Computational and Structural Biotechnology Journal 20 (2022) 1481–1486

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

KEMET – A python tool for KEGG Module evaluation and microbial genome annotation expansion

Matteo Palù [a,1], Arianna Basile [a,1], Guido Zampieri [a,*], Laura Treu [a,**], Alessandro Rossi [a], Maria Silvia Morlino [a], Stefano Campanaro [a,b]

[a] Department of Biology, University of Padova, Via U. Bassi 58/b, 35121 Padova, Italy
[b] CRIBI Biotechnology Center, University of Padova, 35131 Padova, Italy

ARTICLE INFO

Article history:
Received 8 November 2021
Received in revised form 17 March 2022
Accepted 18 March 2022
Available online 26 March 2022

Keywords:
Gene annotation
Microbial genome
Metabolic pathway
Hidden Markov model
Genome-scale metabolic model

ABSTRACT

Background: The rapid accumulation of sequencing data from metagenomic studies is enabling the generation of huge collections of microbial genomes, with new challenges for mapping their functional potential. In particular, metagenome-assembled genomes are typically incomplete and harbor partial gene sequences that can limit their annotation from traditional tools. New scalable solutions are thus needed to facilitate the evaluation of functional potential in microbial genomes.
Methods: To resolve annotation gaps in microbial genomes, we developed KEMET, an open-source Python library devised for the analysis of Kyoto Encyclopedia of Genes and Genomes (KEGG) functional units. KEMET focuses on the in-depth analysis of metabolic reaction networks to identify missing orthologs through hidden Markov model profiles.
Results: We evaluate the potential of KEMET for expanding functional annotations by simulating the effect of assembly issues on real gene sequences and showing that our approach can identify missing KEGG orthologs. Additionally, we show that recovered gene annotations can sensibly increase the quality of draft genome-scale metabolic models obtained from metagenome-assembled genomes, in some cases reaching the accuracy of models generated from complete genomes.
Conclusions: KEMET therefore allows expanding genome annotations by targeted searches for orthologous sequences, enabling a better qualitative and quantitative assessment of metabolic capabilities in novel microbial organisms.

1. Introduction

Metagenomics investigates environmental, engineered, and host-associated microbiomes, stimulating new fast-growing applications in biomedicine and biotechnology [1,2]. The shift towards a holistic approach in microbiome studies can uncover biological activities emerging from synergistic cooperation of microorganisms [3]. Many environments are now being inspected to decipher inhabiting microbial communities, with the aim of predicting their functions and interactions. Thanks to recent improvements in genome-resolved metagenomics, the recovery of metagenome-assembled genomes (MAGs) of high quality is becoming accessible and fast [4]. Functional analysis of genomes derived from metagenomic approaches allows estimating the metabolic potential of species present in a given microbiota. Several dedicated databases, such as the Kyoto Encyclopedia of Genes and Genomes (KEGG), are used as knowledgebases for metabolic pathway inference and reconstruction [5], while tools such as KEGG Mapper [6] and eggNOG-Mapper [7] can assign open reading frames to their function and predict metabolic capabilities at the genome level. However, newly generated metagenomes contain a large number of poorly characterized species, which can be hardly annotated exhaustively with traditional tools [8].

Moreover, genome-scale metabolic models (GSMMs) are now starting to be applied on a metagenome scale [3,9]. GSMM are directly informed by annotation databases and can be automatically reconstructed using tools like CarveMe [10] or gapseq [11]. Such models are useful to infer interactions among microbial species, but the application to uncultured and non-model species can be challenging. In fact, MAG-based GSMMs are especially prone to

* Corresponding author.
** Co-corresponding author.
E-mail addresses: guido.zampieri@unipd.it (G. Zampieri), laura.treu@unipd.it (L. Treu).
[1] Equally contributing authors.

https://doi.org/10.1016/j.csbj.2022.03.015

reconstruction errors due to the gapped nature of metagenomic assemblies. Starting from GSMM reconstructions, several algorithms for network gap-fill enable *in silico* growth simulation and phenotype data fitting. Nevertheless, reactions added this way are not always supported by genomic evidence [12], possibly resulting in erroneous predictions.

To obtain a more exhaustive functional annotation of microbial genomes and improve associated GSMMs, we present KEMET. KEMET - KEgg Module Evaluation Tool - is a command-line, open-source Python toolbox aiming at summarizing and expanding KEGG annotation by comparing microbial sequences to orthologs with curated annotations. With KEMET, annotation recovery from trusted knowledgebases can strengthen the biological fidelity and phenotype prediction in GSMMs and lower the manual refinement effort.

2. Methods and implementation

Starting from genome sequences and associated KEGG annotations, KEMET serves three main goals: functional annotation evaluation, HMM-driven ortholog search in the original sequence, and integration of the corresponding metabolic reactions into GSMMs (Fig. 1). KEMET is a system-independent tool and every dependency is available to UNIX-based and Windows systems. KEMET is freely available and can be downloaded from the GitHub page https://github.com/Matteopaluh/KEMET, where all the procedures to reproduce the results presented in this manuscript are available.

2.1. Module completeness evaluation

The evaluation of metabolic functions present in microbial genomes of interest is performed according to KEGG Modules [5], which consist of manually curated logical expressions of ortholog genes defining the biochemical steps (blocks) of a given function. Functional annotations deriving from different software can directly serve as input data for the Module completeness evaluation, allowing for a flexible downstream implementation of KEMET on pre-existing pipelines. Examples of the supported input files are available in the "toy" folder of the dedicated GitHub repository. At the present time, eggNOG-mapper [7], KofamKOALA [13], and KAAS [14] annotations are supported, and they can be selected through the *-a* parameter. Blocks having KEGG Ortholog (KO) annotations can be identified in target genomes by running KEMET, which allows scaling up the analysis to hundreds of MAGs. Present or missing ortholog blocks in the original annotation can be identified by querying files with KO Module structures. This analysis brings a considerable advantage with respect to the use of KEGG tools alone, allowing to point out single missing orthologs, thus aiding in targeted queries regarding metabolic capabilities of input genomes. The output includes a human-readable tabular file and a flat file indicating the sequential position of missing KOs.

To implement this feature, KEGG Module files are downloaded via the KEGG application-programming interface (API) and parsed to generate intermediate files (*<module_id>.kk* files in the GitHub repository) that are used as Module block structure templates and queried during script usage. The logic behind the block structure in KEMET is devised so as to better identify missing orthologs connected to a single biochemical step. Specifically, the number of blocks in a Module is given by the highest number of individual KOs involved in any alternative reaction path. For example, in Module M00308 the terminal glyceraldehyde-3-phosphate conversion can either be performed via a mechanism involving two KO genes, or via a single dehydrogenase ortholog. While in KEGG Mapper these KOs belong to a single block, KEMET decomposes the longer path into two blocks. These alternative algorithms lead to the same results in terms of numbers of missing orthologs but can give slightly different results when the Module completeness is inspected, as shown in the Results and discussion Section.

2.2. Identification of missing KEGG orthologs

KOs missing from functional annotation can result in incomplete KEGG Modules. This phenomenon can be due to real biological gaps in the species metabolic potential, gene truncation resulting from gaps in the assembly, or limitations of the functional annotation procedure. Missing genes can be sought more in-depth in the genomic sequences, using nucleotidic hidden Markov models (HMM) automatically generated by KEMET, when the *--hmm_mode* parameter is indicated. KEMET has different options for HMM profile generation and for missing KO search. The set of input sequences for the HMM profiles can derive from KOs in an input user-defined list (*--hmm_mode kos*) or from KOs in Modules of interest, e.g. those pointing to specific metabolic functions in the input genomes (*--hmm_mode modules*). Alternatively, HMMs can be built from the KOs of all Modules with one incomplete ortholog block (*--hmm_mode onebm*).

When this analysis is performed, the following workflow is employed with every KO of interest:

1. A taxonomically relevant subset of the KEGG GENES database is downloaded via the KEGG API. This subset includes sequences for every species included in a clade, defined by a C-level KEGG BRITE taxonomical hierarchy (br08601). Such taxonomy is generally almost coincident to that on the phylum level, or to that on the class level for a few specific taxa (e.g. Euryarchaeota).
2. A filtering step is performed to obtain a non-redundant set of sequences. A multiple sequence alignment is built up from these sequences using MAFFT v7.475 [15]. The *--auto* parameter is used here, to choose the appropriate strategy among the possible algorithms according to the size of the alignment dataset.
3. A HMM is generated from the aligned sequences using the *hmmbuild* command from the HMMER suite v3.1b2 [16]. Only the subset of KOs indicated in the *--hmm_mode* argument is utilized.
4. The obtained profiles are searched in the genome of interest with the *nhmmer* program from HMMER version 3.1b2 [16].

The default threshold value depends on the *nhmmer* score divided by the length of the profile HMM. Preliminary tests were performed to fine-tune this value, comparing translated BLASTp hits against the NCBI nr dataset (performed in March 2021), which were manually checked for two different MAG datasets. The threshold identified the highest number of hits with sequence names matching the correct KEGG ortholog gene descriptors, while pointing to the lowest number of false positives. Values obtained from the aforementioned tests resulted in 4.6–7.5% of the hits, depending on the input dataset. Stringency of the scoring for significant hits can be modulated with the *--threshold_value* parameter.

2.3. Integration of recovered biochemical reactions into genome-scale metabolic models

In automated draft GSMM reconstruction, metabolic reactions are collected based on genome or protein sequence alignment scores. Using KEMET, the HMM best scoring hits can be selected, providing new insights into the metabolic network obtained from the initial gene calling process. One option is the generation of a novel GSMM with newly identified orthologs. Alternatively, the HMM prioritization process determines a different set of reactions to be included in an existing GSMM. KEMET implements the --

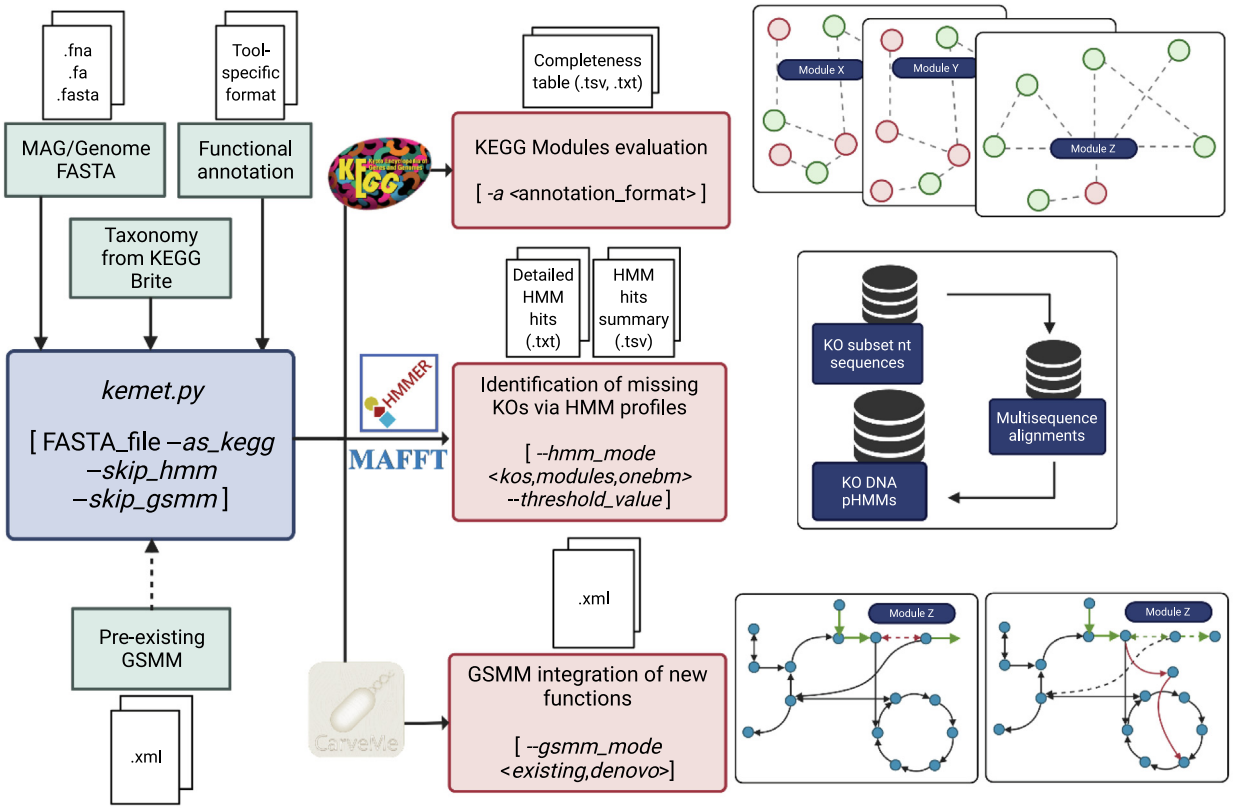

Fig. 1. Workflow of KEMET reporting the input files, outputs, and main parameters for all the tasks that can be executed: KEGG Module evaluation, identification of missing KOs, and integration of identified KOs in GSMMs. On the right side, the rationale of each task is visually outlined.

gsmm_mode parameter to include the newly predicted biochemical functions obtained from genomic evidence into the GSMMs.

KEMET links HMM-identified KOs to their corresponding biochemical reactions present in reference databases for GSMM, namely BiGG [17] and ModelSEED [18]. Their namespaces are adopted by popular GSMM reconstruction tools, such as CarveMe [10] and gapseq [11]. The retrieved reactions can then be incorporated in input GSMMs. As a second option, the translated HMM KO hits can be directly added to the input sequences used for *de novo* genome-scale model generation.

3. Results and discussion

To validate KEMET, we first compared its KEGG Module partitioning with those performed by KEGG Mapper and METABOLIC v4.0 [19] across all the KEGG Modules present at the time of the tests. As shown by Fig. 2A, the three tools interpret the Module block structure in a largely consistent way. However, KEMET is able to capture more Modules in the evaluation and has a block structure that more closely resembles that of KEGG as compared to METABOLIC.

Next, we validated KEMET annotation expansion by two different approaches: (a) an annotation removal strategy to test its ability to identify known KO annotations, and (b) a draft GSMM reconstruction strategy to verify that newly identified annotations produce more sound quantitative models of microbial metabolism, and thus reflect correctly identified functions.

Strategy (a) was used to test *kemet.py --hmm_mode* capability to retrieve the proper annotated sequences when either the original annotation was removed or the sequence was truncated. The rationale was to simulate misassembly-derived gene disruptions and other problems impairing functional prediction in MAGs. KEMET was tested on 12 MAGs derived from a contig-level assembly resulting from a previous work [20] as well as 5 complete genomes downloaded from NCBI (details in Supplementary Data). In terms of taxonomic "novelty", the MAGs were highly different and included species assigned at different levels (spanning from class to species) using GTDB-tk v1.5.0 [21]. The gene calling was performed using Prodigal v2.6.3 [22] with default options. Functional annotations of predicted genes were performed using eggNOG-mapper v2 [7] with default parameters. While in principle alternative gene predictions can impact the subsequent functional annotation, previous empirical investigations found negligible performance variation among different tools [23,24]. For this reason, our tests focused on benchmarking functional annotation prediction by using a single state-of-the-art gene prediction tool.

The test consisted in the removal of three KO annotations from the input set of each genome (i.e. from eggNOG results) before running KEMET with the *--hmm_mode onebm* option. The selected KOs were annotated once per genome, only on a single gene. Moreover, removed KOs were chosen from different Modules marked "Complete" by KEMET, among the mandatory orthologs for a given biochemical step. In this way, removing them would result in the change of Module completeness to "1 block missing". Altogether, 20/36 and 8/12 KO mock removals (55% and 67% true positive rate) resulted in the correct gene and annotation recovery for MAGs and complete genomes, respectively (Fig. 2B and Supplementary Data).

To model MAG construction issues more closely, the removal strategy was repeated two more times by simulating the deletion of tested KO-annotated gene sequences, either by 30% or 70% of their original length. This was done to mimic the typical scenario of a highly fragmented assembly where gene sequences can be

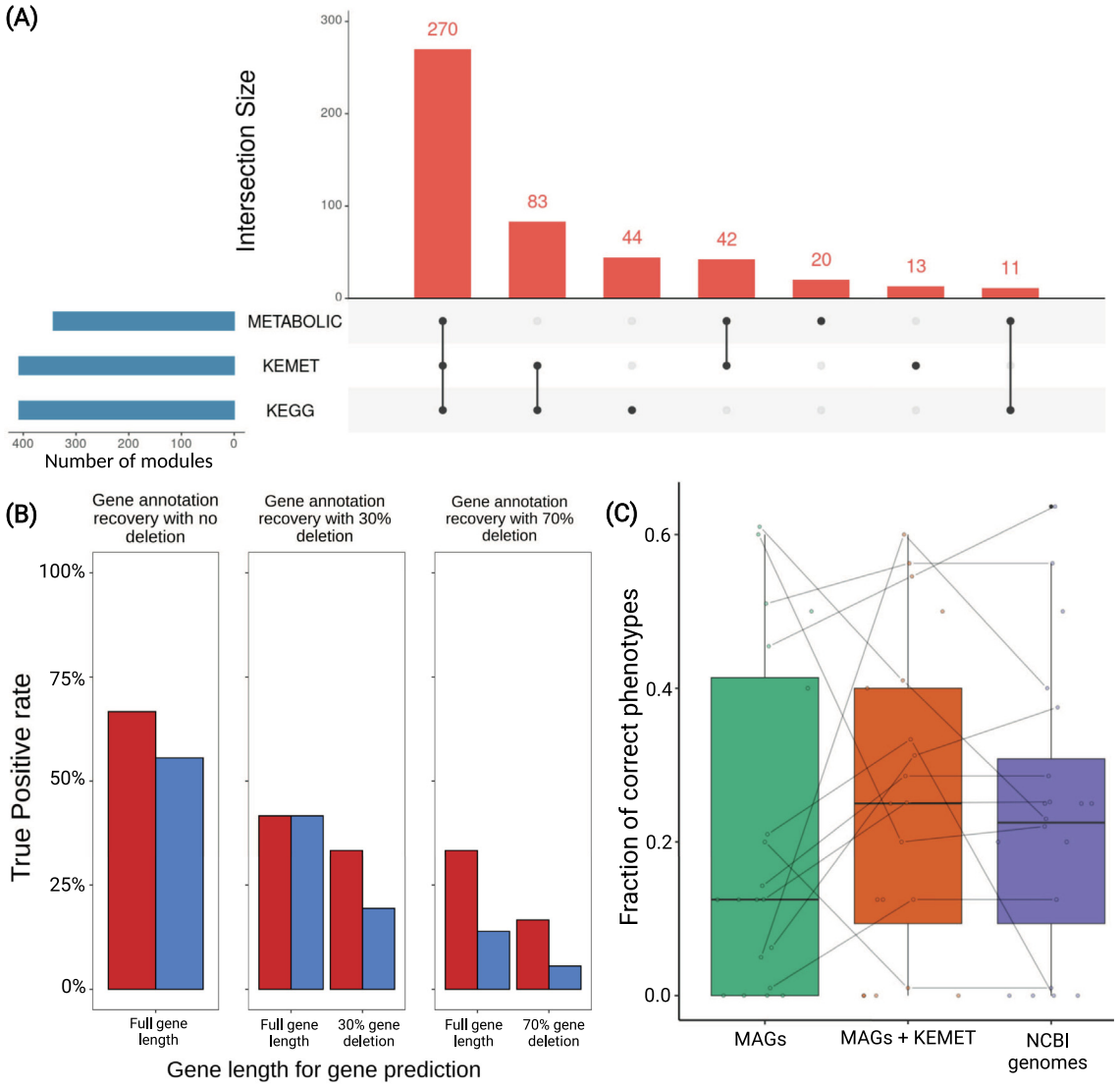

Fig. 2. Results of KEMET quality tests. (A) Comparison between KEMET and METABOLIC in terms of KEGG Module block structure with respect to the original KEGG Modules obtained through KEGG Mapper. The plot shows the intersections among the Module datasets for the three tools, together with the total number of Modules evaluated by each of them. (B) True positive rate for gene sequence identification by HMMs. Results for both isolated genomes (red) and MAGs (blue) are reported. Gene deletions of different extents were performed prior to running KEMET. When deletions were performed, gene annotation recovery was evaluated both with the gene prediction resulting from the original sequences and from those truncated, in order to account for the impact of deletions on gene prediction. (C) Fraction of correct metabolic phenotypes predicted by GSMMs reconstructed from microbial MAGs (green), the same MAGs with an expanded annotation through KEMET (orange), and the corresponding genomes from isolates (purple), based on the literature. The lines track the performance of individual GSMMs corresponding to the same strain. For readability purposes, only lines between points having performance differences across the datasets were drawn. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

split between two different contigs, resulting in a missed gene prediction or improper functional annotation. These additional tests resulted in a decreased performance using both the complete genomes and the MAGs dataset, as expected, but nonetheless gave a significant annotation recovery rate for gene truncations shorter than 50%. Specifically, an annotation recovery between 20% and 33% was achieved when accounting for the impact of sequence truncation on the gene prediction step, whereas a recovery rate of 42% was obtained assuming an unbiased gene prediction. This interval therefore captures KEMET performance in the presence of minor gene deletions. Similarly, for 70% gene truncations the annotation recovery rate further decreases, more clearly for the MAG dataset, as it is sensible with most of the gene sequence lacking. Hence, these results provide a proof-of-principle of KEMET annotation recovery in the occurrence of gene sequence disruption. Detailed results are included in the GitHub page at https://github.com/Matteopaluh/KEMET/blob/main/tests/README.md.

Strategy (b) was implemented to assess the impact of recovering missing KO annotation on downstream metabolic analyses, i.e. via GSMM reconstruction. Specifically, we compared microbial phenotypes recovered from the literature (indicated in Supplementary Data) in terms of metabolite production or consumption capabilities, to their corresponding *in silico* model predictions. This analysis was performed starting from MAGs and their corresponding complete genomes recovered from the NCBI or from the PATRIC database (as pointed by https://github.com/snayfach/IGGdb), by selecting species collected from the anaerobic digestion microbiome [20]. MAG quality metadata were recovered and included genome completeness and contamination. If more than one MAG per species was present in the database, those with $\geq$ 90% com-

pleteness and ≤ 5% contamination were considered for the subsequent analysis. Both MAGs and the complete genomes of isolates were used to check the Module completeness. MAGs were also used to search for missing KOs by using *kemet.py --hmm_mode onebm*. GSMMs were reconstructed from complete genomes and MAGs using CarveMe v1.4.1 [10] with the options *--fbc2 -u*, using as input both the MAG original gene calling and this same data added with the translated nucleotide sequences identified with the HMM via KEMET using the *--gsmm_mode denovo* parameter. Moreover, KEMET performance times were monitored and are included in Supplementary Data.

To benchmark how the addition of newly identified sequences affects GSMM ability to describe *in silico* microbial physiology, metabolic capabilities retrieved from the literature were compared with predictions obtained starting from three types of input for GSMM reconstruction: MAG annotation, MAG annotation expanded with KEMET, and complete genome annotation. Flux variability analysis (FVA) was performed on the obtained GSMMs for assessing such metabolic capabilities, as follows. For each metabolite export reaction, it was determined whether the range of possible fluxes was directed towards metabolite consumption or production (respectively, having flux ranges consisting only of negative or positive values), while maintaining a fixed maximal growth rate. FVA results showing blocked reactions or flux ranging both positive and negative values were considered as incorrect predictions. The results show a nearly 10% improvement in the ability of MAG-derived GSMMs to produce and consume metabolites predicted from wet lab experiments, with an acquired accuracy comparable with the accuracy of GSMMs reconstructed from the genomes of isolates (both around 33%, Fig. 2B and Supplementary Data). On the annotation level, HMMs used on MAGs resulted in 84.76% hits in common with the respective reference isolate genome selected; 7.62% hits were present solely in the MAG dataset (false positives), and 7.62% hits were present in the complete genome dataset alone (false negatives). According to the selected dataset, KEMET HMM predictions therefore display a 91.75% precision and 91.75% sensitivity (Supplementary Data). Despite the addition of a limited number of protein sequences, the resulting models can thus be sensibly more accurate, leading to more precise inferences based on metabolic capabilities. For example, *Selenomonas ruminantium* MAG-derived GSMM (PATRIC genome id: 971.16) phenotype predictions were improved after KEMET usage. The original GSMM could not predict any known metabolic capability of *S. ruminantium*, while the modified GSMM could correctly reproduce metabolic exchanges involving cellobiose, salicin, mannitol, xylose, arabinose, fructose, maltose, lactose, and sucrose. In contrast, the GSMM based on the full genome annotation captured the correct exchanges for glycerol, cellobiose, salicin, mannitol, xylose, and arabinose.

These results demonstrate that KEMET efficiently tackles the summarization of (meta)genomic potential in a user-friendly and scalable way. Other bioinformatics tools allow the evaluation of microbial genome annotation completeness (e.g. METABOLIC [19]). However, to date and up to our knowledge, this is the only tool able to selectively fill the gaps in the annotation, and seamlessly add newly gathered information into GSMMs. At the moment, KEMET relies on KEGG given its structure allowing a systematic pathway completeness evaluation. Further development could include support towards other knowledgebases, such as MetaCyc [25], to further expand the tool compatibility and predictive power. While other published programs, such as DRAM and Anvi'o [26,27] rely on specific KEGG releases, KEGG databases are constantly updated due to newly added sequences, or newly defined KO classifications. In contrast, KEMET allows users to update the downloaded KEGG GENES database through the KEGG API, in order to use the most up-to-date version of KEGG database without relying on fixed versions. The download of such a database represents the only limiting computational factor in KEMET (Supplementary Data), being a mandatory step to comply with the KEGG license. More efficient communication with KEGG servers could be obtained via license, while better solutions will be explored and implemented in future versions of KEMET. Nevertheless, this step is required only once at each database update, which can be decided by the user. Further, KEMET is based on HMMs given their broad applicability in the genomics and metagenomics fields. Other probabilistic graphical models, such as conditional random fields or Bayesian networks could be implemented in future versions of the software.

Altogether, our experiments show that focusing on Module completeness down to single orthologs can aid in identifying missing annotations and enable their correction, not only supporting qualitative evaluation of microbial functions but also improving quantitative models of microbial metabolism. This enables a better mechanistic investigation of microbial ecological roles, allowing us to gather insights without relying necessarily on cultivation or in-depth characterization, which is impractical for most metagenomic studies.

Author Statement

The authors declare that all of them have seen and approved the final version of the manuscript. The manuscript is the authors' original work, has not received prior publication and is not under consideration for publication elsewhere.

Funding

This work was financially supported by the "Budget Integrato della Ricerca Dipartimentale" (BIRD198423) PRID 2019 of the Department of Biology of the University of Padua, entitled "SyMoBio: inspection of Syntrophies with Metabolic Modelling to optimize Biogas Production", by the project "Sviluppo Catalisi dell'Innovazione nelle Biotecnologie" (MIUR ex D.M.738 dd 08/08/19) of the Consorzio Interuniversitario per le Biotecnologie" (CIB), and by the project LIFE20 CCM/GR/001642 – LIFE CO2toCH4 of the European Union LIFE+ program.

CRediT authorship contribution statement

Matteo Palù: Conceptualization, Methodology, Software, Validation, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Arianna Basile:** Conceptualization, Methodology, Software, Writing – original draft, Writing – review & editing, Visualization. **Guido Zampieri:** Methodology, Validation, Supervision, Writing – review & editing, Visualization. **Laura Treu:** Conceptualization, Supervision, Writing – review & editing, Funding acquisition. **Alessandro Rossi:** Methodology, Software. **Maria Silvia Morlino:** Methodology, Writing – review & editing. **Stefano Campanaro:** Conceptualization, Supervision, Project administration, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2022.03.015.

References

[1] Nayfach S, Roux S, Seshadri R, Udwary D, Varghese N, Schulz F, et al. A genomic catalog of Earth's microbiomes. Nat Biotechnol 2021;39:499–509. https://doi.org/10.1038/s41587-020-0718-6.

[2] D'Hondt K, Kostic T, McDowell R, Eudes F, Singh BK, Sarkar S, et al. Microbiome innovations for a sustainable future. Nat Microbiol 2021;6:138–42. https://doi.org/10.1038/s41564-020-00857-w.

[3] Basile A, Campanaro S, Kovalovszki A, Zampieri G, Rossi A, Angelidaki I, et al. Revealing metabolic mechanisms of interaction in the anaerobic digestion microbiome by flux balance analysis. Metab Eng 2020;62:138–49. https://doi.org/10.1016/j.ymben.2020.08.013.

[4] Bowers RM, Kyrpides NC, Stepanauskas R, Harmon-Smith M, Doud D, Reddy TBK, et al. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. Nat Biotechnol 2017;35:725–31. https://doi.org/10.1038/nbt.3893.

[5] Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M. Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Res 2014;42:D199–205. https://doi.org/10.1093/nar/gkt1076.

[6] Kanehisa M, Sato Y, Kawashima M. KEGG mapping tools for uncovering hidden features in biological data. Protein Sci n.d.;n/a. https://doi.org/10.1002/pro.4172.

[7] Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. Mol Biol Evol 2021. https://doi.org/10.1093/molbev/msab293.

[8] Frioux C, Dittami SM, Siegel A. Using automated reasoning to explore the metabolism of unconventional organisms: a first step to explore host-microbial interactions. Biochem Soc Trans 2020;48:901–13. https://doi.org/10.1042/BST20190667.

[9] Zorrilla F, Buric F, Patil KR, Zelezniak A. metaGEM: reconstruction of genome scale metabolic models directly from metagenomes. Nucleic Acids Res 2021. https://doi.org/10.1093/nar/gkab815.

[10] Machado D, Andrejev S, Tramontano M, Patil KR. Fast automated reconstruction of genome-scale metabolic models for microbial species and communities. Nucleic Acids Res 2018;46:7542–53. https://doi.org/10.1093/nar/gky537.

[11] Zimmermann J, Kaleta C, Waschina S. gapseq: informed prediction of bacterial metabolic pathways and reconstruction of accurate metabolic models. Genome Biol 2021;22:81. https://doi.org/10.1186/s13059-021-02295-1.

[12] Bernstein DB, Sulheim S, Almaas E, Segrè D. Addressing uncertainty in genome-scale metabolic model reconstruction and analysis. Genome Biol 2021;22:64. https://doi.org/10.1186/s13059-021-02289-z.

[13] Aramaki T, Blanc-Mathieu R, Endo H, Ohkubo K, Kanehisa M, Goto S, et al. KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. Bioinformatics 2020;36:2251–2. https://doi.org/10.1093/bioinformatics/btz859.

[14] Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M. KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res 2007;35:W182–5. https://doi.org/10.1093/nar/gkm321.

[15] Nakamura T, Yamada KD, Tomii K, Katoh K. Parallelization of MAFFT for large-scale multiple sequence alignments. Bioinformatics 2018;34:2490–2. https://doi.org/10.1093/bioinformatics/bty121.

[16] Wheeler TJ, Eddy SR. nhmmer: DNA homology search with profile HMMs. Bioinformatics 2013;29:2487–9. https://doi.org/10.1093/bioinformatics/btt403.

[17] Norsigian CJ, Pusarla N, McConn JL, Yurkovich JT, Dräger A, Palsson BO, et al. BiGG Models 2020: multi-strain genome-scale models and expansion across the phylogenetic tree. Nucleic Acids Res 2020;48:D402–6. https://doi.org/10.1093/nar/gkz1054.

[18] Seaver SMD, Liu F, Zhang Q, Jeffryes J, Faria JP, Edirisinghe JN, et al. The ModelSEED Biochemistry Database for the integration of metabolic annotations and the reconstruction, comparison and analysis of metabolic models for plants, fungi and microbes. Nucleic Acids Res 2021;49:D575–88. https://doi.org/10.1093/nar/gkaa746.

[19] Zhou Z, Tran PQ, Breister AM, Liu Y, Kieft K, Cowley ES, et al. METABOLIC: high-throughput profiling of microbial genomes for functional traits, metabolism, biogeochemistry, and community-scale functional networks. Microbiome 2022;10:33. https://doi.org/10.1186/s40168-021-01213-8.

[20] Campanaro S, Treu L, Rodriguez-R LM, Kovalovszki A, Ziels RM, Maus I, et al. New insights from the biogas microbiome by comprehensive genome-resolved metagenomics of nearly 1600 species originating from multiple anaerobic digesters. Biotechnol Biofuels 2020;13:25. https://doi.org/10.1186/s13068-020-01679-y.

[21] Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. Bioinformatics 2020;36:1925–7. https://doi.org/10.1093/bioinformatics/btz848.

[22] Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinf 2010;11:119. https://doi.org/10.1186/1471-2105-11-119.

[23] Korandla DR, Wozniak JM, Campeau A, Gonzalez DJ, Wright ES. AssessORF: combining evolutionary conservation and proteomics to assess prokaryotic gene predictions. Bioinformatics 2020;36:1022–9. https://doi.org/10.1093/bioinformatics/btz714.

[24] Dimonaco NJ, Aubrey W, Kenobi K, Clare A, Creevey CJ. No one tool to rule them all: prokaryotic gene prediction tool annotations are highly dependent on the organism of study. Bioinformatics 2022;38:1198–207. https://doi.org/10.1093/bioinformatics/btab827.

[25] Caspi R, Billington R, Keseler IM, Kothari A, Krummenacker M, Midford PE, et al. The MetaCyc database of metabolic pathways and enzymes - a 2019 update. Nucleic Acids Res 2020;48:D445–53. https://doi.org/10.1093/nar/gkz862.

[26] Eren AM, Kiefl E, Shaiber A, Veseli I, Miller SE, Schechter MS, et al. Community-led, integrated, reproducible multi-omics with anvi'o. Nat Microbiol 2021;6:3–6. https://doi.org/10.1038/s41564-020-00834-3.

[27] Shaffer M, Borton MA, McGivern BB, Zayed AA, La Rosa SL, Solden LM, et al. DRAM for distilling microbial metabolism to automate the curation of microbiome function. Nucleic Acids Res 2020;48:8883–900. https://doi.org/10.1093/nar/gkaa621.