metabolisHMM: Phylogenomic analysis for exploration of microbial phylogenies and metabolic pathways

McDaniel, E.A.[1*], Anantharaman, K.[1], McMahon, K.D.[1,2]

[1] Department of Bacteriology, University of Wisconsin – Madison, Madison, WI

[2] Department of Civil and Environmental Engineering, University of Wisconsin – Madison, Madison, WI

[*] Corresponding Author: elizabethmcd93@gmail.com

Abstract

Summary:

Advances in high-throughput sequencing technologies and bioinformatic pipelines have exponentially increased the amount of data that can be obtained from uncultivated microbial lineages inhabiting diverse ecosystems. Various annotation tools and databases currently exist for predicting the functional potential of sequenced genomes or microbial communities based upon sequence identity. However, intuitive, reproducible, and user-friendly tools for further exploring and visualizing functional guilds of microbial community metagenomic sequencing datasets remains lacking. Here, we present metabolisHMM, a series of workflows for visualizing the distribution of curated and user-provided Hidden Markov Models (HMMs) to understand metabolic characteristics and evolutionary histories of microbial lineages. metabolisHMM performs functional annotations with a set of curated or user-defined HMMs to 1) construct ribosomal protein and single marker gene phylogenies, 2) summarize the presence/absence of metabolic pathway markers, and 3) create heatmap visualizations of presence/absence summaries.

Availability and Implementation: metabolisHMM is freely available on Github at https://github.com/elizabethmcd/metabolisHMM and on PyPi at https://pypi.org/project/metabolisHMM/ under the GNU General Public License v3.0.

Contact: elizabethmcd93@gmail.com

1. Introduction

Common comparative genomic approaches for analyzing large metagenomic datasets include analyzing the distribution and evolutionary history of genes of interest, describing the presence/absence of specific metabolic pathways in metagenome assembled genomes (MAGs) or single cell genomes (SAGs), and comparing these results to existing publicly available genomes. Many of these steps require computational expertise in several bioinformatic tools, specific file formats, and sometimes use of expensive, proprietary software platforms. Tools for intuitively summarizing and visualizing the functional potential of sequenced genomes in a high-throughput, user-friendly, reproducible manner that allow for maximum user flexibility (i.e. custom marker sets) and making comparisons among large genome datasets are overall lacking. Here we present metabolisHMM, a set of reproducible workflows for executing common comparative genomics analyses using profile Hidden Markov Models (HMMs). metabolisHMM encompasses a set of easy-to-use and flexible workflows for visualizing phylogenies and metabolic heatmaps from both curated and custom-provided HMM-based profile annotations. We demonstrate the capabilities and output results of metabolisHMM by using publicly available bacterial and archaeal genomes from a subsurface aquifer system (1), available as an online tutorial at https://github.com/elizabethmcd/metabolisHMM/wiki/Subsurface-Aquifer-Tutorial.

2. Implementation

The metabolisHMM package contains two main functionalities: 1) phylogeny construction to visualize evolutionary histories and 2) heatmap-based synthesis of metabolic pathway distributions. The basic input requirements for each of the four embedded workflows are DNA sequences as raw genomic scaffolds in fasta file format for subsequent gene prediction using

72 Prodigal (2). Each workflow uses either curated or user-provided custom profile Hidden Markov
73 Models (HMMs) with a threshold cutoff score provided by the user for performing functional
74 annotation, based on the hmmsearch option of HMMER (3). Additionally, the user defines an
75 output directory in which all intermediate files such as reformatted fasta files, HMM output results,
76 alignments, and final phylogenies and heatmap figures are deposited. The remaining arguments
77 and steps are workflow dependent. Detailed installation instructions and documentation for using
78 the metabolisHMM package and each workflow is provided in the repository wiki:
79 https://github.com/elizabethmcd/metabolisHMM/wiki.

80 ***2.1 Constructing single marker phylogenies***

81 The single-marker-phylogeny workflow searches a panel of genomes for a specific gene
82 marker and builds a phylogenetic tree. Any of the package-provided curated marker sets or a user-
83 provided marker can be used for constructing a single-marker phylogeny. The alignment is
84 constructed using MAFFT (4), and the user can choose to construct the phylogenetic tree using
85 either FastTree (5) or RAxML (6), depending on available computational resources. Due to
86 common issues with MAG gene content redundancy and unknown consequences of copy number
87 variation from uncultivated organisms, metabolisHMM only uses the top-scoring hit for a
88 particular marker within a genome for constructing the final alignment and phylogeny. Given a
89 corresponding metadata file, the user can output data files configured for viewing trees with the
90 interactive Tree Of Life (iTOL) online tool (7).

91 ***2.2 Creating genome phylogenies***

92 The create-genome-phylogeny workflow takes a set of input genomes and creates a
93 ribosomal phylogeny or species tree. We provide a set of 16 single copy ribosomal proteins as part
94 of the metabolisHMM software package release that are specific for archaea or bacteria (32

95 markers total) as described in Hug et al. (8). Alignments and tree construction are performed as
96 described above, with individual alignments concatenated across all genomes. Since
97 metabolisHMM was developed specifically for comparing MAGs and SAGs against isolate
98 genomes, metabolisHMM will warn the user if a genome contains less than 12 or a pre-defined
99 value of ribosomal markers, as confidence in the phylogenetic reconstruction will be low if a
100 genome is missing several markers in the final alignment, due to incompleteness.

101 ***2.3 Summarizing broad metabolic features using curated and custom markers***

102 The summarize-metabolism workflow uses a set of manually curated profiles spanning
103 major transformations in the carbon, nitrogen, sulfur, and hydrogen cycles, that were constructed
104 and made publicly available by Anantharaman et al. (1). Marker descriptions are provided in the
105 ancillary data files of the software distribution. The user also provides a metadata file containing
106 either the specific taxonomical names for each genome, or broad groups by which to aggregate
107 sets of genomes together, such as by phylum-level placement or sample origin. Any marker-
108 genome pair with a value greater than 1 is changed to a value of 1, resulting in a table of 0's and
109 1's for the absence and presence, respectively, of every marker-genome pair. The resulting
110 heatmap shows the presence/absence of all markers spanning broad biogeochemical cycles to show
111 the overall functional guilds of the input genomes. In addition to visualizing curated marker sets
112 provided with the metabolisHMM package, the user can specify any marker sets that are custom-
113 made or from outside databases, such as the PFAM and TIGRFAM databases, and/or the recently
114 released KofamKOALA distribution. (9). The search-custom-markers workflow takes a set of
115 specified markers in a user-provided order and produces a heatmap similar to that of the broad
116 summaries mentioned above.

117 **3. Results and Assessment**

118 To demonstrate the main features of the metabolisHMM workflow, we used a set of 2545
119 publicly available bacterial and archaeal genomes from an aquifer metagenomic dataset (1). All
120 demo figures are available within the tutorial at
121 https://github.com/elizabethmcd/metabolisHMM/wiki/Subsurface-Aquifer-Tutorial. Using the
122 single-marker-phylogeny workflow, we created a phylogeny of the folD marker, part of the
123 reductive acetyl-coA pathway (10). We created a corresponding ribosomal phylogeny of genomes
124 containing the folD marker. Using the FastTree option for constructing phylogenies, to search all
125 2,545 genomes for the folD marker, construct the phylogeny of the single marker, and make a
126 corresponding ribosomal phylogeny of the 610 hits was completed in less than 30 minutes using 1
127 thread on a standard laptop (2015 MacBook Pro). We then characterized the broad metabolic
128 capabilities of a subset of groups of MAGs within the aquifer dataset. Genomes were aggregated
129 by phylum or superphylum group, where the shade of the cell for a specific marker indicates the
130 percentage of genomes within that group that contain each marker. To screen the 874 genomes for
131 all 80 curated markers, this workflow completed in approximately 1 hour using 1 thread. Using
132 the search-custom-markers workflow, we screened 874 genomes for the main steps and subunits
133 that are part of the methyl and carbonyl branches of the reductive acetyl-CoA cycle (11). Markers
134 were accessed from the KofamKOALA KEGG distribution of HMMs and the corresponding
135 threshold cutoffs for each marker was used as suggested (9). For screening 874 genomes with 15
136 markers this workflow completed in less than 30 minutes using 1 thread.

137 We compared the main functionalities and unique capabilities included in metabolisHMM
138 with other recently released and popular software pipelines used for visualizing various functional
139 aspects of sequenced genomes (Table 1). This includes GtoTree, MetaSanity, METABOLIC,
140 KEGG Decoder, and Anvi'o (12–16). Overall, the core functionalities of metabolisHMM are

141 distributed among several existing pipelines. However, metabolisHMM allows for maximum
142 flexibility concerning external HMM profiles, making a corresponding ribosomal phylogeny of
143 genomes with a particular single marker, and customized groupings and orderings of heatmap
144 visualizations. Additionally, metabolisHMM allows for all of these core functions and powerful
145 customized options through simple workflows that are easy to install and use reproducibly.

Tool	Phylogenies of single markers	Phylogenomics/ phylogenies	Corresponding ribosomal tree of single marker hits	Curated metabolic markers/summaries	Custom marker input options	Heatmap visualizations	Custom group aggregating/row ordering
GtoTree	✘	✔	✘	*NA*	✔	*NA*	*NA*
MetaSanity	*NA*	✔	*NA*	✔	✔	✔	✔
METABOLIC	*NA*	*NA*	*NA*	✔	✘	✘	✘
KEGG Decoder	*NA*	*NA*	*NA*	✔	✘	✔	✘
Anvi'o	✘	✔	✘	*NA*	✘	✘	✔
metabolisHMM	✔	✔	✔	✔	✔	✔	✔

146

147 **Table 1: Comparison of metabolisHMM functionalities with other pipelines.** If a particular
148 software pipeline was not intended for certain functions, we denoted that with NA. Pipelines
149 encompassing a functionality but does not include flexible or customizable options, for example,
150 are denoted with an X. Packages with a comparable functionality to metabolisHMM are denotated
151 with a ✔.

152 ## Acknowledgements

153 We would like to thank members of the McMahon lab for testing and providing feedback
154 on the metabolisHMM package. We would like to specifically thank Dr. Sarah Stevens of the
155 Data Science Hub at the University of Wisconsin – Madison for early feedback and
156 troubleshooting.

157

158

159

References

1. Anantharaman K, Brown CT, Hug LA, Sharon I, Castelle CJ, Probst AJ, Thomas BC, Singh A, Wilkins MJ, Karaoz U, Brodie EL, Williams KH, Hubbard SS, Banfield JF. 2016. Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. Nat Commun 7:13219.
2. Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11:119.
3. Eddy SR. 2011. Accelerated Profile HMM Searches. PLoS Comput Biol 7:e1002195.
4. Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 30:3059–66.
5. Price MN, Dehal PS, Arkin AP. 2010. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. PLoS One 5:e9490.
6. Stamatakis A. 2014. The RAxML v8 . 0 . X Manual 1–55.
7. Letunic I, Bork P. 2016. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res 44:W242–W245.
8. Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, Butterfield CN, Hernsdorf AW, Amano Y, Ise K, Suzuki Y, Dudek N, Relman DA, Finstad KM, Amundson R, Thomas BC, Banfield JF. 2016. A new view of the tree of life. Nat Microbiol 1:16048.
9. Aramaki T, Blanc-Mathieu R, Endo H, Ohkubo K, Kanehisa M, Goto S, Ogata H. 2019. KofamKOALA: KEGG ortholog assignment based on profile HMM and adaptive score

183 threshold. bioRxiv 602110.
184 10. Adam PS, Borrel G, Gribaldo S. 2018. Evolutionary history of carbon monoxide
185 dehydrogenase/acetyl-CoA synthase, one of the oldest enzymatic complexes. Proc Natl
186 Acad Sci U S A 115:E1166–E1173.
187 11. Zhuang W-Q, Yi S, Bill M, Brisson VL, Feng X, Men Y, Conrad ME, Tang YJ, Alvarez-
188 Cohen L. 2014. Incomplete Wood-Ljungdahl pathway facilitates one-carbon metabolism
189 in organohalide-respiring Dehalococcoides mccartyi. Proc Natl Acad Sci U S A
190 111:6419–24.
191 12. Lee MD. 2019. GToTree: a user-friendly workflow for phylogenomics. Bioinformatics.
192 13. Neely CJ, Graham ED, Tully BJ. 2019. MetaSanity: An integrated, customizable
193 microbial genome evaluation and annotation pipeline. bioRxiv 789024.
194 14. Zhou Z, Tran P, Liu Y, Kieft K, Anantharaman K. 2019. METABOLIC: A scalable high-
195 throughput metabolic and biogeochemical functional trait profiler based on microbial
196 genomes. bioRxiv 761643.
197 15. Tully BJ, Wheat CG, Glazer BT, Huber JA. 2018. A dynamic microbial community with
198 high functional redundancy inhabits the cold, oxic subseafloor aquifer. ISME J 12:1–16.
199 16. Eren AM, Esen ÖC, Quince C, Vineis JH, Morrison HG, Sogin ML, Delmont TO. 2015.
200 Anvi'o: an advanced analysis and visualization platform for 'omics data. PeerJ 3:e1319.
201