DiSCo: a sequence-based type-specific predictor of Dsr-dependent dissimilatory sulphur metabolism in microbial data

Supplementary Information

Sinje Neukirchen, Filipa L. Sousa*

Department of Functional and Evolutionary Ecology, University of Vienna, Althanstraße 14, 1090 Vienna, Austria

*Correspondence: filipa.sousa@univie.ac.at

DiSCo validation and comparison with other methods in the complete genomes dataset

The complete genomes dataset was analysed with DiSCo. These DiSCo assignments were compared with rBBH and simple best BLAST hits and contingency matrices were created. The accuracy (AC), false discovery rate (FDR), balanced accuracy (BA) [1], recall (RC), and precision (PR) of the methods were calculated and compared (Table S7). Regarding DiSCo assignments, the DsrC models identified DsrC proteins in all of the 103 DsrAB containing micro-organisms (RC=1, AC=1, BA=1) but in one case, with a different predicted type than the one from the DsrAB proteins. Five additional DsrC proteins were found, in micro-organisms devoid of DsrAB proteins, being classified as false positives resulting in a precision of 0.94 (FDR=0.06). The presence of DsrC proteins in micro-organisms without the Dsr-dependent dissimilatory sulphur metabolism was already reported [2], but so far, the function of the protein *in vivo* is not known. Being a small protein (~100 amino acids) involved in sulphur reactions, it might have been recruited to perform a similar reaction, in a different metabolic pathway.

The balanced accuracy, accuracy, recall, and precision of DsrM and DsrK models varied between 1 and 0.96, indicating a good correspondence between DiSCo predictions (in terms of identification and type) and the ones from DsrAB proteins. This means that these models are not only able to distinguish DsrMK proteins from known paralogous protein families [3], but also to distinguish (and independently classify) DsrMK proteins from the reductive and oxidative type. The DsrMK complex was found in three genomes devoid of DsrAB, which were classified as false positives (FDR=0.03-0.04). All three genomes contain DsrC-like proteins with two cysteines, similar to previous observations [2]. The DsrMK complex is thought to recycle the persulphurated DsrC [4], thus, the co-distribution of DsrMK with additional Dsr proteins could indicate a similar role of DsrMK in other sulphur-trafficking processes. A false negative of DsrK was found in *Desulfurella acetivorans*, a known dSDM [5, 6]. Additionally, reductive-type DsrABCD proteins and an oxidative DsrM protein, considered as false positive, were found in this genome. A closer inspection showed, that this DsrM protein is phylogenetically distinct from reductive DsrM proteins and more similar to the ones present in dSOB. Unfortunately, the DsrK protein is encoded on a pseudo gene, and its corresponding protein sequence is missing in the dataset. Hence, the DsrK protein could not be identified by any method. Lower values for the balanced accuracy (BA=0.88-0.89) were obtained in the cases of the DsrJ, DsrO and DsrP proteins. These should be considered to be lower boundaries, since in some micro-organisms, mainly Gram-positive dSRP, instead of the DsrMKJOP complex, the DsrMK version is present [7]. Thus, the values of false negatives obtained for DsrJOP proteins are inflated, affecting the determination of the parameters (RC=0.77-0.79). On the other hand, only one false positive was identified for DsrJ (FDR=0.01), and none in the case of DsrO and DsrP proteins (FDR=0.00). The identified DsrJ protein was found in the genome of *Beggiatoa leptomitoformis* D-402, which lacks the DsrAB complex, but additional DsrC, DsrEFH and DsrMK proteins were found. This species possesses the flavocytochrome c-sulphide dehydrogenase and the Sox system for the oxidation of sulphur to sulphite and the presence of the proteins DsrC and DsrEFH was already reported [8].

The DsrEFH proteins are essential in dSOB [9] and usually used as marker for oxidative processes. In 35 out of the 36 dSOB found in the complete genomes dataset, the

DsrEFH models were able to independently identify the respective protein. The exception was in *Chlorobaculum parvum*, known to have lost the *dsrEFH* genes [10, 11]. In terms of balanced accuracy, accuracy and recall, the performance of the DsrEFH models varied between 0.97 and 1, showing the high sensitivity of the method. Slightly lower values of precision (PC=0.85-0.95) were found due to the identification of the proteins in some genomes devoid of DsrAB proteins (DsrE FDR=0.15, DsrFH FDR=0.05). Even considering that, as in the case of DsrC, the DsrEFH proteins might have been recruited to perform a different function [12], in here they were conservatively considered as false positives. The models for the DsrD protein showed also high precision and accuracy values of 1. DsrD proteins were identified in all genomes containing a reductive DsrAB complex, with exception of members of the family Thermoproteaceae. At least one micro-organism of this archaeal family is able to perform the reduction of sulphate to sulphide without the DsrD protein [13]. This explains, at least partially, the DsrD model's recall of only 0.84 (BA = 0.92, FDR=0.0).

Within the complete genomes dataset, the DsrL protein was found in all micro-organisms containing a oxidative DsrAB complex and in *Desulfurella acetivorans*, an micro-organism with a reductive DsrAB complex, reported to disproportionate elemental sulphur [5, 6] (Table S1). However, a direct connection between the co-distribution of DsrL proteins and reductive DsrAB complexes with the ability to disproportionate S-species was not possible in this dataset. In the S-disproportionating *D. amilsii*, DsrL is proposed to be involved in thiosulphate reduction indicating that, at least for some micro-organisms, DsrL is not involved in S-disproportionation mechanisms [14]. No additional hits of DsrL proteins were observed (BA=1, AC=1, RC=1, PR=1, FDR=0).

The DsrT protein was identified only in DsrAB genomes affiliated with the phylum Chlorobi and in all dSRP possessing the full DsrMKJOP complex. Due to this smaller distribution across DsrAB genomes, the determination of false negatives is influenced leading to a recall of 0.5 (BA=0.75). No false positives were found in genomes devoid of DsrAB proteins and all predicted enzyme types of DsrT proteins were congruent with the DsrAB-type resulting in a precision of 1.0 and an accuracy of 0.99.

The strategy to include also paralogous proteins within the set of models proved to be (partially) functional for the distinction between some protein families, such as the dissimilatory Sat, and AprA and AprB from DsrAB-containing micro-organisms. For the three proteins, around 77 hits were found in genomes containing a DsrAB complex of the same type (BA=0.87). Since not all micro-organisms using this energetic solution have the Sat and AprAB proteins, the number of false positives is overestimated. Of note, while for AprA and AprB proteins, only 7 false positives were found (FDR=0.08), in the case of Sat, 58 additional sequences (FDR=0.43), from related Crenarchaeota, *Mycobacterium* and a few proteobacterial species were also identified.

The AprM protein is a functional alternative of the Qmo complex for interaction with AprAB proteins in dSOB, but AprM is not specific of the Dsr-dependent dissimilatory sulphur metabolism [15, 16]. AprM was found in eleven genomes with oxidative-type DsrAB proteins and oxidative-type AprAB proteins were found to be co-distributed with the AprM protein in all of these genomes. Additional AprM proteins were found in six genomes devoid of DsrAB proteins, in which oxidative-type AprAB proteins were also identified. These cases

were classified as false positive resulting in a precision of 0.65 (FDR=0.35). Although AprM is only present in some dSOB, all genomes with oxidative-type DsrAB proteins lacking AprM were considered as false negatives, which is highly influencing the recall of 0.31. However, the accuracy shows a stable predictability for AprM proteins (AC=0.99, BA=0.65).

Another AprAB's interaction partner is the Qmo complex with its variation of subunits: in some genomes instead of the full QmoABC complex, only the subunits QmoAB are present and QmoC could be replaced by heterodisulphide reductases [7, 15, 17]. Thus, the determined false negatives of QmoC are inflated. In addition, considering dSRP, the QmoABC complex is strictly needed for the complete Dsr-dependent reduction of sulphate to sulphide, but not for Dsr-dependent reduction of sulphite to sulphide [13]. Due to these factors, all statistical measurements are affected by mathematically classified false negatives, which do not reflect the nature of an organism's energy conservation strategy. Overall, QmoABC proteins have a good (high) precision between 0.93 and 1.0, whereas the recall, due to highly overestimated the false negatives, range between 0.4 and 0.57. The accuracy is 0.99 for all three Qmo proteins, while the balanced accuracy is around 0.78 for both QmoAB (FDR=0.0-0.03) and 0.7 for QmoC (FDR=0.07).

We compared DiSCo results of the complete genomes dataset with the ones obtained by BLAST as well as with the rBBH approach. While both DiSCo and the rBBH method identified the DsrAB proteins in 103 micro-organisms, using the defined cutoffs, BLAST led to the identification of 171 assemblies with hits for both DsrAB proteins and 327 additional cases where only one protein was identified. This additional identification of paralogs of DsrAB proteins is not unexpected, since many paralogous sequences share common features, which fulfil the ≥25% local identity cutoff commonly used in BLAST similarity searches. For instance, within those falsely identified protein sequences, we find several anaerobic sulphite reductase, that share the siroheme binding site and the ferredoxin domain with DsrAB proteins. This evolutionary relationships, already identified by [18], are imprinted at the level of the primary sequence and highlights mechanisms of functional diversity that through time, allowed micro-organisms to reuse existing building block, alter them, and evolve the ability to perform different metabolic functions. The reuse of building blocks is a common occurrence in biology, and can be also seen in other protein families, by the high level of false positives identified by BLAST even within completely sequenced genomes (*e.g.* number of false positives identified by BLAST for DsrO: 2,873 (~56% of 5,100 genomes); and DsrL: 3,565 (~70% of 5,100 genomes)). In the cases of the DsrC, DsrM and DsrK proteins, while DiSCo had precision and recall between 1 and 0.94 (FDR=0.03–0.06), in rBBH and BLAST approaches, the identification of many paralogous sequences led to values of precision ranging from 0.81 to 0.25 (FDR=0.75–0.19) in the case of rBBH, and a precision of 0.14 to 0.5 (FDR=0.86–0.50) for the BLAST approach. Due to the identification of highly similar, paralogous sequences, almost no false negatives were identified resulting in a highly overestimated recall for DsrC, DsrMK proteins of 1 in the case of BLAST and 0.99–1 for rBBH (Table S7). Overall, DiSCo highly outperformed both BLAST-based methods in the distinction of orthologous from paralogous proteins in the complete genomes dataset.

References

1. **Feldbauer R, Schulz F, Horn M, Rattei T**. Prediction of microbial phenotypes based on comparative genomics. *BMC Bioinformatics* 2015; 16:S1 DOI:10.1186/1471-2105-16-S14-S1.
2. **Venceslau SS, Stockdreher Y, Dahl C, Pereira IAC**. The 'bacterial heterodisulfide' DsrC is a key protein in dissimilatory sulfur metabolism. *Biochim Biophys Acta* 2014; 1837:1148–1164 DOI:10.1016/j.bbabio.2014.03.007.
3. **Mander GJ, Duin EC, Linder D, Stetter KO, Hedderich R**. Purification and characterization of a membrane-bound enzyme complex from the sulfate-reducing archaeon *Archaeoglobus fulgidus* related to heterodisulfide reductase from methanogenic archaea. *Eur J Biochem*. Epub ahead of print 2002. DOI: 10.1046/j.1432-1033.2002.02839.x DOI:10.1046/j.1432-1033.2002.02839.x.
4. **Santos AA, Venceslau SS, Grein F, Leavitt WD, Dahl C, *et al.*** A protein trisulfide couples dissimilatory sulfate reduction to energy conservation. *Science* 2015; 350:1541–1545 DOI:10.1126/science.aad3558.
5. **Bonch-Osmolovskaya EA, Sokolova TG, Kostrikina NA, Zavarzin GA**. *Desulfurella acetivorans* gen. nov. and sp. nov. - a new thermophilic sulfur-reducing eubacterium. *Arch Microbiol* 1990; 153:151–155 DOI:10.1007/BF00247813.
6. **Florentino AP, Brienza C, Stams AJM, Sánchez-Andrea I**. *Desulfurella amilsii* sp. nov., a novel acidotolerant sulfur-respiring bacterium isolated from acidic river sediments. *Int J Syst Evol Microbiol* 2016; 66:1249–1253 DOI:10.1099/ijsem.0.000866.
7. **Pereira IAC, Ramos AR, Grein F, Marques MC, da Silva SM, *et al.*** A comparative genomic analysis of energy metabolism in sulfate reducing bacteria and archaea. *Front Microbiol* 2011; 2:1–22 DOI:10.3389/fmicb.2011.00069.
8. **Rudenko TS, Tarlachkov S V., Shatskiy ND, Grabovich MY**. Comparative Genomics of *Beggiatoa leptomitoformis* Strains D-401 and D-402T with Contrasting Physiology But Extremely High Level of Genomic Identity. *Microorganisms* 2020; 8:1–11 DOI:10.3390/microorganisms8060928.
9. **Dahl C, Schulte A, Stockdreher Y, Hong C, Grimm F, *et al.*** Structural and Molecular Genetic Insight into a Widespread Sulfur Oxidation Pathway. *J Mol Biol* 2008; 384:1287–1300 DOI:10.1016/j.jmb.2008.10.016.
10. **Kelly DP**. Stable sulfur isotope fractionation by the green bacterium *Chlorobaculum parvum* during photolithoautotrophic growth on sulfide. *Polish J Microbiol* 2008; 57:275–279.
11. **Gregersen LH, Bryant DA, Frigaard N-U**. Mechanisms and Evolution of Oxidative Sulfur Metabolism in Green Sulfur Bacteria. *Front Microbiol* 2011; 2:1–14 DOI:10.3389/fmicb.2011.00116.
12. **Liu LJ, Stockdreher Y, Koch T, Sun ST, Fan Z, *et al.*** Thiosulfate Transfer Mediated by DsrE/TusA Homologs from Acidothermophilic Sulfur-oxidizing Archaeon *Metallosphaera cuprina*. *J Biol Chem* 2014; 289:26949–26959 DOI:10.1074/jbc.M114.591669.
13. **Chernyh NA, Neukirchen S, Frolov EN, Sousa FL, Miroshnichenko ML, *et al.*** Dissimilatory sulfate reduction in the archaeon '*Candidatus* Vulcanisaeta

moutnovskia' sheds light on the evolution of sulfur metabolism. *Nat Microbiol* 2020; 5:1428–1438 DOI:10.1038/s41564-020-0776-z.

14. **Florentino AP, Pereira IAC, Boeren S, van den Born M, Stams AJM, *et al.*** Insight into the sulfur metabolism of *Desulfurella amilsii* by differential proteomics. *Environ Microbiol* 2019; 21:209–225 DOI:10.1111/1462-2920.14442.
15. **Meyer B, Kuever J**. Phylogeny of the alpha and beta subunits of the dissimilatory adenosine-5′ -phosphosulfate (APS) reductase from sulfate-reducing prokaryotes - Origin and evolution of the dissimilatory sulfate-reduction pathway. *Microbiology* 2007; 153:2026–2044 DOI:10.1099/mic.0.2006/003152-0.
16. **Meyer B, Kuever J**. Molecular analysis of the distribution and phylogeny of dissimilatory adenosine-5'-phosphosulfate reductase-encoding genes (*aprBA*) among sulfur-oxidizing. *Microbiology* 2007; 153:3478–3498 DOI:10.1099/mic.0.2007/008250-0.
17. **Junier P, Junier T, Podell S, Sims DR, Detter JC, *et al.*** The genome of the Gram-positive metal- and sulfate-reducing bacterium *Desulfotomaculum reducens* strain MI-1. *Environ Microbiol* 2010; 12:2738–2754 DOI:10.1111/j.1462-2920.2010.02242.x.
18. **Dhillon A, Goswami S, Riley M, Teske A, Sogin ML**. Domain evolution and functional diversification of sulfite reductases. *Astrobiology* 2005; 5:18–29 DOI:10.1089/ast.2005.5.18.

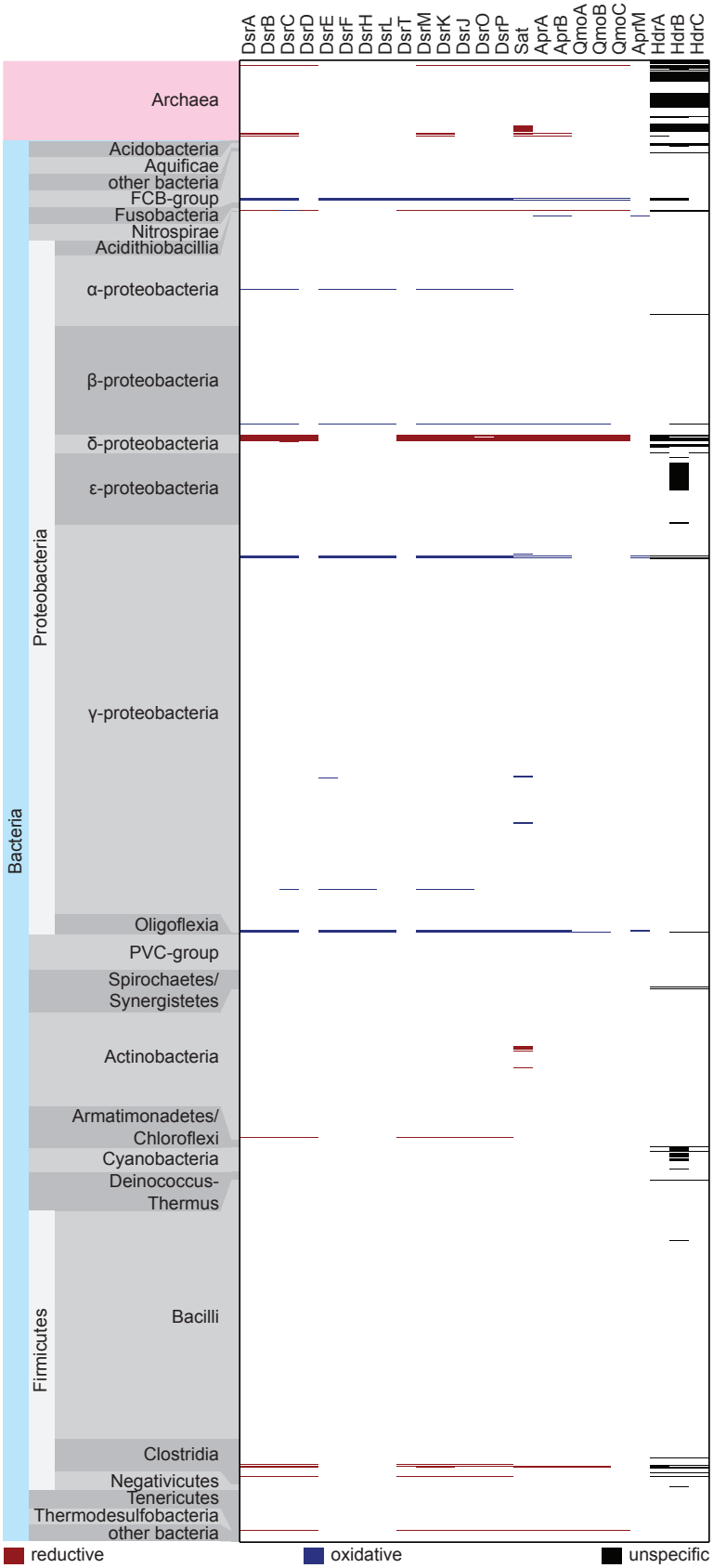

Fig. S1. Distribution of DiSCo hits in all 5,100 genomes across the complete genomes dataset. Each column represents a protein and each row corresponds to a genome. The color code indicates DiSCo predicted enzyme type as in Fig. 4.

Supplementary Tables:

Table S1: Micro-organisms with known phenotype retrieved from the literature search.

Table S2: Complete genomes dataset used in this study. Information regarding taxonomic assignments, download date, database of origin, completeness and redundancy are also included.

Table S3: Metagenomic dataset used in this study. Information regarding taxonomic assignments, download date, database of origin, completeness and redundancy are also included.

Table S4: Selected queries used to search the genomic space.

Table S5: Distribution of sulphur metabolism-related proteins in the complete genomes dataset using similarities. Table representing the distribution of proteins identified within the complete genomes dataset using the rBBH approach followed by a clustering procedure.

Table S6: DiSCo model information. List of protein accessions used to build the models and the profile-specific cutoffs in terms of *E*-value and score.

Table S7: Contingency matrix for type-specific models. Table listing the calculated false positives, true positive, false negatives, and true negatives of DiSCo assignments, rBBH, and BLAST hits in complete genomes and in the independent test set.

Table S8: Jackknife resampling of DsrABCMK models

Table S9: Distribution of DiSCo hits within complete genomes. Matrix showing the results of DiSCo screening across the 5100 complete genomes dataset.

Table S10: Distribution of DiSCo hits within the 195878 metagenomic records.

Table S11: Distribution of DsrABCMK hits within the 195878 genomic records. Only genomes with at least one hit to a DsrA/B/C/M/K protein are represented.

Table S12: Taxonomic summary of DsrAB containing metagenomes. Only genomes with at least one DsrA and/or DsrB hit are considered.

Table S13: Absolute values used to produce Figure 5.