METABOLIC: A scalable high-throughput metabolic and biogeochemical functional trait profiler based on microbial genomes

Zhichao Zhou[1], Patricia Tran[1], Yang Liu[2], Kristopher Kieft[1], Karthik Anantharaman[1,*]

[1] Department of Bacteriology, University of Wisconsin-Madison, Madison, WI, 53706, USA,

[2]Institute for Advanced Study, Shenzhen University, Shenzhen, Guangdong Province, 518060, China

*To whom correspondence should be addressed.

Abstract

Summary: Microbial metabolism mediates fundamental transformations of chemistry and energy that drive biogeochemical cycles on our planet. Increasingly**,** we can read genomic blueprints of microorganisms, decipher their functional capacities and activities, and reconstruct their roles in biogeochemical processes using omic-based techniques such as metagenomics. Currently available tools for analyses of genomic data can annotate and depict metabolic functions to some extent, but they are not comprehensive. No standardized approaches are currently available for bioinformatic validation of metabolic predictions and identifying contributions of microorganisms and genes to biogeochemical cycles. Here we present METABOLIC (**METa**bolic **A**nd **B**ioge**O**chemistry ana**L**yses **I**n mi**C**robes), a scalable metabolic and biogeochemical functional trait profiler to comprehensively study microbial metabolism using genome data. METABOLIC uses metagenome-assembled (MAG), single-cell (SAG), or isolate genomes as input, annotates and processes genomes for identification and characterization of metabolism markers using KEGG and curated custom protein HMM databases, and applies motif confirmation of biochemically validated conserved residues in proteins. The output report includes functionally important HMM hit tables, protein collections for downstream analysis, tables (KEGG modules) and diagrams representing metabolic pathways for individual genomes, and a summary figure representing selected biogeochemical cycling processes on a community scale. We expect that METABOLIC will facilitate the study of genome-informed microbial metabolism and biogeochemistry and transform our understanding of environmental microbiomes.

Availability and implementation: METABOLIC is available on github: https://github.com/AnantharamanLab/METABOLIC.

Contact: karthik@bact.wisc.edu

1. Introduction

Microbially-mediated biogeochemical processes serve as important driving forces for transformation and cycling of elements, energy, and matter among the lithosphere, atmosphere, hydrosphere and biosphere (Madsen, 2011). Metagenomics and single-cell genomics have transformed the field of microbial ecology by revealing a rich diversity of microorganisms from diverse settings, including terrestrial and marine environments and human body (Anantharaman, et al., 2016; Dombrowski, et al., 2018; Parks, et al., 2017; Pasolli, et al., 2019). These approaches can provide an unbiased and insightful view into microorganisms mediating and contributing to the biogeochemical activities at a number of scales ranging from individual organisms to communities (Anantharaman, et al., 2016; Bowers, et al., 2017; Hug, et al., 2016; Parks, et al., 2017). Prediction of microbial metabolism relies on annotation of protein function for microorganisms using a number of established databases, e.g., KEGG (Kanehisa and Goto, 2000), MetaCyc (Caspi, et al., 2006), Pfam (Finn, et al., 2014), TIGRfam (Selengut, et al., 2007), SEED (Overbeek, et al., 2013), and eggNOG (Huerta-Cepas, et al., 2016). However, these results are often highly detailed. Obtaining a functional profile and identifying metabolic pathways in a microbial genome can involve manual inspection of thousands of genes. Interpreting, organizing and visualizing such datasets remains a challenge and is often untenable, and there is a critical need for a tool to identify and validate the presence of metabolic pathways and genes of biogeochemical function in a user-friendly manner. Such a tool would also allow standardization and easy integration of genome-informed metabolism into biogeochemical models which currently rely primarily on physico-chemical data and treats microorganisms as black boxes.

Here we present the software METABOLIC, a tool to profile metabolic and biogeochemical functional traits based on microbial genomes. It integrates annotation of proteins using KEGG (Kanehisa and Goto, 2000), TIGRfam (Selengut, et al., 2007), Pfam (Finn, et al., 2014), and custom HMM databases (Anantharaman, et al., 2016), incorporates a motif validation step to accurately identify proteins based on prior biochemical validation, determines presence or absence of metabolic pathways based on KEGG modules, and produces user-friendly outputs in the form of tables and figures including a summary of biogeochemically-relevant pathways and their abundance for individual genomes and at the community scale.

2. Methods

METABOLIC is written in Perl and R and is expected to run in Unix/Linux and MacOS. The prerequisites are described on METABOLIC's GitHub page (https://github.com/AnantharamanLab/METABOLIC). The input folder requires microbial genome sequences in FASTA format and an optional set of metagenomic reads in which were used to reconstruct those genomes (Supplementary Figure S1). Genomic sequences are annotated by Prodigal (Hyatt, et al., 2010), or a user can provide self-annotated proteins (with extensions of ".faa") in order to facilitate incorporation into existing pipelines. Proteins will be queried against hidden Markov model (HMM) databases using hmmsearch implemented within HMMER (Finn, et al., 2011) which implements methods to detect remote homologs as sensitively and efficiently as possible. The HMM databases include Kofam prokaryotic (KEGG) (Aramaki, et al., 2019), TIGRfam (Selengut, et al., 2007), Pfam (Finn, et al., 2014) and custom metabolic HMM files (Anantharaman, et al., 2016). The cutoff threshold values for HMM databases were used as follows: Kofam - Kofam suggested values; TIGRfam/Pfam/Custom databases - Manually curated by adjusting noise cutoffs (NC) and trusted cutoffs (TC) to avoid potential false positive hits; detailed curation methods are described previously (Anantharaman, et al., 2016).

To computationally validate protein hits and avoid false positives, we have introduced a motif validation step that including comparison of protein motifs against a manually curated set of highly conserved residues in important proteins. As an example, DsrC (sulfite reductase subunit C) and TusE (tRNA 2-thiouridine synthesizing protein E) are similar proteins that are routinely misannotated. Both are assigned to the family KO:K11179 in the KEGG database. To avoid assigning TusE as a sulfite reductase, we identified a specific motif for DsrC but not TusE (GPXKXXCXXXGXPXPXXCX" where "X" stands for any amino acid) (Venceslau, et al., 2014). We use these specific motifs to filter for proteins which have high sequence similarity but functionally divergent homologs.

The software output integrates the presence and absence of genes from the outputs of individual HMM runs and relates them to microbial functional traits. Individual KEGG annotations are inferred in the context of KEGG modules for better interpretation of metabolic pathways. A KEGG module is a collection of manually defined functional units. A module is comprised of multiple steps with each step representing a distinct metabolic function. Since genomes can often have incomplete metabolic pathways, we determine the completeness of specific metabolic pathways by parsing KEGG module IDs. A user-defined cutoff is used to estimate the completeness of a given module (the default value is 75%), which is then used to produce KEGG module presence/absence table. All modules exceeding the cutoff are determined to be complete in the given genome. Outputs consist of four different results that are reported in an Excel spreadsheet (Supplementary Figure S2). These contain details of HMM hits (Supplementary Figure S2A), presence/absence of functional traits (Supplementary Figure S2B), presence/absence of KEGG modules (Supplementary Figure S2C), and presence/absence of KEGG module steps (Supplementary Figure S2D). Each collection of HMM hits can be extracted from input genomes for the downstream phylogenetic analysis. A detailed workflow of METABOLIC is available in Supplementary Figure S1.

To visualize pathways of biogeochemical importance, the software draws schematic profiles for nitrogen, carbon, sulfur and other element cycles for each genome. A summary schematic diagram at the scale of a microbial community integrates results from all genomes from a given dataset (Figure 1) and includes computed abundances for each step in a biogeochemical cycle if the metagenomic reads datasets are provided.

3. Results

METABOLIC has been successfully applied on a metagenomic dataset which includes 98 MAGs from a deep-sea hydrothermal plume at Guaymas Basin in the Pacific Ocean, and two sets of metagenomic reads (that are subsets of original reads with 10 million read numbers for each pair comprising ~10% of the total reads). The total run time was ~8 hours using 25 CPU threads in a Linux version 4.15.0-48-generic server (Ubuntu v5.4.0). The resulting summary scheme on various biogeochemical cycling processes reflects the pattern on a community scale (Figure 1) (Supplementary Data S1 contains tables and figures from the METABOLIC output).

In order to test the accuracy of the results predicted by METABOLIC, we picked 15 bacterial and archaeal genomes from *Chloroflexi*, *Thaumarchaeota*, and *Crenarchaeota* which are reported to have 3 Hydroxypropionate cycle (3HP) or 3-hydroxypropionate/4-hydroxybutyrate cycle (3HP/4HB) for carbon fixation. METABOLIC predicts results in line with KEGG genome database annotations and can also be visualized with the KEGG Mapper (Supplementary Table S1). Our predictions are also in accord with biochemical evidence of the existence of corresponding carbon fixation pathways in each microbial group: only organisms from the phylum *Chloroflexi* are known to possess the 3HP pathway and 3HP/4HB pathway could only be detected in *Crenarchaeota* and *Thaumarchaeota* (Supplementary Table S1 and references therein). These results suggest that METABOLIC can provide accurate annotations and genomic profiles of metabolism and serve as a good functional predictor for microbial genomes at the individual and community scales.

Funding

We thank the University of Wisconsin - Office of the Vice Chancellor for Research and Graduate Education, University of Wisconsin – Department of Bacteriology, Wisconsin Alumni Research Foundation, and University of Wisconsin – College of Agriculture and Life Sciences for their support.

Conflict of interest: none declared.

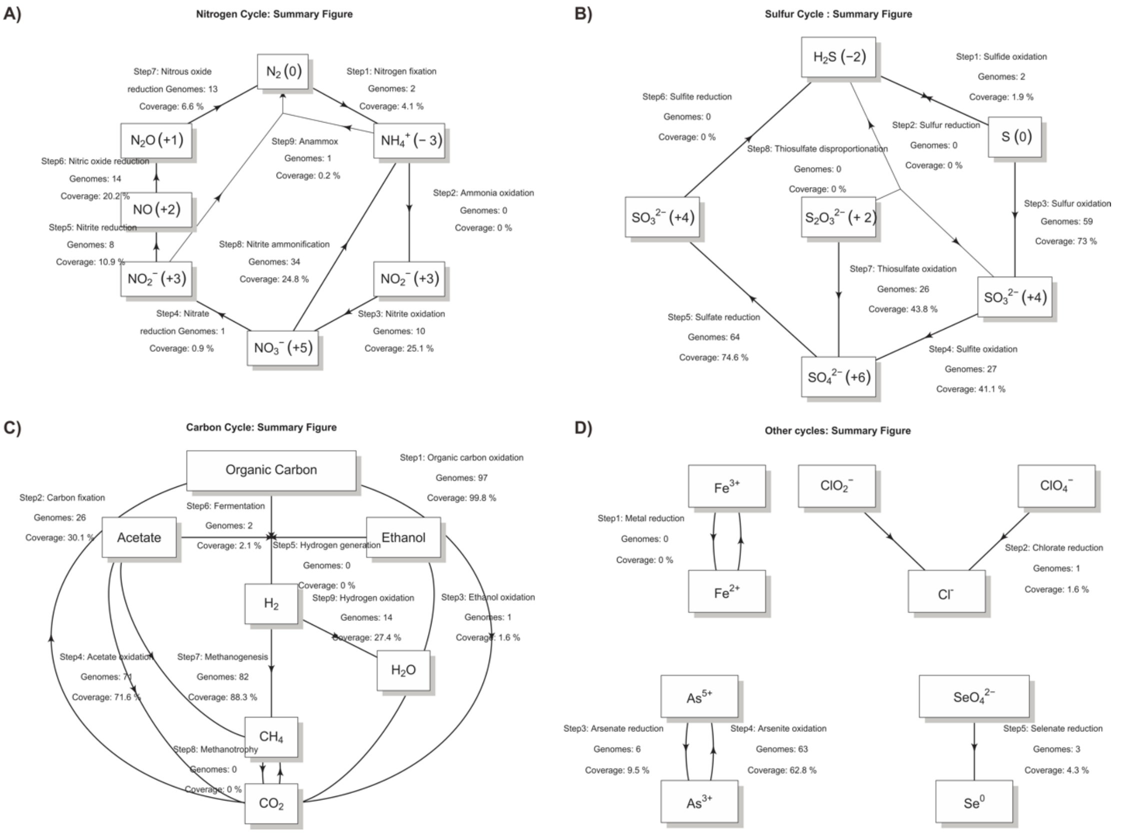
A)
Nitrogen Cycle: Summary Figure
N_2 (0)
Step7: Nitrous oxide reduction Genomes: 13 Coverage: 6.6 %
Step1: Nitrogen fixation Genomes: 2 Coverage: 4.1 %
N_2O (+1)
NH_4^+ (− 3)
Step9: Anammox Genomes: 1 Coverage: 0.2 %
Step6: Nitric oxide reduction Genomes: 14 Coverage: 20.2 %
NO (+2)
Step2: Ammonia oxidation Genomes: 0 Coverage: 0 %
Step5: Nitrite reduction Genomes: 8 Coverage: 10.9 %
Step8: Nitrite ammonification Genomes: 34 Coverage: 24.8 %
NO_2^- (+3)
NO_2^- (+3)
Step4: Nitrate reduction Genomes: 1 Coverage: 0.9 %
Step3: Nitrite oxidation Genomes: 10 Coverage: 25.1 %
NO_3^- (+5)
B)
Sulfur Cycle : Summary Figure
H_2S (−2)
Step1: Sulfide oxidation Genomes: 2 Coverage: 1.9 %
Step6: Sulfite reduction Genomes: 0 Coverage: 0 %
Step2: Sulfur reduction Genomes: 0 Coverage: 0 %
S (0)
Step8: Thiosulfate disproportionation Genomes: 0 Coverage: 0 %
SO_3^{2-} (+4)
$S_2O_3^{2-}$ (+ 2)
Step3: Sulfur oxidation Genomes: 59 Coverage: 73 %
Step7: Thiosulfate oxidation Genomes: 26 Coverage: 43.8 %
SO_3^{2-} (+4)
Step5: Sulfate reduction Genomes: 64 Coverage: 74.6 %
Step4: Sulfite oxidation Genomes: 27 Coverage: 41.1 %
SO_4^{2-} (+6)
C)
Carbon Cycle: Summary Figure
Organic Carbon
Step1: Organic carbon oxidation Genomes: 97 Coverage: 99.8 %
Step2: Carbon fixation Genomes: 26 Coverage: 30.1 %
Step6: Fermentation Genomes: 2 Coverage: 2.1 %
Acetate
Ethanol
Step5: Hydrogen generation Genomes: 0 Coverage: 0 %
H_2
Step9: Hydrogen oxidation Genomes: 14 Coverage: 27.4 %
Step3: Ethanol oxidation Genomes: 1 Coverage: 1.6 %
Step4: Acetate oxidation Genomes: 71 Coverage: 71.6 %
Step7: Methanogenesis Genomes: 82 Coverage: 88.3 %
H_2O
CH_4
Step8: Methanotrophy Genomes: 0 Coverage: 0 %
CO_2
D)
Other cycles: Summary Figure
Fe^{3+}
ClO_2^-
ClO_4^-
Step1: Metal reduction Genomes: 0 Coverage: 0 %
Step2: Chlorate reduction Genomes: 1 Coverage: 1.6 %
Fe^{2+}
Cl^-
As^{5+}
SeO_4^{2-}
Step3: Arsenate reduction Genomes: 6 Coverage: 9.5 %
Step4: Arsenite oxidation Genomes: 63 Coverage: 62.8 %
Step5: Selenate reduction Genomes: 3 Coverage: 4.3 %
As^{3+}
Se^0

Figure 1. The summary scheme of biogeochemical cycling processes on a community scale. Above each arrow (which represent each step within a cycle) there are three lines. The first one indicates the step name and the reaction, the second one indicates the number of genomes that acquire these reactions, the third one indicates the percentage of metagenomic coverage on each step.

References

Anantharaman, K.*, et al.* (2016) Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system, *Nat. Commun.*, **7**, 13219.

Aramaki, T.*, et al.* (2019) KofamKOALA: KEGG ortholog assignment based on profile HMM and adaptive score threshold, *bioRxiv*, 602110.

Bowers, R.M.*, et al.* (2017) Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea, *Nat. Biotechnol.*, **35**, 725.

Caspi, R.*, et al.* (2006) MetaCyc: a multiorganism database of metabolic pathways and enzymes, *Nucleic Acids Res.*, **34**, D511-516.

Dombrowski, N., Teske, A.P. and Baker, B.J. (2018) Expansive microbial metabolic versatility and biodiversity in dynamic Guaymas Basin hydrothermal sediments, *Nat. Commun.*, **9**, 4999.

Finn, R.D.*, et al.* (2014) Pfam: the protein families database, *Nucleic Acids Res.*, **42**, D222-230.

Finn, R.D., Clements, J. and Eddy, S.R. (2011) HMMER web server: interactive sequence similarity searching, *Nucleic Acids Res.*, **39**, W29-37.

Huerta-Cepas, J.*, et al.* (2016) eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences, *Nucleic Acids Res.*, **44**, D286-D293.

Hug, L.A.*, et al.* (2016) A new view of the tree of life, *Nat. Microbiol.*, **1**, 16048.

Hyatt, D.*, et al.* (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification, *BMC Bioinformatics*, **11**, 119.

Kanehisa, M. and Goto, S. (2000) KEGG: kyoto encyclopedia of genes and genomes, *Nucleic Acids Res.*, **28**, 27-30.

Madsen, E.L. (2011) Microorganisms and their roles in fundamental biogeochemical cycles, *Curr. Opin. Biotechnol.*, **22**, 456-464.

Overbeek, R.*, et al.* (2013) The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST), *Nucleic Acids Res.*, **42**, D206-D214.

Parks, D.H.*, et al.* (2017) Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life, *Nat. Microbiol.*, **2**, 1533-1542.

Pasolli, E.*, et al.* (2019) Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle, *Cell*, **176**, 649-662 e620.

Selengut, J.D.*, et al.* (2007) TIGRFAMs and Genome Properties: tools for the assignment of molecular function and biological process in prokaryotic genomes, *Nucleic Acids Res.*, **35**, D260-264.

Venceslau, S.S.*, et al.* (2014) The "bacterial heterodisulfide" DsrC is a key protein in dissimilatory sulfur metabolism, *Biochim. Biophys. Acta*, **1837**, 1148-1164.