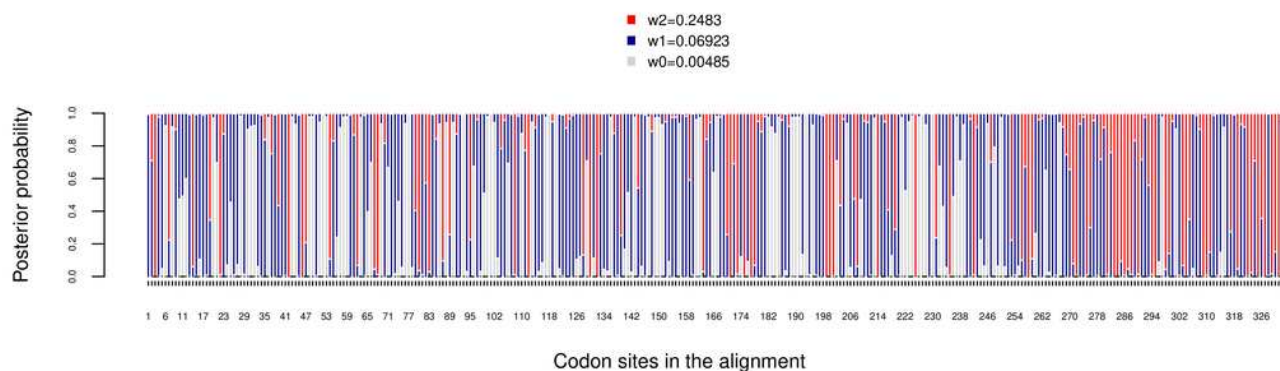


Figure 6

Tertiary structure of DmdA (PDB: 3tfh) with sites under episodic positive selection mapped in yellow color through Pymol.

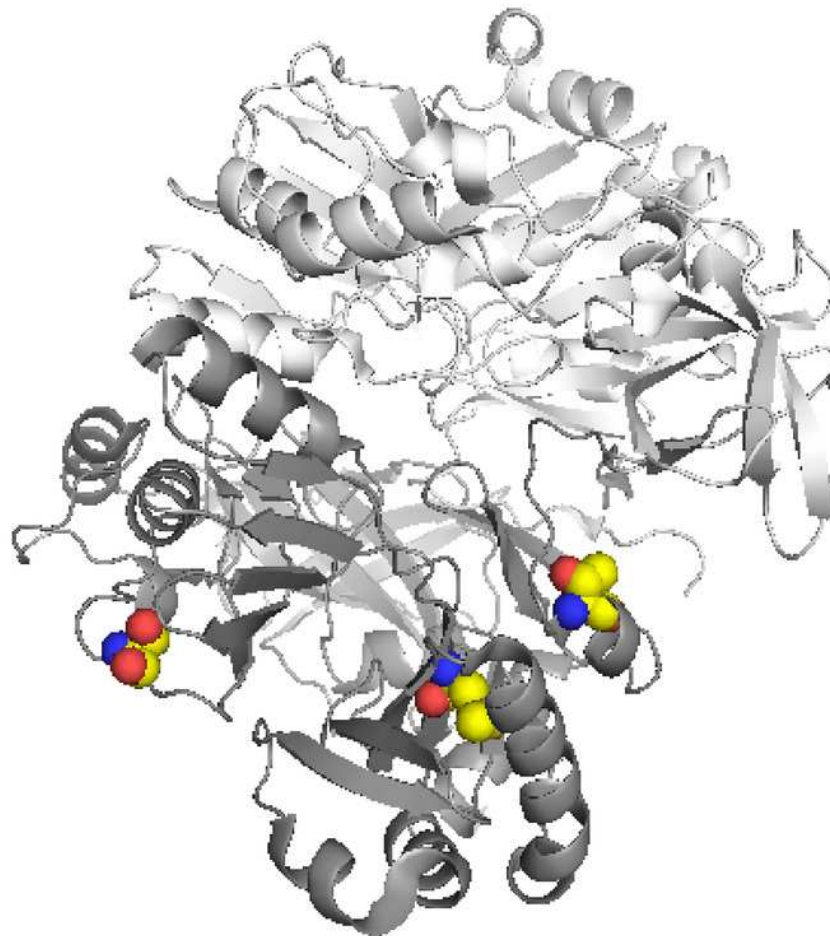


Figure 7

Hypothesis of DmdA evolution. BI phylogeny under uncorrelated relaxed clock model and Birth-death tree model.

Node names represent the ancestral sequences reconstructed; GcvT prior to main duplication, DmdA for DmdA clade and DmgdH for non-DmdA clade. In DmdA clade, blue color represents ecoparalogs where pI is < 5.7 and they are adapted to less concentration of DMSP in comparison with DmdA paralogs (red color) which have $pI \Rightarrow 6.5$. In non-DmdA clade, yellow branches represents paralogs with DmgdH tertiary structure and black branches paralogs with DmdA tertiary structure.

