Bioinformatics, 37(18), 2021, 3029–3031
doi: 10.1093/bioinformatics/btab184
Advance Access Publication Date: 18 March 2021
Applications Note

Sequence analysis

Fast and sensitive taxonomic assignment to metagenomic contigs

M. Mirdita [1], **M. Steinegger** [2,3,4], **F. Breitwieser** [5], **J. Söding** [1,6,*] **and E. Levy Karin** [1,*]

[1]Quantitative and Computational Biology, Max Planck Institute for Biophysical Chemistry, Göttingen, Germany, [2]School of Biological Sciences, Seoul National University, Seoul, South Korea, [3]Institute of Molecular Biology and Genetics, Seoul National University, Seoul, South Korea, [4]Artificial Intelligence Institute, Seoul National University, Seoul, South Korea, [5]Center for Computational Biology, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins School of Medicine, Baltimore, MD 21205, USA and [6]Campus-Institut Data Science (CIDAS), Göttingen, Germany

*To whom correspondence should be addressed.

Associate Editor: Janet Kelso

Received on November 16, 2020; revised on February 26, 2021; editorial decision on March 11, 2021; accepted on March 16, 2021

Abstract

Summary: MMseqs2 taxonomy is a new tool to assign taxonomic labels to metagenomic contigs. It extracts all possible protein fragments from each contig, quickly retains those that can contribute to taxonomic annotation, assigns them with robust labels and determines the contig's taxonomic identity by weighted voting. Its fragment extraction step is suitable for the analysis of all domains of life. MMseqs2 taxonomy is 2–18× faster than state-of-the-art tools and also contains new modules for creating and manipulating taxonomic reference databases as well as reporting and visualizing taxonomic assignments.

Availability and implementation: MMseqs2 taxonomy is part of the MMseqs2 free open-source software package available for Linux, macOS and Windows at https://mmseqs.com.

Contact: soeding@mpibpc.mpg.de or eli.levy.karin@gmail.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Metagenomic studies shine a light on previously unstudied parts of the tree of life. However, unraveling taxonomic composition accurately and quickly remains a challenge. While most methods label short metagenomic reads (reviewed in Sczyrba *et al.*, 2017), only a handful (e.g. Huson *et al.*, 2018) assign entire contigs, even though this should lead to improved accuracy.

Recently, von Meijenfeldt *et al.* (2019) developed CAT, a tool for taxonomic annotation of contigs based on protein homologies to a reference database. It combines Prodigal (Hyatt *et al.*, 2010) for predicting open reading frames (ORFs), DIAMOND (Buchfink *et al.*, 2015) to search with the translated ORFs, and logic to aggregate individual ORF annotations. CAT achieved higher precision than state-of-the-art tools on bacterial benchmarks. Despite its advantage over existing methods, CAT has limitations: (i) Prodigal was designed for prokaryotes and not eukaryotes (West *et al.*, 2018); (ii) Prodigal runs single-threaded, limiting applicability to metagenomics; (iii) CAT's *r* parameter determines the cut-off score below each ORF's top-hit above which hits are included in the ORF's lowest common ancestor (LCA) computation. Although the authors provide guidelines to set *r*, it is unclear how general they are.

Here, we present MMseqs2 taxonomy, a novel protein-search-based tool for taxonomy assignment to contigs. It overcomes the aforementioned limitations by extracting all possible protein fragments, covering the coding repertoire of all domains of life. It quickly eliminates fragments that do not bear minimal similarity to the reference database, and searches with the remaining ones. MMseqs2 taxonomy uses an approximate 2bLCA (Hingamp *et al.*, 2013) strategy to assign translated fragments to taxonomic nodes (Supplementary Material). The hits for the approximate 2bLCA computation are determined automatically, saving the need to tune an equivalent of CAT's *r* parameter. It outperforms CAT on bacterial and eukaryotic datasets.

2 Materials and methods

Input. Contigs are provided as (compressed) FASTA/Q files. As reference, the *databases* workflow can download and prepare various public taxonomy databases, such as, nr (Agarwala *et al.*, 2018), UniProt (Bateman, 2019) or GTDB (Parks *et al.*, 2020). Alternatively, users can prepare their own taxonomic reference database (see MMseqs2 wiki).

Algorithm. The four main steps are described in Figure 1A.

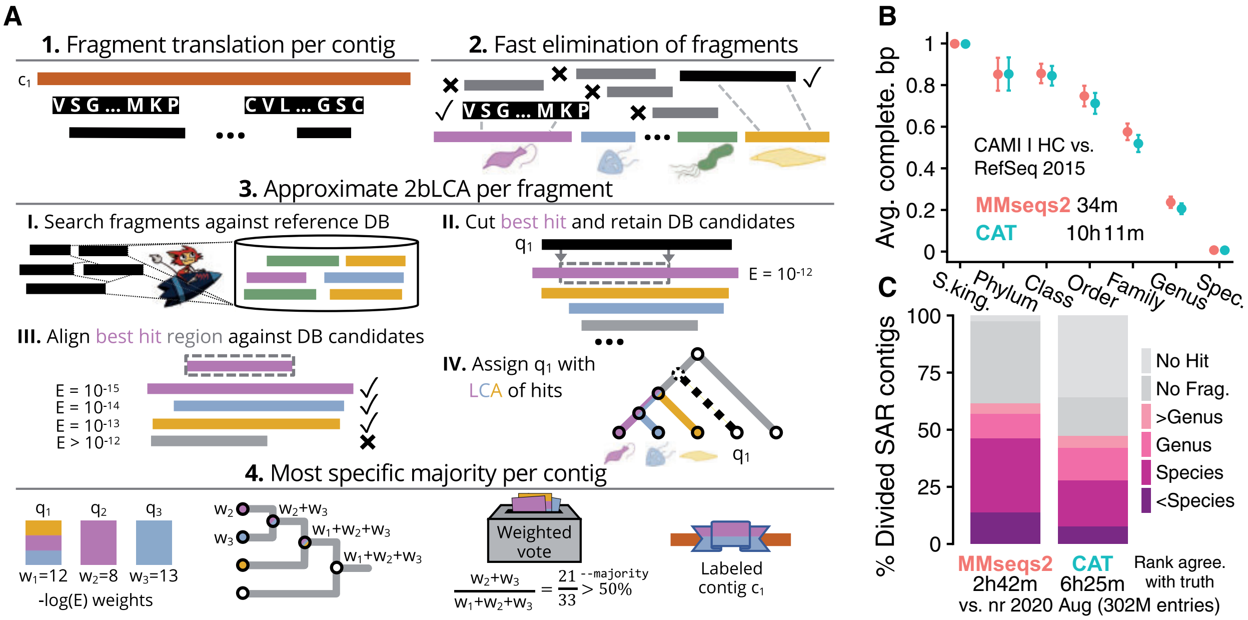
A
1. Fragment translation per contig
c_1
V S G ... M K P
C V L ... G S C
2. Fast elimination of fragments
V S G ... M K P
3. Approximate 2bLCA per fragment
I. Search fragments against reference DB
II. Cut best hit and retain DB candidates
q_1
E = 10^{-12}
III. Align best hit region against DB candidates
E = 10^{-15}
E = 10^{-14}
E = 10^{-13}
E > 10^{-12}
IV. Assign q_1 with LCA of hits
q_1
4. Most specific majority per contig
q_1 q_2 q_3
w_1=12 w_2=8 w_3=13
-log(E) weights
w_2
w_3
w_2+w_3
$w_1+w_2+w_3$
$w_1+w_2+w_3$
Weighted vote
$\frac{w_2+w_3}{w_1+w_2+w_3} = \frac{21}{33} \overset{\text{--majority}}{>} 50\%$
Labeled contig c_1
B
Avg. complete. bp
1
0.8
0.6
0.4
0.2
0
CAMI I HC vs. RefSeq 2015
MMseqs2 34m
CAT 10h11m
S.king.
Phylum
Class
Order
Family
Genus
Spec.
C
% Divided SAR contigs
100
75
50
25
0
No Hit
No Frag.
>Genus
Genus
Species
<Species
MMseqs2 2h42m
CAT 6h25m
Rank agree. with truth
vs. nr 2020 Aug (302M entries)

Fig. 1. (**A**) Taxonomy assignment algorithm in four steps: (1) Translate all possible protein fragments in six frames from all contigs. (2) Reject fragments unlikely to find a taxonomic hit in later stages (full details in Supplementary Material). (3) Assign taxonomic nodes using an approximate 2bLCA procedure. Each query fragment *q* is searched against the reference database, resulting in a list *l* of all its homologous targets. The aligned region between *q* and the best hit *t* [with *E*-value *E*(*q*, *t*)] is aligned against all targets in *l*. Assign *q* the LCA of the taxonomic lables of all target sequences that have an *E*-value lower than *E*(*q*, *t*). Realigning *l* allows avoiding the costly second search of 2bLCA. (4) Each assigned *q* contributes its weight (*–log E*(*q*, *t*)) to its taxonomic label and all labels above it, up to the root. The contig's taxonomic node is determined as the most specific taxonomic label, which has a support of at least the −−majority parameter. The support of a label is the sum of its contributing weights divided by the total sum of weights. (**B**) MMseqs2 taxonomy (red) is ∼18× faster and achieves similar average completeness to CAT (turquoise) on a bacterial benchmark. (**C**) MMseqs2 assigns taxa to eukaryotic SAR contigs more accurately than CAT across all phylogenetic levels, at twice the speed. At species level, MMseqs2 taxonomy classifies 46% contigs correctly versus 28% for CAT. Runtimes measured on a 2×14-core Intel E5-2680v4 server with 768 GB RAM

Output. MMseqs2 taxonomy returns the following eight fields for each contig accession: (i) the taxonomic identifier (taxid) of the assigned label, (ii) rank, (iii) name, followed by the number of fragments: (iv) retained, (v) taxonomically assigned, and (vi) in agreement with the contig label (i.e. same taxid or have it as an ancestor), (vii) the support the taxid received and, optionally, (viii) the full lineage. The result can be converted to a TSV-file, and to a Kraken (Wood *et al.*, 2019) report or a Krona (Ondov *et al.*, 2011) visualization (Supplementary Material).

3 Results

Bacterial dataset. The CAMI-I high-complexity challenge and its accompanying RefSeq 2015 reference database (Sczyrba *et al.*, 2017) were given to MMseqs2 and CAT. AMBER v2 (Meyer *et al.*, 2018) was used to assess the taxonomic assignment by computing the average completeness (Fig. 1B) and purity (Supplementary Fig. S1) bp using its taxonomic binning benchmark mode. At similar assignment quality, MMseqs2 taxonomy is 18× faster than CAT. Using the nr, MMseqs2 is 10× faster (Supplementary Fig. S2).

Eukaryotic dataset. All 57 SAR (taxid 2698737) RefSeq assemblies and their taxonomic labels were downloaded from NCBI in 08/2020. To resemble metagenomic data, their scaffolds were randomly divided following the length distribution of contigs assembled for sample ERR873969 of eukaryotic Tara Oceans (Carradec *et al.*, 2018), resulting in 2.7 million non-overlapping contigs with a minimal length of 300 bp. Using nr from 08/2020, MMseqs2 classified more contigs than CAT (62% versus 47%). For 36%, CAT extracted a fragment that did not hit the reference, suggesting fragments extracted by MMseqs2 are more informative for eukaryotic taxonomic annotation (Fig. 1C, Supplementary Fig. S3).

4 Conclusion

MMseqs2 taxonomy is as accurate as CAT on a bacterial dataset while being 3–18× faster and requiring fewer parameters. Its extracted fragments make it suitable for analyzing eukaryotes. It is accompanied by several taxonomy utility modules to assist with taxonomic analyses.

Funding

E.L.K. is a FEBS long-term fellowship recipient and an EMBO nonstipendiary long-term fellow. The work was supported by the BMBF CompLifeSci project horizontal4meta; the ERC's Horizon 2020 Framework Programme ['Virus-X', project no. 685778]; the National Research Foundation of Korea grant funded by the Korean government (MEST) [2019R1A6A1A10073437, NRF-2020M3A9G7103933]; and the Creative-Pioneering Researchers Program through Seoul National University.

Conflict of Interest: none declared.

Data availability

The data used to benchmark MMseqs2 taxonomy in this study are openly available from https://data.cami-challenge.org/participate at the Databases section. The SAR assemblies were downloaded from NCBI in 08/2020 and processed as described.

References

Agarwala,R. *et al.* (2018) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **46**, D8–D13.

Bateman,A. (2019) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.*, **47**, D506–D515.

Buchfink,B. *et al.* (2015) Fast and sensitive protein alignment using DIAMOND. *Nat. Methods*, **12**, 59–60.

Carradec,Q. *et al.*, Tara Oceans Coordinators. (2018) A global ocean atlas of eukaryotic genes. *Nat. Commun.*, **9**, 373.

Hingamp,P. *et al.* (2013) Exploring nucleo-cytoplasmic large DNA viruses in Tara Oceans microbial metagenomes. *ISME J.*, **7**, 1678–1695.

Huson,D.H. *et al.* (2018) MEGAN-LR: new algorithms allow accurate binning and easy interactive exploration of metagenomic long reads and contigs. *Biol. Direct.*, **13**, 6.

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

Meyer,F. *et al.* (2018) AMBER: Assessment of Metagenome BinnERs. *Gigascience*, **7**, giy069.

Ondov,B.D. *et al.* (2011) Interactive metagenomic visualization in a Web browser. *BMC Bioinform.*, **12**, 385.

Parks,D.H. *et al.* (2020) A complete domain-to-species taxonomy for Bacteria and Archaea. *Nat. Biotechnol.*, **38**, 1079–1086.

Sczyrba,A. *et al.* (2017) Critical assessment of metagenome interpretation—a benchmark of metagenomics software. *Nat. Methods*, **14**, 1063–1071.

von Meijenfeldt,F.A.B. *et al.* (2019) Robust taxonomic classification of uncharted microbial sequences and bins with CAT and BAT. *Genome Biol.*, **20**, 217.

West,P.T. *et al.* (2018) Genome-reconstruction for eukaryotes from complex natural microbial communities. *Genome Res.*, **28**, 569–580.

Wood,D.E. *et al.* (2019) Improved metagenomic analysis with Kraken 2. *Genome Biol.*, **20**, 257.