Posted on 12 May 2023 — The copyright holder is the author/funder. All rights reserved. No reuse without permission. — https://doi.org/10.22541/au.168388964.46166236/v1 — This a preprint and has not been peer reviewed. Data may be preliminary.

HMSS2: an advanced tool for the analysis of sulfur metabolism, including organosulfur compound transformation, in genome and metagenome assemblies

Tomohisa Tanabe[1] and Christiane Dahl[1]

[1]Rheinische Friedrich-Wilhelms-Universität Bonn

May 12, 2023

Abstract

The global sulfur cycle has implications for human health, climate change, biogeochemistry, and bioremediation. The organosulfur compounds that participate in this cycle not only represent a vast reservoir of sulfur, but are also used by prokaryotes as sources of energy and/or carbon. Closely linked to the inorganic sulfur cycle, it involves the interaction of prokaryotes, eukaryotes, and chemical processes. However, ecological and evolutionary studies of the conversion of organic sulfur compounds are hampered by the poor conservation of the relevant pathways and their variation even within strains of the same species. In addition, several proteins involved in the conversion of sulfonated compounds are related to proteins involved in sulfur dissimilation or turnover of other compounds. Therefore, the enzymes involved in the metabolism of organic sulfur compounds are usually not correctly annotated in public databases. To address this challenge, we have developed HMSS2, a profiled Hidden Markov Model-based tool for rapid annotation and synteny analysis of organic and inorganic sulfur cycle proteins in prokaryotic genomes. Compared to its previous version (HMS-S-S), HMSS2 includes several new features. HMM-based annotation is now supported by non-homology criteria and covers the metabolic pathways of important organosulfur compounds, including dimethylsulfopropionate, taurine, isethionate, and sulfoquinovose. In addition, the calculation speed has been increased by a factor of four and the available output formats have been extended to include iTol compatible datasets, and customised sequence FASTA files

Title: HMSS2: an advanced tool for the analysis of sulfur metabolism, including organosulfur compound transformation, in genome and metagenome assemblies

Authors: **Tomohisa Sebastian Tanabe*** **and Christiane Dahl***

Institut für Mikrobiologie & Biotechnologie, Rheinische Friedrich-Wilhelms-Universität Bonn, Bonn, Germany

Running head: HMSS2, a sulfur metabolism analysis tool

*Correspondence: Tomohisa Sebastian Tanabe and Christiane Dahl
Institut für Mikrobiologie & Biotechnologie
Rheinische Friedrich Wilhelms-Universität Bonn
Meckenheimer Allee 168
53115 Bonn, Germany
Tel. +49-228-735591
Fax +49-228-737476
E-mails: s6totana@uni-bonn.de (T.S.T.), chdahl@uni-bonn.de (C.D.)

Keywords: Hidden Markov model (HMM) database, organosulfur compounds, sulfur metabolism, dimethylsulfpopropionate, sulfoquinovose

Abbreviations:

Abstract

The global sulfur cycle has implications for human health, climate change, biogeochemistry, and bioremediation. The organosulfur compounds that participate in this cycle not only represent a vast reservoir of sulfur, but are also used by prokaryotes as sources of energy and/or carbon. Closely linked to the inorganic sulfur cycle, it involves the interaction of prokaryotes, eukaryotes, and chemical processes. However, ecological and evolutionary studies of the conversion of organic sulfur compounds are hampered by the poor conservation of the relevant pathways and their variation even within strains of the same species. In addition, several proteins involved in the conversion of sulfonated compounds are related to proteins involved in sulfur dissimilation or turnover of other compounds. Therefore, the enzymes involved in the metabolism of organic sulfur compounds are usually not correctly annotated in public databases. To address this challenge, we have developed HMSS2, a profiled Hidden Markov Model-based tool for rapid annotation and synteny analysis of organic and inorganic sulfur cycle proteins in prokaryotic genomes. Compared to its previous version (HMS-S-S), HMSS2 includes several new features. HMM-based annotation is now supported by non-homology criteria and covers the metabolic pathways of important organosulfur compounds, including dimethylsulfpopropionate, taurine, isethionate, and sulfoquinovose. In addition, the calculation speed has been increased by a factor of four and the available output formats have been extended to include iTol compatible datasets, and customised sequence FASTA files.

1 INTRODUCTION

The global organic sulfur cycle occurs in both terrestrial and aquatic environments and involves the interplay of prokaryotes, eukaryotes, and chemical processes. Millions of megatonnes of sulfonated compounds are produced annually by biological and industrial processes. These compounds not only represent a vast reservoir of sulfur but can also be used by prokaryotes as a sources of energy and carbon (Moran & Durham, 2019). Understanding the mechanisms and ecological interactions of prokaryotes in the organic sulfur cycle is of great importance because the decomposition of organic sulfur compounds affects human health, bacterial virulence in infection (Dhouib et al., 2021), global warming, bioremediation processes such as wastewater treatment (Schäfer et al., 2010), and is linked to the biogeochemical cycling of sulfur between habitats (Koch & Dahl, 2018).

Sulfonated compounds can range from small size with only a C_1 carbon skeleton up to sulfonated lipids with long-chain alkanes, amino acids such as cysteine, or sulfur-containing cofactors with complex structures such as lipoate (Boden & Hutt, 2019; Goddard-Borger & Williams, 2017; Moran & Durham, 2019). While chemistry offers an infinite number of possible sulfonated compounds and new ones are being discovered all the time, these compounds often lack a described metabolic function or the pathways for their synthesis or degradation have not been elucidated (Thume et al., 2018). Only the most abundant sulfonated compounds, such as sulfoquinovose, dimethylsulfopropionate (DMSP), taurine, isethionate, cysteine, and methionine, have been studied biochemically in terms of synthesis and degradation pathways.

In aquatic environments, the anti-stress molecule DMSP is the most well-known organosulfur compound (Kiene et al., 2000). Mainly produced by macroalgae and phytoplankton, it is emitted by around 600 million tonnes per year. Bacterial DMSP degradation in the oceans, salt marshes, and coastal regions is the major source of dimethylsulfide (DMS), which is released at a rate of about 300 million tonnes per year (Moran & Durham, 2019). As a volatile compound, DMS affects atmospheric chemistry and global warming by forming cloud condensation nuclei that increase the reflection of solar radiation (Schäfer et al., 2010). In the context of the global sulfur cycle, DMS acts as a link between the terrestrial, atmospheric and aquatic environments (Lovelock et al., 1972). DMS-derived

carbon and sulfur are used as electron acceptors or donors during dissimilation, or are assimilated via the intermediates dimethlysulfone and methanesulfinate (Fig. 1).

Sulfonated lipids are estimated to be the largest reservoir of sulfur in terrestrial ecosystems (Goddard-Borger & Williams, 2017). Sulfoquinovose is a sulfonated glucose derivate and the most common part of the head group of sulfolipids which are integral part of thylakoid membranes of chloroplasts and photosynthetic systems. Mainly produced by plants, algae, and cyanobacteria its turnover rate has been estimated at around 10 billion tonnes per year (Goddard-Borger & Williams, 2017). The bacterial decomposition of sulfoquinovose involves several different pathways similar to the degradation of glucose (Fig. 2a), with the exception that smaller sulfonated compounds are often released, since complete utilisation with release of free sulfur by a single organism is often not possible (Wei et al., 2022). Release and scavenging of sulfonated intermediates is achieved by various transport systems (Fig. 2b). Sulfoquinovose decomposition and release of inorganic sulfur is then completed by pathways linked to taurine, isethionate and/or sulfoacetate (Fig. 2c). In summary, prokaryotic utilization of these organic compounds as sources of sulfur, carbon, and energy is far from being a uniform process and new metabolic pathways for the degradation of sulfonated compound are constantly being discovered (Boden et al., 2010; Koch & Dahl, 2018; Sharma et al., 2022; Wolf et al., 2022).

These processes are also closely linked to the availability of inorganic sulfur as the released sulfur is either assimilated or excreted as sulfate (Ruff et al., 2003), sulfite (Koch & Dahl, 2018; Li et al., 2022; Sharma et al., 2022), thiosulfate (De Zwart et al., 1997), tetrathionate (Boden et al., 2010) or sulfide (Peck et al., 2019). Indeed, the complete consumption of the volatile sulfonated C_1-compound DMS coupled with the oxidation of the thiosulfate formed as an intermediate, has been reported for a single organism, providing a new link between the organic and inorganic sulfur cycles (Koch & Dahl, 2018). However, the fate of the sulfur released from sulfonated compounds is often not known or assumed to be the same as in dissimilatory sulfur oxidation or reduction. The physiology and interactions of bacterial communities that release sulfur from sulfonated carbon compounds have been sparsely explored and the few existing studies are based on, or assume, sulfur cycling via dissimilatory sulfite reductases (Burrichter et al., 2021; Hanson et al., 2021; Wolf et al., 2022).

Ecological studies of organic sulfur compounds are difficult because their metabolism is poorly conserved across bacterial phylogeny and can even vary between strains of the same species. Thus, even within a species, predictions based on taxonomic assignment are not possible (Schäfer et al., 2010). As the functional annotation pipelines of public databases mainly focus on the synthesis of methionine and cysteine, the enzymes involved in the metabolism of organic sulfur compounds are usually not correctly annotated. Inaccurate annotation in public databases is exacerbated by the fact that several proteins involved in the conversion of sulfonated compounds are related to proteins involved in sulfur dissimilation or the turnover of other compounds e.g. the DMSO reductase family (Leimkühler & Iobbi-Nivol, 2016) or quinone oxidoreductase complexes (Duarte et al., 2021). For these reasons, the abundance of microbes utilising organic sulfur compounds is likely to be underestimated (Carrion et al., 2019) and the role of sulfonated compounds is understudied (Wolf et al., 2022). Thus, there is a knowledge gap of the link between inorganic and organic sulfur cycling in ecological systems.

To fill this gap, we have extended HMS-S-S (Tanabe & Dahl, 2022). This tool was originally developed for rapid detection and annotation of inorganic sulfur dissimilation in prokaryotic genomes. With the substantial extension presented here, it now includes not only inorganic sulfur metabolism enzymes, but also enzymes with characterized or at least strongly indicated function in the metabolism of sulfonated sulfur compounds. These include sulfoquinovose synthesis and degradation pathways, DMSP metabolism, taurine and isethionate conversion, and transport systems for various sulfonated compounds. For all these pathways, we developed individual profiled hidden Markov Models (HMM) and validated score thresholds by cross-validation and with an independent test dataset. HMS-S-S itself has been completely redesigned, improving usability and output formats, and extending the file manipulation tool. By optimising the underlying algorithms, the overall computing speed has been increased by a factor of four. Due to the complete overhaul, we have renamed the tool “HMSS2”. HMSS2 now covers the known metabolism of inorganic and organic sulfur compounds, facilitating the exploration of the microbe-driven natural sulfur cycle.

2 METHODS

2.1 HMSS2 improvements and workflow

Algorithmic improvements were made on the speed and user-friendliness by process optimization and the implementation of additional features. HMSS2 algorithms are now completely written in Python and precompiled versions are available. In this way, the number of dependencies required to be installed by the user has been greatly reduced to just two external programs. HMMER and Prodigal are still required but installing and configuring of MySQL is no longer necessary. The installation was further simplified by preparation of a pre-compiled executable, that will run directly on a Unix system.

HMSS2 includes the basic design of HMS-S-S with further automation. User-supplied input requires a directory containing files in FASTA nucleotide format, consisting of scaffolds or contigs. Alternatively, it is possible to provide amino acid sequences in FASTA files and the corresponding features in GFF3 formatted files. All files in the directory will then be processed in consecutive order. Nucleotide input files are first searched for open-reading frames and translated into protein sequences by Prodigal. This step is omitted if protein sequences are provided. Profile hidden Markov Models (HMM) are then queried against the protein sequences of the current file with validated bit score cutoffs via hmmsearch. Hits are saved in a local database together with corresponding genomic features and protein amino acid sequences. The local database now uses the SQLite database engine and an improved database table structure that allows to save multi-domain proteins with all domains. In the next step, the detected proteins are searched for genetic co-localization. This is done via the genomic features and a maximum nucleotide distance between two genes to be syntenic. Syntenic gene clusters are then compared with a set of predefined and named gene patterns. A new feature of HMSS2 is the detection of co-linear gene clusters. This is a special type of synteny where the genes occur in exactly the same order as the gene pattern. Gene clusters that are similar to the pattern(s) provided are then named by characteristic keywords. NCBI, GTDB taxonomy files or custom files with a similar format can be used to assign taxonomic information. As the taxonomy may change over time, it is recommended that the user updates this information locally as required. Results can be retrieved from the local database filtered by protein domains and/or keywords via HMSS2. The standard output now includes FASTA formatted files and iTol datasets.

2.1 Training dataset generation, annotation and HMM development

Datasets were generated from genomic data downloaded from NCBI RefSeq (Haft et al., 2018) or GenBank (Sayers et al., 2019) as of September 2022. The HMM training dataset contained all assemblies from the NCBI RefSeq database with an assembly level of a complete chromosome. The independent test data consisted of assemblies originating from GenBank, again with an assembly level of the complete chromosome. GenBank covers a greater number of phyla and a wider range of quality and is therefore not entirely similar to the training data from RefSeq. Sequence annotation for Hidden-Markov-model generation was performed using the training dataset and list of reference proteins for organic sulfur metabolism (Table S1). Methods for annotating the training and independent test datasets and for HMM generation were used as described previously (Tanabe & Dahl, 2022).

2.4 Performance metric calculation

Performance was determined using balanced accuracy (Brodersen et al., 2010), F1-score (Forman & Scholz, 2010), and the Matthew-correlation-coefficient (MCC) (Chicco & Jurman, 2020). The metric values were additionally corrected for the dataset's skewness (Jeni et al., 2013) (Table S2). Values for each Hidden Markov Model were calculated from a confusion matrix obtained by comparing the annotation of the training/test dataset and annotation assigned by the HMMs. Matching assignments were considered as true positives (TP), while mismatching assignments were considered as false positives (FP), if the HMM recognised a sequence unrelated to the HMM training sequences. All sequences that were not recognized by the HMM but matched the annotation were counted as false negative (FN), and all other sequences were recorded as true negatives (TN).

2.5 Thresholding and cross-validation

Thresholding and cross-validation were executed as previously described (Tanabe & Dahl, 2022). For each HMM, bit scores for noise cutoff, trusted cutoff, and an optimized threshold were determined prior to cross-validation. The noise cutoff corresponded to the score of the lowest scoring TP hit. The trusted cutoff corresponded to the score of the highest scoring FP hit. The optimized cutoff was computed during a nested cross-validation procedure with a 10-fold outer loop and a 5-fold inner loop (Varma & Simon, 2006). The optimized cutoff

corresponded to the median of the thresholds with the highest F1 scores across all inner folds. Outer folds were analyzed after all thresholds were set.

Each cross-validation fold was generated from the HMM training data. Sequences were randomly sorted into the 10 outer folds of equal size, followed by the equal deviation of each outer fold into 5 inner folds. A cross-validation procedure was then performed on all folds. The inner folds were used to determine the optimized thresholds. The overall performance of each HMM was then done with a confusion matrix created for the outer folds using the optimized thresholds as a cutoff. Balanced accuracy was calculated as the average of all accuracies from each fold. F1 score and MCC were calculated as the sum of the confusion matrices from all folds (Forman & Scholz, 2010). The same procedure without fold generation was performed for the independent test dataset (Chicco, 2017).

2.6 Performance testing

The performance of HMSS2 was compared with that of HMS-S-S version 1 (Tanabe & Dahl, 2022). The HMM library included all 164 HMMs of the original library, detecting dissimilatory sulfur metabolism. A quadratic increasing number of randomly selected genomes ranging from 2 to 64 were chosen from the training dataset described for version 1 and used as input for the performance comparison. The input data were in FASTA nucleotide format. Each run was repeated three times with newly randomised input data to reduce performance bias caused by the input data. Both program versions were benchmarked for the execution time required for the workflow from data ientry to the final annotated hits with appropriately named gene clusters, but without taxonomy assignment. Time was measured as the required wall-clock runtime when running HMS-S-S or HMSS2 with four parallel threads on an Intel Core i7-6700 CPU.

3 RESULTS

Here, we created a comprehensive database of reliable hidden Markov models (HMMs) based on archaeal and bacterial proteins associated with organic sulfur metabolism. The same approach has already been used for the enzymes of dissimilatory metabolism of inorganic sulfur compounds (Tanabe & Dahl, 2022). Not only sequence similarity, but also integrated synteny was considered to assign a protein to a specific functional group. The HMMs created

here focus on the most abundant organic sulfur compounds in terrestrial and aquatic environments. The compounds covered here include dimethylsulfoniopropionate (DMSP), dimethyl sulfide (DMS), dimethyl sulfoxide (DMSO), dimethyl sulfone ($DMSO_2$) (Fig. 1), 2,3-dihydroxypropane-1-sulfonate (DHPS), isethionate, taurine, and membrane sulfolipids (Fig. 2). The HMMs for the enzymes of the metabolic pathways for degradation of individual compounds are described in full below. Normally, prokaryotes do not code for the entire degradation pathways, but only for parts of them.

3.1 HMM Development: DMSP degradation

DMSP is primarily produced by single-celled phytoplankton and algal seaweeds, where it acts as an osmolyte and anti-stress molecule (Kiene et al., 2000). Degradation of DMSP either requires a demethylation pathway or a DMSP lyase (Fig. 1). The demethylation pathway is encoded by the *dmdABCD* gene cluster and starts with the demethylation of DMSP via DmdA to form methylmercaptopropionate. This intermediate is further catabolized by DmdB, DmdC and finally DmdD with the release of acetaldehyde and methanethiol (Bullock et al., 2014; Reisch et al., 2011). For each of the enzymes, one HMM was generated, making four in total. Several non-orthologous DMSP lyases, DddL, DddP, DddQ, DddW and DddY, have been characterised which convert DMSP to acrylate with the release of DMS and acrylate. The latter is then converted to 3-hydroxypropionate by AcuNK (Curson et al., 2011) or to propionyl-CoA by AcuI (Todd et al., 2012). DMSP lyase DddD catalyzes formation of propionyl-CoA and DMS from DMSP in a single reaction without the formation of an acrylate intermediate. 3-hydroxypropionate can be further converted to acetyl-CoA via DddA and DddC (Curson et al., 2011). HMMs were generated for AcuI, AcuN, AcuK, DddA, and all DMSP lyases. As there were less than ten sequences identified for DddQ, DddW and DddC, HMMs could not be constructed for these three enzymes.

3.2 HMM development: Assimilation of methanethiol and DMS

DMS and methanethiol are C_1-organosulfur compounds derived mainly from the degradation of DMSP. Both can be assimilated by bacteria as a source of sulfur and carbon, where methanethiol is first converted to DMS, followed by oxidation and assimilation (Fig. 1). The conversion of methanethiol to DMS is catalyzed by methanethiol S-methyltransferase, MddA. This membrane-bound enzyme transfers a single sulfur atom from S-

adenosylmethionine to methanethiol (Carrion et al., 2015). The resulting DMS can be further oxidized by either dimethylsulfide cytochrome *c* reductase, DdhABCD, also known as dimethylsulfide dehydrogenase (McDevitt, Hanson, et al., 2002), or by multicomponent DMS monooxygenase DsoABCDEF (Horinouchi et al., 1999). The periplasmic DdhABC dimethylsulfide dehydrogenase couples the oxidation of DMS to the reduction of two *c*-type cytochromes, producing DMSO as the final product. DdhD is a cytoplasmic protein that is not part of the DMS dehydrogenase but has a proposed function in the assembly of the DdhAB complex and its secretion via the Tat pathway (McDevitt, Hugenholtz, et al., 2002). For DdhA and DdhB, it was possible to generate individual HMMs, while this was not the case for DdhC and DdhD which had less than ten validly annotated sequences in the training dataset. The multicomponent DMS monooxygenase DsoABCDEF oxidizes DMS in a two-step reaction to $DMSO_2$ with DMSO as intermediate. As the sulfur moiety is specifically oxidised, this enzyme is also referred to in the literature as assimilatory DMS S-monooxygenase (Boden & Hutt, 2019). A total of six HMMs were generated for this complex. After the oxidation of DMS to $DMSO_2$, the next step in sulfur assimilation is the oxygen-dependent conversion of $DMSO_2$ to methanesulfinate, catalyzed by FMN-dependent $DMSO_2$ monooxygenase SnfG (Wicht, 2016). SnfG was represented by a single HMM. Methanesulfinate is chemically oxidized to methanesulfonate, which is further oxidized to sulfite and formaldehyde by the assimilatory methanesulfonate monooxygenase MsuDE in a NADH- and oxygen-dependent reaction. For MsuDE, a HMM was trained for each subunit.

3.3 HMM development: Dissimilation of $DMSO_2$

Dimethylsulfone is mainly derived from oxidation of DMS. The degradation of dimethyl sulfone ($DMSO_2$) begins with its reduction to dimethyl sulfoxide (DMSO) by a $DMSO_2$ reductase in an NADH-dependent reaction (Fig. 1). Although the activity has been measured in crude extracts of some methylotrophic Actinobacteria and Alphaprotebacteria (Borodina et al., 2000; Borodina et al., 2002), the enzyme has not been characterized. DMSO is then further reduced to dimethylsulfide (DMS). Two types of DMSO reductases have so far been characterized (Boden & Hutt, 2019). The first, membrane-bound enzyme is composed of the three subunits, DmsABC, and uses electrons from the quinol pool for DMSO reduction (Bilous & Weiner, 1985). For this enzyme one HMM for each subunit was trained. The second DMSO reductase uses NADH for this purpose and probably consists of only one subunit with high

similarity to DmsA, indicated by its cross-reaction with DmsA antibodies. A separate HMM could not be trained for this enzyme, because it is only known by its activity in crude extracts (Borodina et al., 2002). In addition to the Dms-type DMSO reductases, a soluble periplasmic DMSO reductase, DorCAD, has been characterized (A. G. McEwan et al., 1998). The corresponding genes are regulated by DorS and DorR (Kappler & Schäfer, 2014). For each of these five proteins/subunits, we constructed one HMM. The DMS, which is released by DMSO reductase of both types, is oxidized to methanethiol (CH_3SH) and formaldehyde by a DMS monooxygenase, DmoAB, in another NADH-consuming reaction (Boden et al., 2011). As only *dmoA* has been validly identified so far, we trained a HMM specifically for DmoA, but not for DmoB. Further oxidation of methanethiol by a methanethiol oxidase MtoX leads to the final release of sulfide and another molecule of formaldehyde (Eyice et al., 2017). A single HMM was trained for MtoX.

3.4 HMM development: Dissimilation of methanesulfonate

Methanesulfonate is formed by spontaneous chemical oxidation of DMS in the atmosphere (Fig. 1). It is used by diverse aerobic bacteria as a sulfur source and by some specialized methylotrophic prokaryotes as a source of carbon and energy (Kelly & Murrell, 1999). The dissimilatory methanesulfonate monooxygenase catalyzes the conversion of methanesulfonate to formaldehyde and sulfite (Henriques & De Marco, 2015). This enzyme is encoded by the *msmABCD* operon, which is often located adjacent to the *msmEFGH* operon, usually in the opposite direction. The latter encodes a putative ABC-type transporter (Fig. 2b) proposed to facilitate the import of methanesulfonate into to the cytoplasm (Henriques & De Marco, 2015). Six HMMs were developed to represent each of these proteins. MsmC and MsmD had to be excluded due to the small number of sequences in the training datasets.

3.5 HMM development: Alkanesulfonate oxidation and transporters

The *ssuEADCB* gene cluster encodes the two-component alkanesulfonate monooxygenase SsuDE and the alkanesulfonate ABC-transporter SsuABC (Fig. 2b). Alkanesulfonate monooxygenase catalyzes the oxidation of various sulfonated alkanes as substrates with variable affinity, including phenylated organic compounds like N-phenyltaurine. After transport into the cell via SsuABC, the sulfonate is cleaved by SsuDE in a reaction dependent on NADH and molecular oxygen (Eichhorn et al., 1999). Electrons are

provided by SsuE via an FMN cofactor. SsuD then cleaves the sulfonate group and oxidizes the terminal carbon atom. For this pathway five HMMs, one for each encoded protein, were created.

3.6 HMM development: Sulfoquinovose synthesis

Sulfoquinovose (SQ) is a sulfonated derivate of glucose where the 6-hydroxyl group is substituted by a sulfonate group. SQ is a constituent of the unique head group of the membrane-bound glycolipid sulfoquinovosyl diacylglycerol (SQDG) present in thylakoid membranes and photosynthetic prokaryotes. On a genetic level, five genes *sqdA, sqdB, sqdC, sqdD* and *sqdX* have been described to be involved in SQDG synthesis in bacteria so far (Benning & Somerville, 1992a, 1992b; Guler et al., 2000; Rossak et al., 1995). The functions of SqdA and SqdC have not been completely resolved (Benning & Somerville, 1992b; Rossak et al., 1997). The synthesis begins with the exchange of the 6-hydroxyl group of uridine-diphosphate (UDP)-glucose for a sulfonate group by UDP-sulfoquinovose synthase, SqdB. The formation of SQDG is then catalyzed SQDG synthase, SqdD or SqdX (Rossak et al., 1995). A total of five HMMs was trained to detect the enzymes of this pathway.

3.7 HMM development: Sulfoquinovose degradation and transport

As sulfoquinovose is a sulfonated derivate of glucose, it is catabolized in a similar manner and can serve as a carbon and energy source (Hanson et al., 2021). Several pathways resembling glucose degradation have been characterized, including the Sulfo-Embden-Meyerhof-Parnas pathway (Denger et al., 2014), the Sulfo-Entner–Doudoroff pathway (Felux et al., 2015), the transaldolase-based pathway related to the pentose phosphate pathway (Frommeyer et al., 2020) and a complete degradation pathway based on a sulfoquinovose monooxygenase (Sharma et al., 2022) (Fig. 2a).

The Sulfo-Embden-Meyerhof-Parnas pathway (Fig. 2a) begins with import of sulfoquinovose by the transporter YihO. A sulfolipid α-glucosidase YihQ may also be involved and other SQ derivatives may also be imported. Analogous to the EMP pathway, SQ is then cleaved to dihydroxyacteonephosphate (DHAP) and 3-sulfolactaldehyde (SLA) via the isomerase YihS, kinase YihV and aldolase YihT. In an NADH-dependent reaction, the reductase YihU then reduces SLA to the final product 2,3-dihydroxypropane sulfonate (DHPS), which is

transported out of the cell again via YihP. A separate HMM was created for each of the Yih proteins.

The Sulfo-Entner–Doudoroff is analogous to the ED pathway (Fig. 2a). As there was no specific abbreviated name assigned to these enzymes by the original publication (Felux et al., 2015), we assigned names to enhance HMSS2 output readability. SQ is cleaved by a dehydrogenase SedA, a lactonase SedB, a dehydratase SedC and an aldolase SedD to pyruvate and SLA. Another dehydrogenase, SedE, then oxidizes 3-sulfolactaldehyde (SLA) in an NAD-dependent reaction to 3-sulfolactate (SL), which is then exported. A separate HMM was generated for each of the proteins mentioned, for a total of 5 HMMs.

The third SQ degradation pathway contains a transaldolase as the key enzyme (Fig. 2a) (Frommeyer et al., 2020). SQ is imported into this pathway via the transporter SftA and converted to sulfofructose by the isomerase SftI. This product, together with glycerine-aldehyde-3-phosphate, is then converted by the transaldolase SftT to SLA and fructose-6-phosphate. SLA, in turn, is converted to SL in an NAD-dependent reaction by the dehydrogenase SftD and exported via the transporter SftE or reduced to 2,3-dihydroxypropane sulfonate (DHPS) in an NADH-dependent reaction by the reductase SftR. A separate HMM was generated for each of the Sft proteins, for a total of 6 HMMs.

The fourth known degradation pathway for SQ (Fig. 2a) differs from the others described so far, because it involves oxidation of the entire molecule, including cleavage of sulfur (Sharma et al., 2022). The pathway described begins with the import of sulfoquinovosylglycerol by an ABC transporter called SmoEFGH. In the cytoplasm, sulfoquinovosyl glycerol is cleaved by the sulfoquinovosidase SmoI to SQ. In contrast to the other pathways, SQ is now transformed to 6-oxo-glucose and sulfite by an alkanesulfonate monooxygenase, SmoC. The electrons for this reaction come from NADPH via the flavin reductase SmoA. 6-oxo-glucose is converted in another NADPH-dependent reaction by SmoB into glucose, which is then available for glycolysis. Eight HMMs were generated for this pathway, one for each protein. An additional HMM was trained for SmoD ,a putative regulator encoded in the *smo* operon.

3.8 HMM development: 2,3-dihydroxypropane sulfonate transporters and degradation

According to the postulated pathway for degradation of 2,3-dhydroxy propane sulfonate (DHPS) (Fig. 2c), the compound is either taken up by the TRAP transporter HpsKLM or by HpsU (Fig. 2b). The DHPS-3-dehydrogenase HpsN then converts (R)-DHPS to sulfolactate with concomitant formation of two equivalents of NADH. For (S)-DHPS, it was postulated that this compound is first converted to the (R)-DHPS enantiomer via (R)-DHPS-2-dehydrogenase HpsP and (S)-DHPS-2-dehydrogenase HpsO (Mayer et al., 2010). The resulting (R)-sulfolactate can be further converted in several ways: The (R)-sulfolactate sulfolyase SuyAB catalyzes a desulfonation reaction, releasing sulfite and pyruvate. The (S)-enantiomer of sulfolactate is first converted to sulfopyruvate by SlcC and then to (R)-sulfolactate by ComC (Mayer et al., 2010). Both enantiomers were postulated to be transported by the exporter SlcHFG (Mayer et al., 2010) (Fig. 2b). On HMM was created for each protein/subunit of the DHPS degradation pathway.

3.9 HMM development: Isethionate and taurine degradation

Isethionate and taurine are C_2-sulfonates which are produced by eukaryotes from cysteine or methionine (Moran & Durham, 2019). Bacterial degradation of these compounds includes sulfoacetaldehyde as an intermediate which is a point of convergence with sulfoacetate degradation (Weinitschke, Hollemeyer, et al., 2010) (Fig. 2c). Two different transporters are proposed for the import of isethionate (Fig. 2b). These are the TRAP transporters IseKLM and IseU from the major facilitator superfamily. After import into the cytoplasm, isethionate is oxidized to sulfoacetaldehyde by the isethionate dehydrogenase IseJ (Weinitschke, Sharma, et al., 2010). In some organisms, isethionate is not converted, but the sulfonate group is cleaved off by isethionate sulfite lyase IslAB, releasing sulfite and acetaldehyde (Peck et al., 2019).

Taurine import is postulated to be facilitated by the ABC transporter TauAB1B2C or the TRAP transporter TauKLM (Fig.2b). There are several possibilities for the further pathway. Taurine can either be oxygenated by TauD to form 1-hydroxy-2-aminoethane sulfonic acid, which decomposes to aminoacetaldehyde and sulfite (Eichhorn et al., 1999), or it is oxidized in NADH-dependent reaction by the taurine dehydrogenase TauXY, which produces

sulfoacetaldehyde. The same product is also produced by the transfer of the amino group to pyruvate by taurine:pyruvate aminotransferase Tpa (Bruggemann et al., 2004)) or to 2-oxoglutarate by taurine:2-oxoglutarate aminotransferase Toa (Krejcik et al., 2010).

Sulfoacetaldehyde can be converted by the NADPH-dependent sulfoacetaldehyde reductase IsfD to isethionate which is then exported by the IsfE transporter (Krejcik et al., 2010). Another possible fate of sulfoacetyladehyde is desulfonation coupled to a phophorylation by sulfoacetaldehyde acetyltransferase Xsc to acetyl phosphate which is further converted to acetyl-CoA by phosphate acetyltransferase Pta (Weinitschke, Sharma, et al., 2010) Sulfite released in the each of these processes is exported via TauE (Weinitschke et al., 2007). An individual HMM was developed for each individual protein/subunit mentioned here. An exception was made for TauB1 and TauB2, which were combined into a single HMM due to their similarity. Additionally, we trained an HMM for TauZ, a protein of unknown function, and the regulator TauR. Both are commonly found genetically associated with other *tau* genes.

3.10 HMM development: Sulfoacetaldehyde formation

Sulfoacetaldehyde is not only produced by taurine and isethionate degradation but also by the dissimilation of sulfoacetate (Weinitschke, Hollemeyer, et al., 2010). The transporter SauU is hypothesised to facilitate the entry of sulfoacetate into the cell (Fig. 2b). Subsequently, sulfoacetate is activated by sulfoacetate-CoA ligase, SauT, and finally reduced to sulfoacetaldehyde via sulfoacetaldehyde dehydrogenase, SauS, consuming NADPH. SauS, SauT and SauU (Weinitschke, Hollemeyer, et al., 2010) were each represented by a HMM respectively. Sulfoacetaldehyde can also be produced by decarboxylation of sulfopyruvate (Fig. 2c) catalyzed by ComDE (Denger et al., 2009). These two subunits are each represented by a HMM.

3.11 HMM development: Cysteine synthesis

Cysteine is an essential amino acid with a thiol side chain. Here, we started to cover the relevant proteins with HMMs primarily based on knowledge collected with enterobacterial model organisms. Biosynthesis begins with the import of sulfate or thiosulfate into the bacterial cell via CysUWA (Aguilar-Barajas et al., 2011) or YeeE/YedE-like (Tanaka et al., 2020) transporters. Sulfate is reduced to sulfide which is then incorporated into O-acetylserine to

synthesize cysteine (Kredich, 1996). In *E. coli*, sulfate is activated by ATP sulfurylase CysDN (Leyh et al., 1988) to adenosine 5'-phosphosulfate (APS), which can be further activated by APS kinase CysC to 3'-phosphoadenosine-5'-phosphosulfate (PAPS). PAPS reductase CysH then reduces the activated compound to sulfite. In some bacteria, including most cyanobacteria, APS can be reduced to sulfite directly, without phosphorylation to PAPS (Bick et al., 2000). The assimilatory APS reductases catalyzing this reaction exhibit similarity to the assimilatory PAPS reductases (Abola et al., 1999; Bick et al., 2000) and are covered by the same HMM (CysH) in this work. In Enterobacteria, sulfite is reduced to sulfide via CysIJ. Finally, cysteine is synthesized from sulfide and O-acetyl-L-serine by the cysteine synthase CysK. A total of 10 new HMMs was generated for the mentioned proteins/subunits. An HMM for YeeE/YedE-like transporters was already available through HMS-S-S (Tanabe & Dahl, 2022)

3.12 HMM validation: cross validation and independent test data set

The HMMs developed were validated by cross-validation and with an independent test data set. In cross-validation, sequences unrelated to the tested HMM training data were added as true negative examples in addition to the omitted training sequences (Chicco, 2017; Refaeilzadeh et al., 2009). The omitted sequences from each fold served as true positive examples. Cross-validation was performed using the optimized thresholds calculated prior to cross-validation. Thus, the threshold values should also be checked for their suitability. Performance was measured using the Matthews Correlation Coefficient (MCC). This metric ranges from -1 to 1, with 0 corresponding to random assignment, 1 corresponding to perfect assignment with no misclassification, and -1 corresponding to complete misclassification. Here, the individual occurrence of FP or FN lowers the score on the MCC, while the combination of both misclassifications lowers the score more dramatically than the single occurrence of either type of error (Chicco & Jurman, 2020).

The majority of the HMMs developed showed high precision and recall in the cross-validation and on the test dataset (Fig. 3). Of the 134 HMMs covering proteins of organic sulfur compound metabolism, 127 stayed above an MCC of 0.80 during the cross-validation (Fig. 3, Table S2). The evaluation of the 134 HMMs against the independent test dataset resulted in 120 HMMs with an MCC of 0.80 or higher. HMMs for the alkanesulfonate transporter subunits SsuB and SsuC failed the cross-validation threshold of 0.8 slightly by 0.02 points but performed better on the independent test dataset. These were the only cases where the cross-validation

performance was insufficient but the performance on the test dataset was above the threshold. From the HMMs with an MCC > 0.8 during cross-validation, seven scored below 0.8 in the test dataset. These were MsmG with an MCC of 0.78, SmoI (0.76), MsmB (0.66), DddA (0.62), DorA (0.46) and SftD (0.03). For SftD, MsmB, MsmG and DddA this was due to a high number of sequences which were falsely classified as negative, probably due to a low training sequence diversity. Thus, these HMMs had a high precision and did not generate high numbers of false positive hits, but they performed low in recognition resulting in a high number of unrecognized sequences. The opposite was the case for the DorA HMM, which generated too many false positive hits but no false negative ones. Sulfoquinovosidase SmoI interfered in the detection with sulfoquinovosidase named YihQ. The same holds true for transporters HpsU and IseU. All sequences that were falsely classified by one of these two HMMs belonged to the other HMM. Together these two HMMs performed well in detecting of isethionate and DHPS transporters of the major facilitator superfamily. The situation was similar for YihO and SftA which are both postulated sulfoquinovose importers that catalyse the same function in the context of sulfoquinovose degradation. In summary, 112 of 134 HMMs were successfully tested via cross-validation and with an independent dataset. Two other pairs of HMMS can be used together, for the safe detection of sulfoquinovosidase and the transporters YihO and SftA.

3.13 HMM validation: Case study

HMSS2 was also validated with 24 complete genomes from bacteria with organic sulfur compound metabolism (Table S3), which were screened for the presence of enzymes for the utilisation of taurine, isethionate, DHPS, sulfoquinovose and DMS (Fig. 4).

Proteins for taurine utilization were found mainly in the known taurine-utilizing genera *Octadecabacter*, *Roseobacter*, *Roseovarius* and *Ruegeria* of the Rosebacterales, including the taurine degraders *Roseovarius nubinhibens* (Denger et al., 2009) and *Ruegeria pomeroyi* (Gorzynska et al., 2006). These strains encoded for the TauABC taurine importer, Tpa and Xsc constituting the complete degradation pathway from free taurine via sulfoacetaldehyde to acetyl phosphate with the release of sulfite. *Roseobacter denitrificans* additionally possessed genes for the taurine dehydrogenase TauXY and the taurine:2-oxoglutarate aminotransferase Toa, which can also convert taurine to sulfoacetaldehyde. The sulfoacetaldehyde acetyltransferase Xsc was present in all genomes examined. This is probably due to the fact

that sulfoacetaldehyde is not exclusively an intermediate of taurine degradation but also of isethionate, sulfoacetate and DHPS degradation, and possibly of other as yet unknown pathways (Weinitschke, Hollemeyer, et al., 2010). In line with this possibility, genes encoding isethionate dehydrogenase IseJ, which converts isethionate to sulfoacetaldehyde, were found in almost all analyzed Rhodobacterales, Hyphomicrobiales and Gammaproteobacteria genomes, consistent with earlier reports (Weinitschke, Sharma, et al., 2010). *Leminorella grimontii*, *Hyphomicrobium denitrificans* and all *Methylophaga* species were exceptions, consistent with the inability of *H. denitrificans* and *Methylophaga* to consume organosulfur compounds with more than one carbon atom.

Isethionate desulfonation via isethionate sulfite-lyase IslAB has been found in microcompartments of *Bilophila wadsworthia* (Burrichter et al., 2021). In accordance, HMSS2 detected the importer IseU and IslAB in this organism. A similar desulfonation pathway without microcompartments was postulated for *Desulfovibrio alaskensis* and *D. desulfuricans* (Burrichter et al., 2021). In *D. desulfuricans*, HMSS2 also found IseU and IslAB, suggesting that this organisms, like *B. wadsworthia*, may scavenge free isethionate via IseU. In contrast, *D. alaskensis* encodes IslAB but not IseU. Instead, it contains sulfocacetaldehyde reductase IsfD (or SarD), which is also present in *Bilophila wadsworthia*. In both cases, this enzyme may provide an endogenous source of isethionate (Burrichter et al., 2021).

Most analysed genomes possessed the potential for sulfopyruvate and (R)-sulfolactate generation from DHPS and (L)-sulfolactate. The potential of (R)-DHPS oxidation via HpsN generating 2 NADH equivalents was found in all analysed strains and most lso encoded for isomerization of (S)-DHPS to (R)-DHPS via HpsP (17/24 genomes). The predicted presence of genes for desulfonation of sulfopyruvate by ComDE and sulfolactate by SuyAB as found here is also in accordance with previous reports for the Roseobacterales clade (Chen et al., 2021; Denger et al., 2009), the Hyphomicrobiales (Chen et al., 2021), *Desulfovibrio desulfuricans* and *B. wadsworthia* (Hanson et al., 2021). Even without the ability to desulfonate sulfopyruvate or sulfolactate, the conversion of DHPS to sulfopyruvate or sulfolactate and export of these as end products provides 2-3 NADH equivalents and thus a growth advantage for the organism.

Sulfoquinovose degradation via the Sulfo-Entner-Doudoroff pathway is present in eight bacteria, including *Pseudomonas putida* and other bacteria for which this pathway has been described or postulated (Felux et al., 2015). The complete sulfoquinovose degradation

pathway based on a sulfoquinovose monooxygenase was found in seven proteobacteria in accordance with previous reports (Sharma et al., 2022). The other known sulfoquinovose degradation pathways were not detected, which is likely due to the presence of the Sulfo-Embden-Meyerhof-Parnas pathway (Denger et al., 2014) primarly in Enterobacterales and the transaldolase-dependent sulfoquinovose degradation in Firmicutes (Frommeyer et al., 2020). Bacteria from these taxonomic groups were not included in the case study.

DMS degradation has been described for *Methylophaga thiooxydans*, *Methylophaga sulfidovorans* (Kröber & Schäfer, 2019), *Hyphomicrobium denitrificans* (Koch & Dahl, 2018), and *Hyphomicrobium sulfonivorans* (Boden et al., 2011). According to our HMSS2 analysis, *H. sulfonivorans* encoded for DmoA, while all other three encoded only for methanethiol oxidase MtoX. DmoA was missing and the organisms must contain a so far unknown DMS monooxygenase. In accordance with previous reports, MtoX was also found in *Methylacidiphilum fumariolicum* (Schmitz et al., 2022), and several Rosebacterales, including *Ruegeria pomeroyi* (Eyice et al., 2017). The latter is a known degrader of DMSP to methanethiol via DmdA, B, C and DmdD (Reisch et al., 2011) which were all detected by the HMMs created here.

In summary, our case study on characterized organosulfur compound degraders has shown that in all cases the detection by HMSS2 agrees with the published analyses of other authors.

3.14 HMSS2 improvements

HMSS2 has a redesigned engine and additional features for protein annotation and output format customisation (Fig. 5). Proteins with multiple domains are now stored with all domains and not just the domain with the highest score. This was accomplished by improving the local relational database structure. This requires that the recognised domain regions in the primary sequence do not overlap, so that domains with high scores are not overwritten by lower scores. On the other hand, high-scoring domains may still overwrite one or more lower-scoring domains during annotation.

Gene arrangement can now be used by HMSS2 for annotation as a non-homologous criterion. Hits below the threshold are also considered and annotated if they lie within a gene cluster and the potentially assigned annotation would complete a known gene cluster

arrangement. Thus, a gene that highly likely occurs within a gene cluster must reach a lower cutoff than normal to be detected if it is encoded within such a cluster.

The output formats have been greatly expanded, and new features were added to improve usability and readability. It is still possible to retrieve sequences filtered by protein type, the genomic proximity and the presence of proteins or gene clusters in the same genome. HMSS2 automatically recovers a list of all hits with genomic features and a separate protein sequence file in FASTA format. Additionally, two subsets of the latter file are created. One subset includes all hits that are unique to their genome respectively, while another subset includes all hits that occur at least twice in the same genome. Multi-domains proteins, retrieved by the requested protein type, are listed separately if at least one other domain has been annotated.

An output module for iTol compatible datasets was also included. This module integrates the generation of iTol datasets for presence/absence of the keywords/domains for each genome. Range datasets, which mark specific proteins in a phylogenetic tree, can now also be generated by HMSS2, as well as iTol compatible datasets for displaying gene clusters. HMSS2 also comes with several utilities to modify the output protein FASTA files. It is now possible to assign the taxonomic names of the source organism to each sequence. Files can now be filtered by length, merged without duplicating sequence identifiers and sequences from multiple FASTA files originating from the same organism can be concatenated into a single sequence. With a FASTA-formatted file as input, a list of neighboring genes is now accessible to support searches for conserved but previously undiscovered gene constellations.

The execution time of the HMSS2 was compared to that of HMS-S-S to demonstrate the scalability and efficiency of HMSS2. For this test, increasing numbers of genomes were randomly selected from the assemblies of the training dataset and gene clusters were annotated and determined with the 164 HMMs of the original library. Time measurements were performed in triplicate with random selection of input assemblies for each replicate. The execution time was then averaged over all replicates. Comparison between the two versions showed a large difference in the required execution time (Fig. 6, Table S4). The observed increase in execution speed for HMSS2 became more significant as the number of genomes processed increased and scaled linearly with the number of input assemblies. While HMS-S-S required around 26 minutes to process 64 assemblies, HMSS2 needed only 7 minutes for this

task. Thus, the introduced improvements led to a fourfold accelerated computation speed for HMSS2.

4 DISCUSSION

Here, we present a substantial update that provides an HMM-based search tool for proteins involved in the metabolism of inorganic and organic sulfur compounds. The high accuracy of the advanced tool presented here provides a reliable basis for genome analysis and is further supported by the genomic context detection. The HMSS2 algorithm now uses homologous and non-homologous criteria already in the protein annotation step, not just for the later identification of gene clusters. In addition, the overall execution time was accelerated by fourfold compared to the previous version, further speeding up the detection of sulfur metabolism pathways in genomes and metagenomes. With the increasing number of available genomes, faster protein annotation is required to handle the immense amount of available data.

We also significantly broadened the applicability of HMSS2 by adding the conversion of sulfonated carbon compounds. HMSS2 now covers pathways from the entire sulfur cycle, enabling studies on the link between the cycles of inorganic and organic sulfur compounds. In addition to providing operon structure information to support equivalence prediction, the accessibility and display of the annotated proteins has been greatly enhanced. Not only can sequences now be filtered by annotation, but also the presence of genes and genomic context can be displayed using other specialised applications, further extending the capabilities of synteny analysis. Such analyses are not limited to studies of the ecological role of prokaryotes but also include the evolution of metabolic pathways (Garcia et al., 2022), distribution of new pathways (Sharma et al., 2022) and genomic context visualization (Garcia et al., 2019; Letunic & Bork, 2021).

The expansion to the metabolism of organic sulfur compounds resulted in the generation of 134 additional HMMs in addition to the 164 HMMs previously included in HMS-S-S, almost doubling the total number of proteins included. The accuracy of the newly generated HMMs and the respective thresholds were demonstrated by cross-validation and a test dataset. Observed deviations between both testing methods are likely due to an uneven distribution and abundance of protein sequences influencing the number and diversity of

testable sequences. The quality of the 134 novel HMMs was ensured by selection of high-quality genomes derived from the RefSeq and GenBank databases. The overall development process had already been successfully applied for the proteins of inorganic sulfur metabolism (Tanabe & Dahl, 2022). The test dataset was obtained from the full diversity of phyla accessible from GenBank and should therefore reflect the widest possible range of sequence variation. However, although the cutoff values have been validated, they are likely to need adjustment for newly discovered phyla (Anantharaman et al., 2018; Jaffe et al., 2020).

The diversity of proteins involved in the metabolism of organic sulfur compounds covered by HMSS2 also includes less prominent pathways for degradation and conversion of compounds such as sulfoquinovose or DMS. Although a considerable proportion of sulfur in the biosphere is bound in substrates or intermediates of these pathways, they are not commonly included in annotation pipelines and often unrecognized or incorrectly annotated. This is illustrated by fact that only 16 of the 124 proteins included here for the conversion of sulfoquinovose, taurine, isethionate or DMSP have an exact counterpart in PFAM (El-Gebali et al., 2019) or TIGRFAMs. In contrast, eight of ten HMMs covering sulfate assimilation for cysteine biosynthesis have a TIGRFAM equivalent. A common problem in the functional annotation of enzymes involved in metabolism of organic sulfur compounds are enzymes, such as DmsA or DorA, that belong to the DMSO reductase superfamily. This family includes tetrathionate reductase, polysulfide reductase and thiosulfate reductase, as well as several other proteins unrelated to sulfur metabolism. Tertiary structure and complex composition is conserved throughout all members of this family (Alastair G. McEwan et al., 2010) and substrate specificity may only arise through a small number of conserved amino acids at the active site (Struwe et al., 2021). The validation performed here showed that related complexes in the DMSO reductase family did not negatively affect the HMMs for DmsA and DorA. Furthermore, the reliability of prediction is raised when genomic context is paired with the prediction made by the HMM detection as already discussed above.

5 CONCLUSIONS

In summary, HMSS2 is an advanced comprehensive HMM-based tool for annotation and synteny analysis of prokaryotic sulfur metabolism. It has a higher speed and a much wider coverage than its predecessor HMS-S-S and now includes proteins involved in the metabolism

of inorganic and organic sulfur compounds. The use of curated functionally equivalent sequences for HMM training resulted in HMMs with high precision and recall. This also fills a gap in the coverage of sulfur metabolism prediction by HMMs. The application possibilities also include the combination with other HMMs from public databases or user-defined models and can therefore be extended according to the user's needs. The improved output formats are also applicable to ecology and evolutionary research.

ACKNOWLEDGEMENTS

This work was supported by the Deutsche Forschungsgemeinschaft (grant Da 351/13-1 to CD). TST received a scholarship from the Studienstiftung des Deutschen Volkes.

CONFLICT OF INTEREST

The authors declare that they have no competing interests.

AUTHOR CONTRIBUTIONS

TST and CD conceived the study. TST developed and implemented the method and performed the analyses. TST analysed and interpreted the data. Both authors wrote and approved the final version of the manuscript.

DATA AVAILABILITY STATEMENT

HMSS2 program files are available at https://github.com/TSTanabe/HMSS2.

References

Abola, A. P., Willits, M. G., Wang, R. C., & Long, S. R. (1999). Reduction of adenosine-5 '-phosphosulfate instead of 3 '- phosphoadenosine-5 '-phosphosulfate in cysteine biosynthesis by *Rhizobium meliloti* and other members of the family Rhizobiaceae. *Journal of Bacteriology, 181*, 5280-5287. https://doi.org/10.1128/JB.181.17.5280-5287.1999

Aguilar-Barajas, E., Diaz-Perez, C., Ramirez-Diaz, M. I., Riveros-Rosas, H., & Cervantes, C. (2011). Bacterial transport of sulfate, molybdate, and related oxyanions. *Biometals, 24*(4), 687-707. https://doi.org/10.1007/s10534-011-9421-x

Anantharaman, K., Hausmann, B., Jungbluth, S. P., Kantor, R. S., Lavy, A., Warren, L. A., Rappé, M. S., Pester, M., Loy, A., Thomas, B. C., & Banfield, J. F. (2018). Expanded diversity of microbial groups that shape the dissimilatory sulfur cycle. *ISME Journal, 12*, 1715-1728. https://doi.org/10.1038/s41396-018-0078-0

Benning, C., & Somerville, C. R. (1992a). Identification of an operon involved in sulfolipid biosynthesis in *Rhodobacter sphaeroides*. *Journal of Bacteriology, 174*(20), 6479-6487. https://doi.org/10.1128/jb.174.20.6479-6487.1992

Benning, C., & Somerville, C. R. (1992b). Isolation and genetic complementation of a sulfolipid-deficient mutant of *Rhodobacter sphaeroides*. *Journal of Bacteriology, 174*(7), 2352-2360. https://doi.org/10.1128/jb.174.7.2352-2360.1992

687 Bick, J. A., Dennis, J. J., Zylstra, G. J., Nowack, J., & Leustek, T. (2000). Identification of a new class of 5
688 '-adenylylsulfate (APS) reductases from sulfate-assimilating bacteria. *Journal of Bacteriology,*
689 *182*, 135-142.
690 Bilous, P. T., & Weiner, J. H. (1985). Dimethyl sulfoxide reductase activity by anaerobically grown
691 *Escherichia coli* HB101. *Journal of Bacteriology, 162*(3), 1151-1155.
692 https://doi.org/10.1128/jb.162.3.1151-1155.1985
693 Boden, R., Borodina, E., Wood, A. P., Kelly, D. P., Murrell, J. C., & Schäfer, H. (2011). Purification and
694 characterization of dimethylsulfide monooxygenase from *Hyphomicrobium sulfonivorans*.
695 *Journal of Bacteriology, 193*(5), 1250-1258.
696 Boden, R., & Hutt, L. P. (2019). Bacterial metabolism of C_1 sulfur compounds. In F. Rojo (Ed.), *Aerobic*
697 *utilization of hydrocarbons, oils and lipids. Handbook of hydrocarbon and lipid microbiology*
698 (pp. 1-43). Cham: Springer Nature Switzerland AG.
699 Boden, R., Kelly, D. P., Murrell, J. C., & Schäfer, H. (2010). Oxidation of dimethylsulfide to tetrathionate
700 by *Methylophaga thiooxidans* sp. nov.: a new link in the sulfur cycle. *Environmental*
701 *Microbiology, 12*(10), 2688-2699. https://doi.org/10.1111/j.1462-2920.2010.02238.x
702 Borodina, E., Kelly, D. P., Rainey, F. A., Ward-Rainey, N. L., & Wood, A. P. (2000). Dimethylsulfone as a
703 growth substrate for novel methylotrophic species of *Hyphomicrobium* and *Arthrobacter*.
704 *Archives of Microbiology, 173*(5-6), 425-437.
705 Borodina, E., Kelly, D. P., Schumann, P., Rainey, F. A., Ward-Rainey, N. L., & Wood, A. P. (2002). Enzymes
706 of dimethylsulfone metabolism and the phylogenetic characterization of the facultative
707 methylotrophs *Arthrobacter sulfonivorans* sp. nov., *Arthrobacter methylotrophus* sp. nov., and
708 *Hyphomicrobium sulfonivorans* sp. nov. *Archives of Microbiology, 177*(2), 173-183.
709 https://doi.org/10.1007/s00203-001-0373-3
710 Brodersen, K. H., Ong, C. S., Stephan, K. E., & Buhmann, J. M. (2010). The balanced accuracy and its
711 posterior distribution. *2010 International Conference on Pattern Recognition*, 3121-3124.
712 https://doi.org/10.1109/icpr.2010.764
713 Bruggemann, C., Denger, K., Cook, A. M., & Ruff, J. (2004). Enzymes and genes of taurine and
714 isethionate dissimilation in *Paracoccus denitrificans*. *Microbiology (Reading), 150*(Pt 4), 805-
715 816. https://doi.org/10.1099/mic.0.26795-0
716 Bullock, H. A., Reisch, C. R., Burns, A. S., Moran, M. A., & Whitman, W. B. (2014). Regulatory and
717 functional diversity of methylmercaptopropionate coenzyme A ligases from the
718 dimethylsulfoniopropionate demethylation pathway in *Ruegeria pomeroyi* DSS-3 and other
719 proteobacteria. *Journal of Bacteriology, 196*(6), 1275-1285.
720 https://doi.org/10.1128/JB.00026-14
721 Burrichter, A. G., Dorr, S., Bergmann, P., Haiss, S., Keller, A., Fournier, C., Franchini, P., Isono, E., &
722 Schleheck, D. (2021). Bacterial microcompartments for isethionate desulfonation in the
723 taurine-degrading human-gut bacterium *Bilophila wadsworthia*. *BMC Microbiology, 21*(1),
724 340. https://doi.org/10.1186/s12866-021-02386-w
725 Carrion, O., Curson, A. R. J., Kumaresan, D., Fu, Y., Lang, A. S., Mercade, E., & Todd, J. D. (2015). A novel
726 pathway producing dimethylsulphide in bacteria is widespread in soil environments. *Nature*
727 *Communications, 6*, 6579. https://doi.org/10.1038/ncomms7579
728 Carrion, O., Pratscher, J., Richa, K., Rostant, W. G., Farhan Ul Haque, M., Murrell, J. C., & Todd, J. D.
729 (2019). Methanethiol and dimethylsulfide cycling in Stiffkey saltmarsh. *Frontiers in*
730 *Microbiology, 10*, 1040. https://doi.org/10.3389/fmicb.2019.01040
731 Chen, X., Liu, L., Gao, X., Dai, X., Han, Y., Chen, Q., & Tang, K. (2021). Metabolism of chiral sulfonate
732 compound 2,3-dihydroxypropane-1-sulfonate (DHPS) by Roseobacter bacteria in marine
733 environment. *Environment International, 157*, 106829.
734 https://doi.org/10.1016/j.envint.2021.106829
735 Chicco, D. (2017). Ten quick tips for machine learning in computational biology. *BioData Mining, 10*,
736 35. https://doi.org/10.1186/s13040-017-0155-3

737 Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over
738 F1 score and accuracy in binary classification evaluation. *BMC Genomics, 21*(1), 6.
739 https://doi.org/10.1186/s12864-019-6413-7
740 Curson, A. R., Todd, J. D., Sullivan, M. J., & Johnston, A. W. (2011). Catabolism of
741 dimethylsulphoniopropionate: microorganisms, enzymes and genes. *Nature Reviews*
742 *Microbiology, 9*(12), 849-859. https://doi.org/10.1038/nrmicro2653
743 De Zwart, J., Sluis, J., & Kuenen, J. G. (1997). Competition for dimethyl sulfide and hydrogen sulfide by
744 *Methylophaga sulfidovorans* and *Thiobacillus thioparus* T5 in continuous Ccltures. *Applied and*
745 *Environmental Microbiology, 63*(8), 3318-3322. https://doi.org/10.1128/aem.63.8.3318-
746 3322.1997
747 Denger, K., Mayer, J., Buhmann, M., Weinitschke, S., Smits, T. H., & Cook, A. M. (2009). Bifurcated
748 degradative pathway of 3-sulfolactate in *Roseovarius nubinhibens* ISM via sulfoacetaldehyde
749 acetyltransferase and (S)-cysteate sulfolyase. *Journal of Bacteriology, 191*(18), 5648-5656.
750 https://doi.org/10.1128/JB.00569-09
751 Denger, K., Weiss, M., Felux, A. K., Schneider, A., Mayer, C., Spiteller, D., Huhn, T., Cook, A. M., &
752 Schleheck, D. (2014). Sulphoglycolysis in *Escherichia coli* K-12 closes a gap in the
753 biogeochemical sulphur cycle. *Nature, 507*(7490), 114-117.
754 https://doi.org/10.1038/nature12947
755 Dhouib, R., Nasreen, M., Othman, D., Ellis, D., Lee, S., Essilfie, A. T., Hansbro, P. M., McEwan, A. G., &
756 Kappler, U. (2021). The DmsABC sulfoxide reductase supports virulence in non-typeable
757 *Haemophilus influenzae*. *Frontiers in Microbiology, 12*, 686833.
758 https://doi.org/10.3389/fmicb.2021.686833
759 Duarte, A. G., Barbosa, A. C. C., Ferreira, D., Manteigas, G., Domingos, R. M., & Pereira, I. A. C. (2021).
760 Redox loops in anaerobic respiration - The role of the widespread NrfD protein family and
761 associated dimeric redox module. *Biochimica et Biophysica Acta Bioenergetics, 1862*(7),
762 148416. https://doi.org/10.1016/j.bbabio.2021.148416
763 Eichhorn, E., van der Ploeg, J. R., & Leisinger, T. (1999). Characterization of a two-component
764 alkanesulfonate monooxygenase from *Escherichia coli*. *Journal of Biological Chemistry,*
765 *274*(38), 26639-26646. https://doi.org/10.1074/jbc.274.38.26639
766 El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., Qureshi, M., Richardson, L. J.,
767 Salazar, G. A., Smart, A., Sonnhammer, E. L. L., Hirsh, L., Paladin, L., Piovesan, D., Tosatto, S. C.
768 E., & Finn, R. D. (2019). The Pfam protein families database in 2019. *Nucleic Acids Research,*
769 *47*(D1), D427-D432. https://doi.org/10.1093/nar/gky995
770 Eyice, O., Myronova, N., Pol, A., Carrión, O., Todd, J. D., Smith, T. J., Gurman, S. J., Cuthbertson, A.,
771 Mazard, S., Mennink-Kersten, M. A., Bugg, T. D., Andersson, K. K., Johnston, A. W., Op den
772 Camp, H. J., & Schäfer, H. (2017). Bacterial SBP56 identified as a Cu-dependent methanethiol
773 oxidase widely distributed in the biosphere. *ISME Journal, 12*(1), 145-160.
774 https://doi.org/10.1038/ismej.2017.148
775 Felux, A. K., Spiteller, D., Klebensberger, J., & Schleheck, D. (2015). Entner-Doudoroff pathway for
776 sulfoquinovose degradation in *Pseudomonas putida* SQ1. *Proceedings of the National*
777 *Academy of Sciences of the United States of America, 112*(31), E4298-4305.
778 https://doi.org/10.1073/pnas.1507049112
779 Forman, G., & Scholz, M. (2010). Apples-to-apples in cross-validation studies. *ACM SIGKDD*
780 *Explorations Newsletter, 12*(1), 49-57. https://doi.org/10.1145/1882471.1882479
781 Frommeyer, B., Fiedler, A. W., Oehler, S. R., Hanson, B. T., Loy, A., Franchini, P., Spiteller, D., &
782 Schleheck, D. (2020). Environmental and intestinal phylum Firmicutes bacteria metabolize the
783 plant sugar sulfoquinovose via a 6-deoxy-6-sulfofructose transaldolase pathway. *iScience,*
784 *23*(9), 101510. https://doi.org/10.1016/j.isci.2020.101510
785 Garcia, P. S., D'Angelo, F., Ollagnier de Choudens, S., Dussouchaud, M., Bouveret, E., Gribaldo, S., &
786 Barras, F. (2022). An early origin of iron-sulfur cluster biosynthesis machineries before Earth
787 oxygenation. *Nature Ecology & Evolution, 6*(10), 1564-1572. https://doi.org/10.1038/s41559-
788 022-01857-1

789 Garcia, P. S., Jauffrit, F., Grangeasse, C., & Brochier-Armanet, C. (2019). GeneSpy, a user-friendly and
790 flexible genomic context visualizer. *Bioinformatics, 35*(2), 329-331.
791 https://doi.org/10.1093/bioinformatics/bty459
792 Goddard-Borger, E. D., & Williams, S. J. (2017). Sulfoquinovose in the biosphere: occurrence,
793 metabolism and functions. *Biochemical Journal, 474*(5), 827-849.
794 https://doi.org/10.1042/BCJ20160508
795 Gorzynska, A. K., Denger, K., Cook, A. M., & Smits, T. H. M. (2006). Inducible transcription of genes
796 involved in taurine uptake and dissimilation by *Silicibacter pomeroyi* DSS-3^T. *Archives of*
797 *Microbiology, 185*(5), 402-406. https://doi.org/10.1007/s00203-006-0106-8
798 Guler, S., Essigmann, B., & Benning, C. (2000). A cyanobacterial gene, *sqdX*, required for biosynthesis
799 of the sulfolipid sulfoquinovosyldiacylglycerol. *Journal of Bacteriology, 182*(2), 543-545.
800 https://doi.org/10.1128/JB.182.2.543-545.2000
801 Haft, D. H., DiCuccio, M., Badretdin, A., Brover, V., Chetvernin, V., O'Neill, K., Li, W., Chitsaz, F.,
802 Derbyshire, M. K., Gonzales, N. R., Gwadz, M., Lu, F., Marchler, G. H., Song, J. S., Thanki, N.,
803 Yamashita, R. A., Zheng, C., Thibaud-Nissen, F., Geer, L. Y., Marchler-Bauer, A., & Pruitt, K. D.
804 (2018). RefSeq: an update on prokaryotic genome annotation and curation. *Nucleic Acids*
805 *Research, 46*(D1), D851-D860. https://doi.org/10.1093/nar/gkx1068
806 Hanson, B. T., Dimitri Kits, K., Loffler, J., Burrichter, A. G., Fiedler, A., Denger, K., Frommeyer, B.,
807 Herbold, C. W., Rattei, T., Karcher, N., Segata, N., Schleheck, D., & Loy, A. (2021).
808 Sulfoquinovose is a select nutrient of prominent bacteria and a source of hydrogen sulfide in
809 the human gut. *ISME J*. https://doi.org/10.1038/s41396-021-00968-0
810 Henriques, A. C., & De Marco, P. (2015). Methanesulfonate (MSA) catabolic genes from marine and
811 estuarine bacteria. *PLoS One, 10*(5), e0125735.
812 https://doi.org/10.1371/journal.pone.0125735
813 Horinouchi, M., Yoshida, T., Nojiri, H., Yamane, H., & Omori, T. (1999). Polypeptide requirement of
814 multicomponent monooxygenase DsoABCDEF for dimethyl sulfide oxidizing activity.
815 *Bioscience, biotechnology and biochemistry, 63*(10), 1765-1771.
816 https://doi.org/10.1271/bbb.63.1765
817 Jaffe, A. L., Castelle, C. J., Matheus Carnevali, P. B., Gribaldo, S., & Banfield, J. F. (2020). The rise of
818 diversity in metabolic platforms across the Candidate Phyla Radiation. *BMC Biology, 18*(1), 69.
819 https://doi.org/10.1186/s12915-020-00804-5
820 Jeni, L. A., Cohn, J. F., & De La Torre, F. (2013). Facing imbalanced data recommendations for the use
821 of performance metrics. *Int Conf Affect Comput Intell Interact Workshops, 2013*, 245-251.
822 https://doi.org/10.1109/ACII.2013.47
823 Kappler, U., & Schäfer, H. (2014). Transformations of dimethylsulfide. *Metal Ions in Life Sciences, 14*,
824 279-313.
825 Kelly, D. P., & Murrell, J. C. (1999). Microbial metabolism of methanesulfonic acid. *Archives of*
826 *Microbiology, 172*(6), 341-348. https://doi.org/10.1007/s002030050770
827 Kiene, R. P., Linn, L. J., & Bruton, J. A. (2000). New and important roles for DMSP in marine microbial
828 communities. *Journal of Sea Research, 43*(3-4), 209-224. https://doi.org/10.1016/s1385-
829 1101(00)00023-x
830 Koch, T., & Dahl, C. (2018). A novel bacterial sulfur oxidation pathway provides a new link between the
831 cycles of organic and inorganic sulfur compounds. *ISME Journal, 12*(10), 2479-2491.
832 https://doi.org/10.1038/s41396-018-0209-7
833 Kredich, N. M. (1996). Biosynthesis of cysteine. In F. C. Neidhardt (Ed.), *Escherichia coli and Salmonella*
834 *typhimurium. Cellular and molecular biology* (pp. 514-527). Washington D.C.: American Society
835 for Microbiology.
836 Krejcik, Z., Hollemeyer, K., Smits, T. H. M., & Cook, A. M. (2010). Isethionate formation from taurine in
837 *Chromohalobacter salexigens*: purification of sulfoacetaldehyde reductase. *Microbiology*
838 *(Reading), 156*(Pt 5), 1547-1555. https://doi.org/10.1099/mic.0.036699-0

Kröber, E., & Schäfer, H. (2019). Identification of proteins and genes expressed by *Methylophaga thiooxydans* during growth on dimethylsulfide and their presence in other members of the genus. *Frontiers in Microbiology, 10*, 1132. https://doi.org/10.3389/fmicb.2019.01132

Leimkühler, S., & Iobbi-Nivol, C. (2016). Bacterial molybdoenzymes: old enzymes for new purposes. *FEMS Microbiology Reviews, 40*(1), 1-18. https://doi.org/10.1093/femsre/fuv043

Letunic, I., & Bork, P. (2021). Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Research, 49*(W1), W293-W296. https://doi.org/10.1093/nar/gkab301

Leyh, T. S., Taylor, J. C., & Markham, G. D. (1988). The sulfate activation locus of *Escherichia coli* K12: cloning, genetic, and enzymatic characterization. *Journal of Biological Chemistry, 263*, 2409-2416.

Li, J., Koch, J., Flegler, W., Garcia Ruiz, L., Hager, N., Ballas, A., Tanabe, T. S., & Dahl, C. (2022). A metabolic puzzle: consumption of C_1 compounds and thiosulfate in *Hyphomicrobium denitrificans* X^T. *Biochimica et Biophysica Acta (BBA) - Bioenergetics, 1864*, 148932. https://doi.org/10.1016/j.bbabio.2022.148932

Lovelock, J. E., Maggs, R. J., & Rasmussen, R. A. (1972). Atmospheric cimethyl sulphide and the natural sulphur cycle. *Nature, 237*(5356), 452-453. https://doi.org/10.1038/237452a0

Mayer, J., Huhn, T., Habeck, M., Denger, K., Hollemeyer, K., & Cook, A. M. (2010). 2,3-Dihydroxypropane-1-sulfonate degraded by *Cupriavidus pinatubonensis* JMP134: purification of dihydroxypropanesulfonate 3-dehydrogenase. *Microbiology (Reading), 156*(Pt 5), 1556-1564. https://doi.org/10.1099/mic.0.037580-0

McDevitt, C. A., Hanson, G. R., Noble, C. J., Cheesman, M. R., & McEwan, A. G. (2002). Characterization of the redox centers in dimethyl sulfide dehydrogenase from *Rhodovulum sulfidophilum*. *Biochemistry, 41*(51), 15234-15244.

McDevitt, C. A., Hugenholtz, P., Hanson, G. R., & McEwan, A. G. (2002). Molecular analysis of dimethyl sulphide dehydrogenase from *Rhodovulum sulfidophilum*: its place in the dimethyl sulphoxide reductase family of microbial molybdopterin-containing enzymes. *Molecular Microbiology, 44*(6), 1575-1587.

McEwan, A. G., Hanson, G. R., & Bailey, S. (1998). Dimethylsulphoxide reductase from purple phototrophic bacteria: structures and mechanism(s). *Biochemical Society Transactions, 26*, 390-396.

McEwan, A. G., Ridge, J. P., McDevitt, C. A., & Hugenholtz, P. (2010). The DMSO reductase family of microbial molybdenum enzymes; molecular properties and role in the dissimilatory reduction of toxic elements. *Geomicrobiology Journal, 19*(1), 3-21. https://doi.org/10.1080/014904502317246138

Moran, M. A., & Durham, B. P. (2019). Sulfur metabolites in the pelagic ocean. *Nature Reviews Microbiology*. https://doi.org/10.1038/s41579-019-0250-1

Peck, S. C., Denger, K., Burrichter, A., Irwin, S. M., Balskus, E. P., & Schleheck, D. (2019). A glycyl radical enzyme enables hydrogen sulfide production by the human intestinal bacterium *Bilophila wadsworthia*. *Proceedings of the National Academy of Sciences of the United States of America, 116*(8), 3171-3176. https://doi.org/10.1073/pnas.1815661116

Refaeilzadeh, P., Tang, L., & Liu, H. (2009). Cross-validation. In L. Liu & T. Özsu (Eds.), *Encyclopedia of database systems* (pp. 532-538). Boston, MA: Springer.

Reisch, C. R., Stoudemayer, M. J., Varaljay, V. A., Amster, I. J., Moran, M. A., & Whitman, W. B. (2011). Novel pathway for assimilation of dimethylsulphoniopropionate widespread in marine bacteria. *Nature, 473*(7346), 208-211. https://doi.org/10.1038/nature10078

Rossak, M., Schafer, A., Xu, N., Gage, D. A., & Benning, C. (1997). Accumulation of sulfoquinovosyl-1-O-dihydroxyacetone in a sulfolipid-deficient mutant of *Rhodobacter sphaeroides* inactivated in *sqdC*. *Archives of Biochemistry and Biophysics, 340*(2), 219-230. https://doi.org/10.1006/abbi.1997.9931

Rossak, M., Tietje, C., Heinz, E., & Benning, C. (1995). Accumulation of UDP-sulfoquinovose in a sulfolipid-deficient mutant of *Rhodobacter sphaeroides*. *Journal of Biological Chemistry, 270*(43), 25792-25797. https://doi.org/10.1074/jbc.270.43.25792

Ruff, J., Denger, K., & Cook, A. M. (2003). Sulphoacetaldehyde acetyltransferase yields acetyl phosphate: purification from *Alcaligenes defragrans* and gene clusters in taurine degradation. *Biochemical Journal, 369*(Pt 2), 275-285. https://doi.org/10.1042/BJ20021455

Sayers, E. W., Cavanaugh, M., Clark, K., Ostell, J., Pruitt, K. D., & Karsch-Mizrachi, I. (2019). GenBank. *Nucleic Acids Research, 47*(D1), D94-D99. https://doi.org/10.1093/nar/gky989

Schäfer, H., Myronova, N., & Boden, R. (2010). Microbial degradation of dimethylsulphide and related C_1-sulphur compounds: organisms and pathways controlling fluxes of sulphur in the biosphere. *Journal of Experimental Botany, 61*(2), 315-334. https://doi.org/10.1093/jxb/erp355

Schmitz, R. A., Mohammadi, S. S., van Erven, T., Berben, T., Jetten, M. S. M., Pol, A., & Op den Camp, H. J. M. (2022). Methanethiol consumption and hydrogen sulfide production by the thermoacidophilic methanotroph *Methylacidiphilum fumariolicum* SolV. *Frontiers in Microbiology, 13*, 857442. https://doi.org/10.3389/fmicb.2022.857442

Sharma, M., Lingford, J. P., Petricevic, M., Snow, A. J. D., Zhang, Y., Jarva, M. A., Mui, J. W., Scott, N. E., Saunders, E. C., Mao, R., Epa, R., da Silva, B. M., Pires, D. E. V., Ascher, D. B., McConville, M. J., Davies, G. J., Williams, S. J., & Goddard-Borger, E. D. (2022). Oxidative desulfurization pathway for complete catabolism of sulfoquinovose by bacteria. *Proceedings of the National Academy of Sciences of the United States of America, 119*(4). https://doi.org/10.1073/pnas.2116022119

Struwe, M. A., Kalimuthu, P., Luo, Z., Zhong, Q., Ellis, D., Yang, J., Khadanand, K. C., Harmer, J. R., Kirk, M. L., McEwan, A. G., Clement, B., Bernhardt, P. V., Kobe, B., & Kappler, U. (2021). Active site architecture reveals coordination sphere flexibility and specificity determinants in a group of closely related molybdoenzymes. *Journal of Biological Chemistry, 296*, 100672. https://doi.org/10.1016/j.jbc.2021.100672

Tanabe, T. S., & Dahl, C. (2022). HMS-S-S: a tool for the identification of sulphur metabolism-related genes and analysis of operon structures in genome and metagenome assemblies. *Molecular Ecology Resources, 22*(7), 2758-2774. https://doi.org/10.1111/1755-0998.13642

Tanaka, Y., Yoshikaie, K., Takeuchi, A., Ichikawa, M., Mori, T., Uchino, S., Sugano, Y., Hakoshima, T., Takagi, H., Nonaka, G., & Tsukazaki, T. (2020). Crystal structure of a YeeE/YedE family protein engaged in thiosulfate uptake. *Science Advances, 6*(35), eaba7637. https://doi.org/10.1126/sciadv.aba7637

Thume, K., Gebser, B., Chen, L., Meyer, N., Kieber, D. J., & Pohnert, G. (2018). The metabolite dimethylsulfoxonium propionate extends the marine organosulfur cycle. *Nature, 563*(7731), 412-415. https://doi.org/10.1038/s41586-018-0675-0

Todd, J. D., Curson, A. R., Sullivan, M. J., Kirkwood, M., & Johnston, A. W. (2012). The *Ruegeria pomeroyi acuI* gene has a role in DMSP catabolism and resembles *yhdH* of *E. coli* and other bacteria in conferring resistance to acrylate. *PLoS One, 7*(4), e35947. https://doi.org/10.1371/journal.pone.0035947

Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics, 7*, 91. https://doi.org/10.1186/1471-2105-7-91

Wei, Y., Tong, Y., & Zhang, Y. (2022). New mechanisms for bacterial degradation of sulfoquinovose. *Bioscience Reports, 42*(10). https://doi.org/10.1042/BSR20220314

Weinitschke, S., Denger, K., Cook, A. M., & Smits, T. H. M. (2007). The DUF81 protein TauE in *Cupriavidus necator* H16, a sulfite exporter in the metabolism of C_2 sulfonates. *Microbiology, 153*, 3055-3060. https://doi.org/10.1099/mic.0.2007/009845-0

Weinitschke, S., Hollemeyer, K., Kusian, B., Bowien, B., Smits, T. H., & Cook, A. M. (2010). Sulfoacetate is degraded via a novel pathway involving sulfoacetyl-CoA and sulfoacetaldehyde in *Cupriavidus necator* H16. *Journal of Biological Chemistry, 285*(46), 35249-35254. https://doi.org/10.1074/jbc.M110.127043

939 Weinitschke, S., Sharma, P. I., Stingl, U., Cook, A. M., & Smits, T. H. (2010). Gene clusters involved in
940 isethionate degradation by terrestrial and marine bacteria. *Applied and Environmental*
941 *Microbiology, 76*(2), 618-621. https://doi.org/10.1128/AEM.01818-09
942 Wicht, D. K. (2016). The reduced flavin-dependent monooxygenase SfnG converts dimethylsulfone to
943 methanesulfinate. *Archives of Biochemistry and Biophysics, 604*, 159-166.
944 https://doi.org/10.1016/j.abb.2016.07.001
945 Wolf, P. G., Cowley, E. S., Breister, A., Matatov, S., Lucio, L., Polak, P., Ridlon, J. M., Gaskins, H. R., &
946 Anantharaman, K. (2022). Diversity and distribution of sulfur metabolic genes in the human
947 gut microbiome and their association with colorectal cancer. *Microbiome, 10*(1), 64.
948 https://doi.org/10.1186/s40168-022-01242-x

949

950

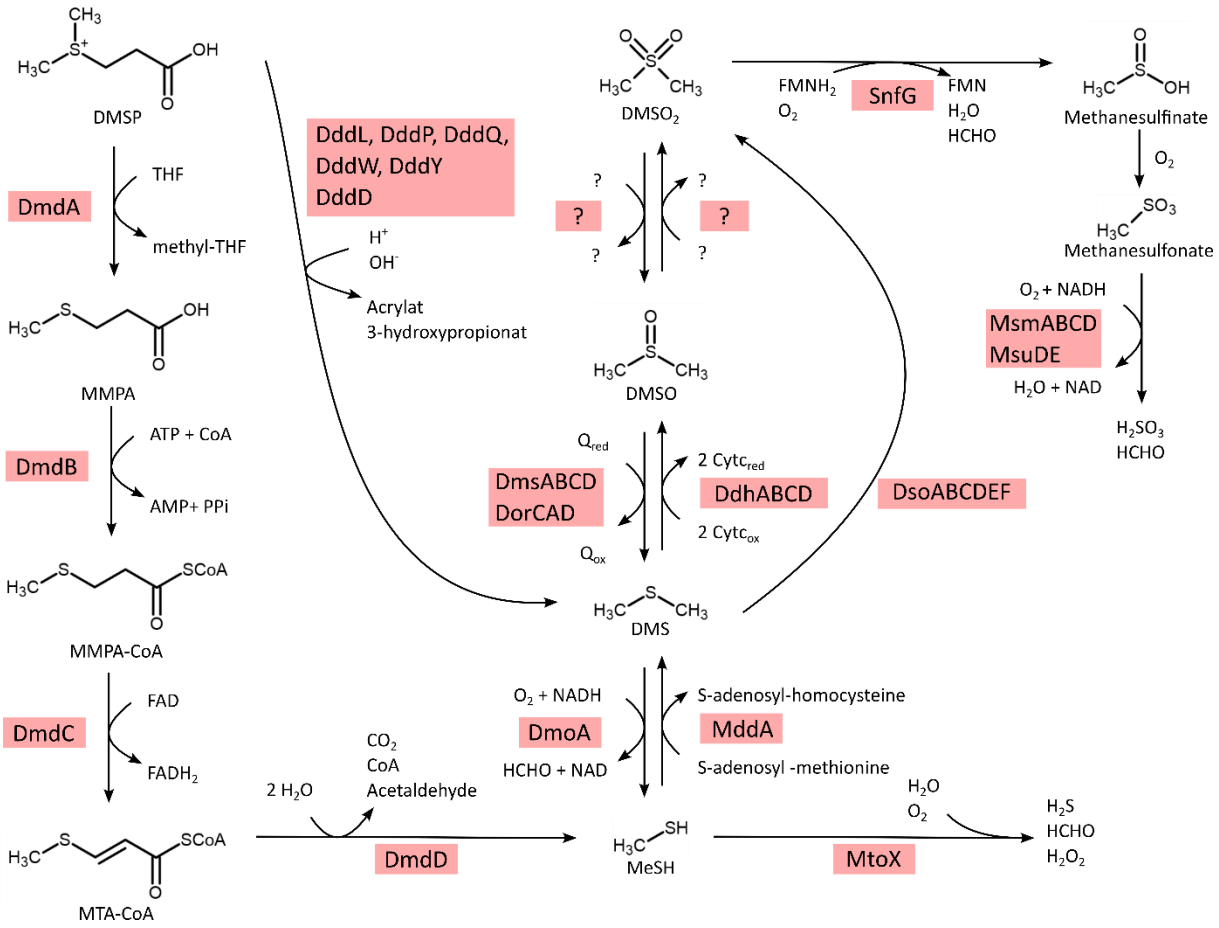


Figure 1 Prokaryotic metabolism of C_1 organosulfur compounds. All proteins shown have a corresponding HMM in HMSS2. Cytc, Cytochrome *c*; DMSP, dimethylsulfoniopropionate; DHPS, 2,3-dihydroxypropane-1-sulfonate; DMS, dimethylsulfide; DMSO, dimethylsulfone, $DMSO_2$ dimethylsulfoxide; FMN, flavin mononucleotide; $FMNH_2$, reduced flavin mononucleotide; MeSH, methanethiol; MMPA, methylmercaptopropionate; MMPA-CoA, 3-methylmercaptopropionyl-CoA; MTA-CoA, methylthioacryloyl-CoA; THF, tetrahydrofolate.

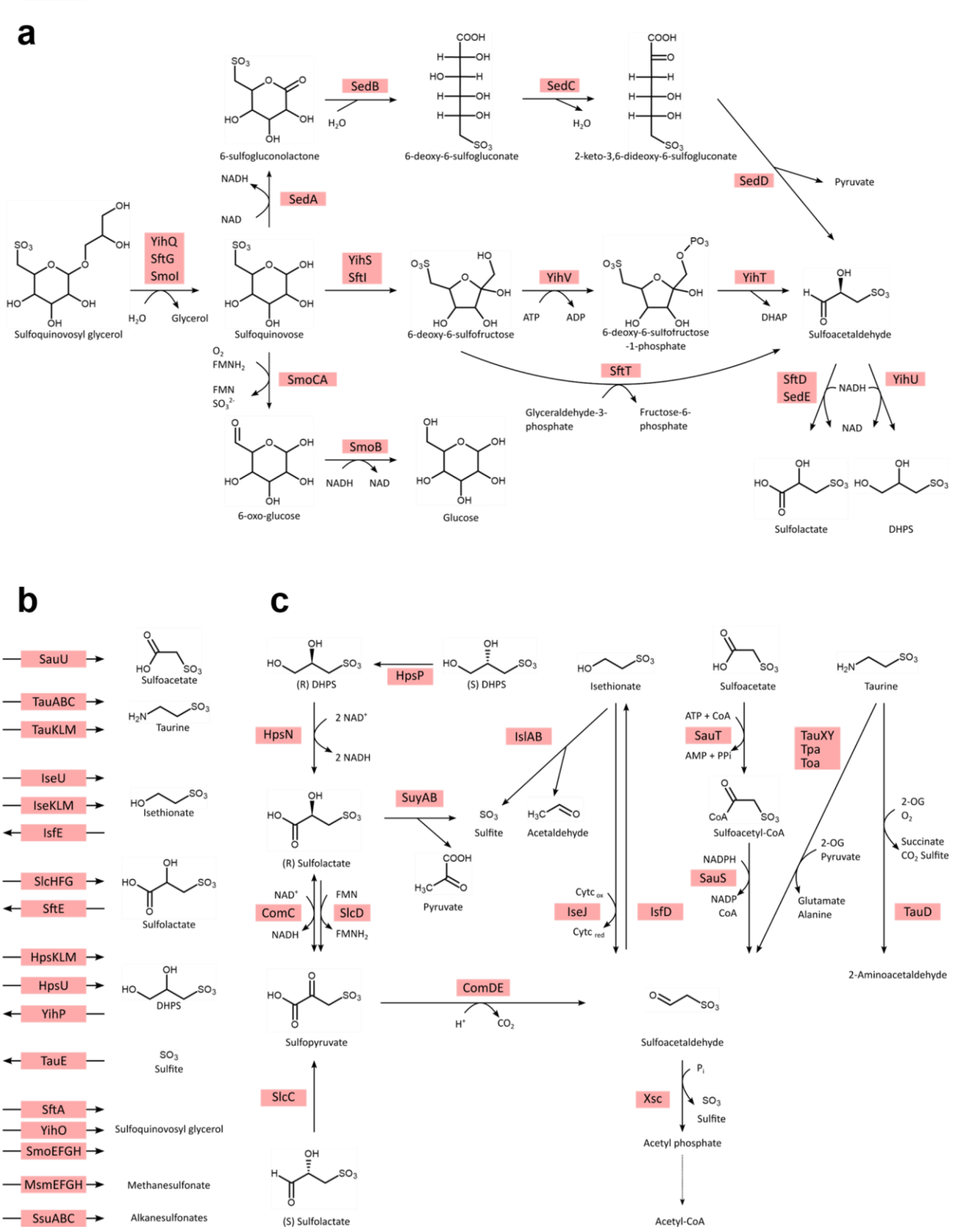
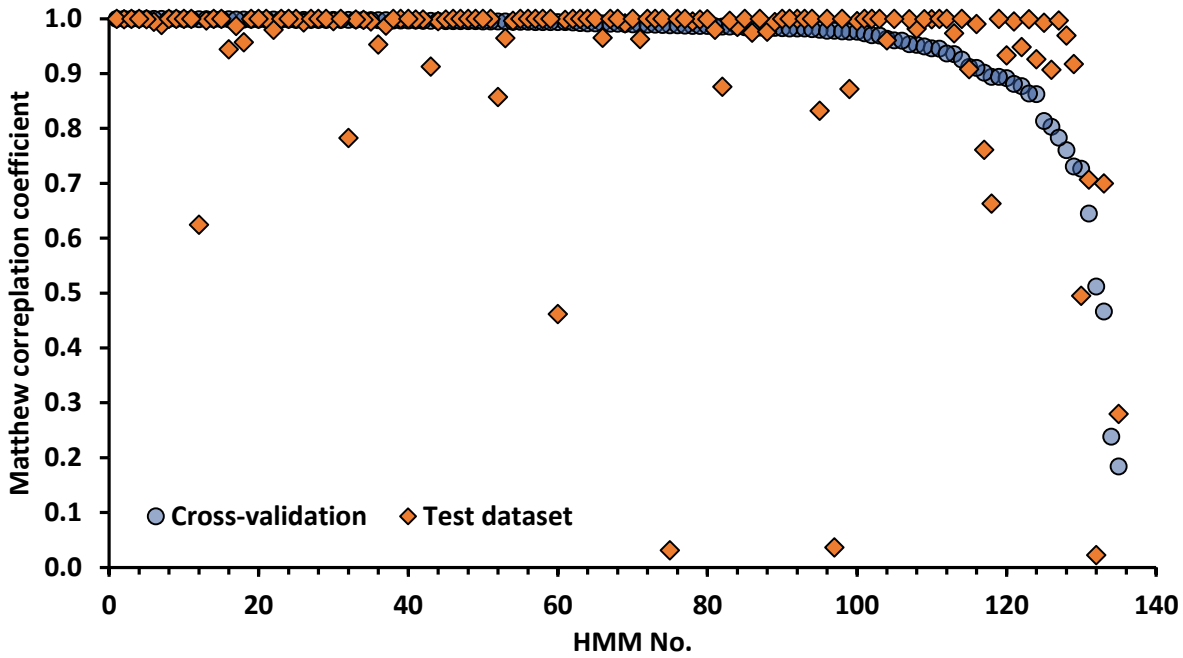


Figure 2. Prokaryotic metabolism of organosulfur compounds with two or more carbon atoms and relevant transporters. (a) Pathways of sulfoquinovosyl glycerol degradation. (b) Transport systems for import and export of organic sulfur compounds. (c) Degradation pathways of C_2 and C_3 organosulfur compounds. Usually, the same cell does not contain all the pathways. All proteins show have a corresponding HMM in HMSS2. Cytc, Cytochrome c; DHPS, 2,3-dihydroxypropane-1-sulfonate; FMN, flavin mononucleotide; $FMNH_2$, reduced flavin mononucleotide; 2-OG, 2-oxoglutarate.

966



967

968 **Figure 3 Validation of the 134 HMMs generated in this work.** Performance was assessed by cross-
969 validation (blue dots) and on an independent test dataset (red diamonds). For each HMM Matthew
970 correlation coefficient was calculated. HMMs were ranked by their performance in cross-validation.

971

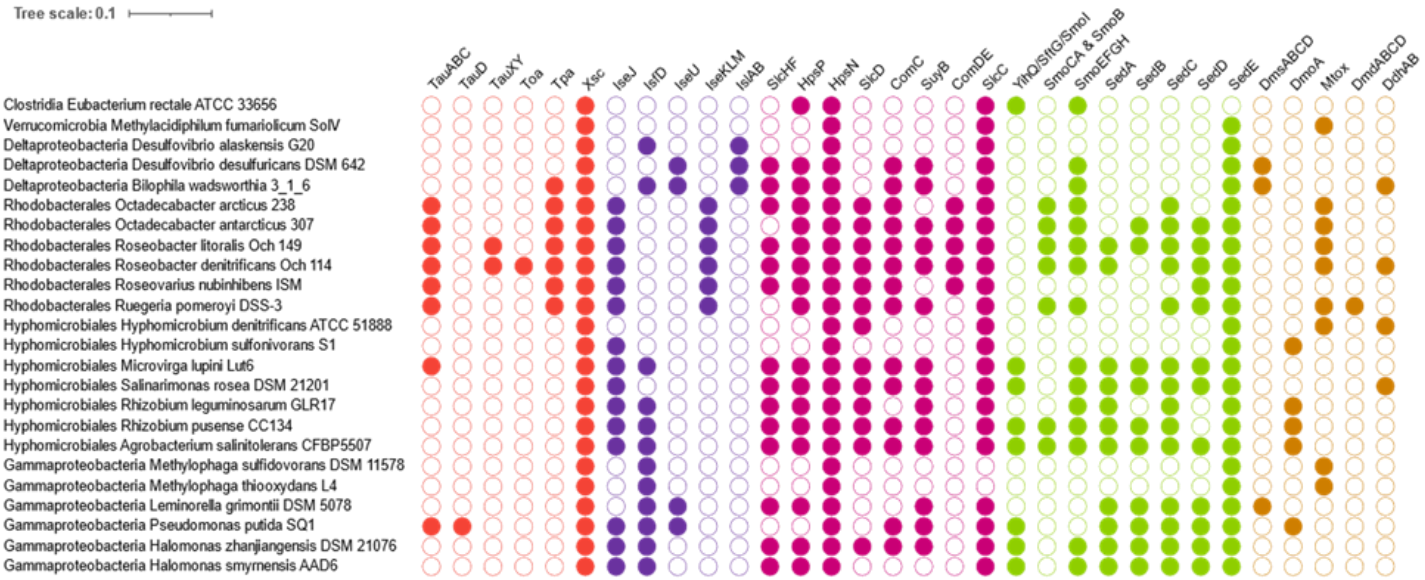


Figure 4. Presence/absence of proteins involved in the metabolism of organic sulfur compounds. Occurrence of genes for proteins involved in taurine degradation, isethionate degradation, 2,3-dihydroxypropane-1-sulfonate, sulfoquinovose and DMS metabolism, is indicated by filled orange, violet, purple, green and light brown circles, respectively. The function of the individual proteins can be deduced from Figures 1 and 2.

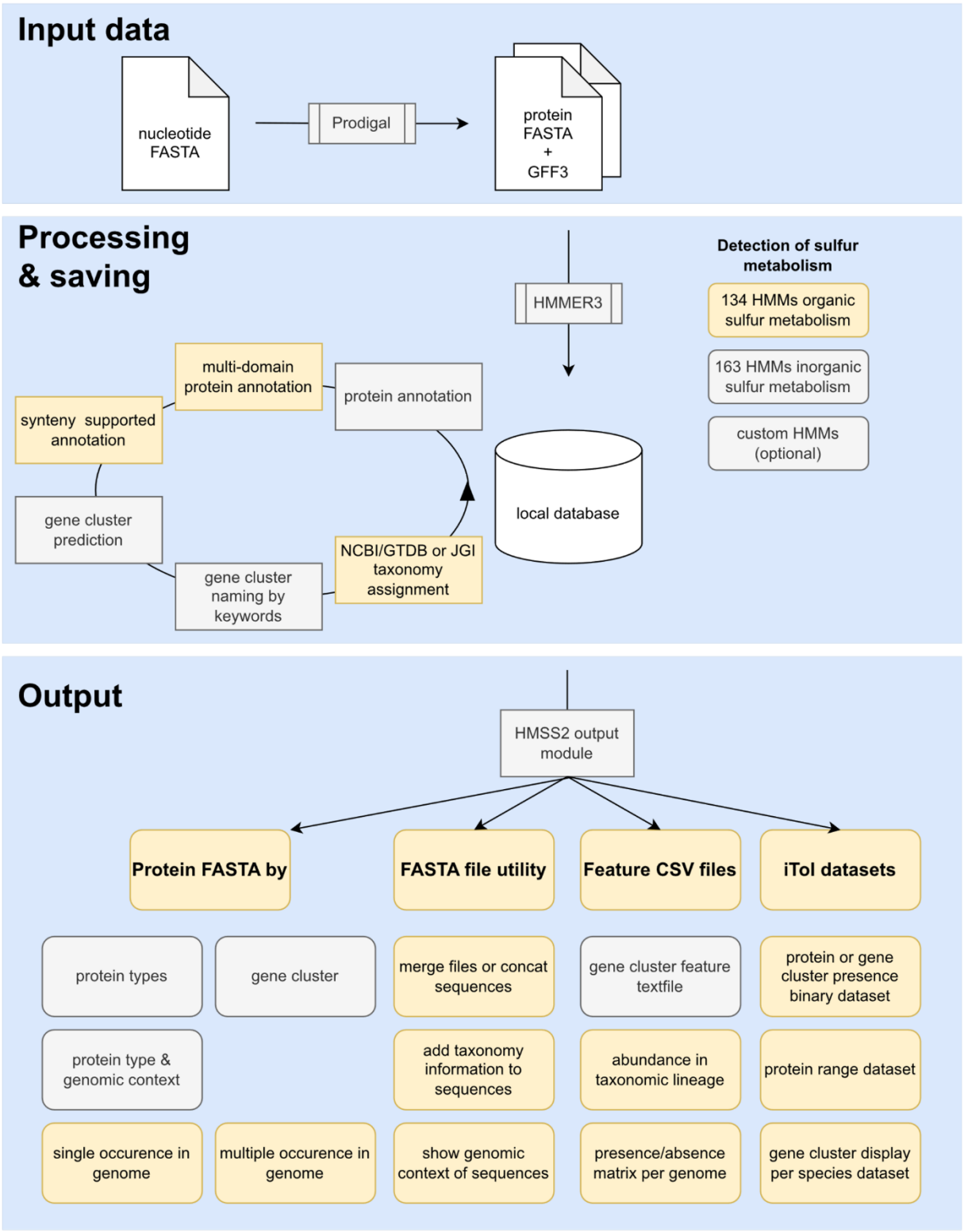
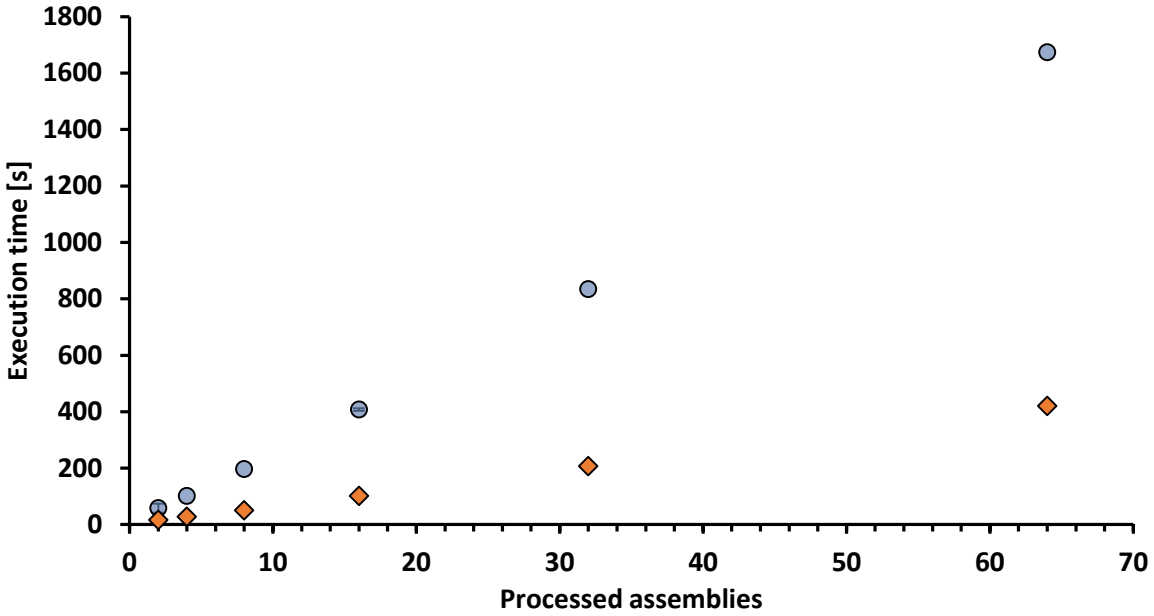


Figure 5. Algorithm overview of HMSS2. New features added in HMSS2 are highlighted in yellow. The only external programs required are HMMER3 and Prodigal.



984

985 **Figure 6. Computing time required by HMS-S-S compared to HMSS2.** Test were performed in triplicate
986 with defined numbers of randomly selected sulfur-oxidizing or sulfur-reducing prokaryotes and 164
987 HMMs. Blue circles: HMS-S-S, organe diamonds: HMSS2

988
989

Supplementary Tables and Figures

Table S1. Reference proteins for dataset annotation

Table S2. HMM performance evaluation

Table S3. HMS-S-S vs. HMSS2 Benchmark

Table S4. Organisms for case study