EsMeCaTa: Estimating metabolic capabilities from taxonomic affiliations

Arnaud Belcour [1], Baptiste Ruiz [1], Clémence Frioux [2], Samuel Blanquart [1*] and Anne Siegel [1*]

[1]Univ Rennes, Inria, CNRS, IRISA, F-35000 Rennes, France.
[2]Inria, INRAE, Université de Bordeaux, France.
* Co-last authors.

March 16, 2022

Abstract

Summary: Predicting the functional potential of microorganisms in environmental samples from cultivation-independent techniques is a major challenge. A persistent difficulty lies in associating taxonomic profiles obtained from metabarcoding experiment with accurate functional profiles, particularly for poorly-resolved taxonomic groups. In this paper, we present EsMeCaTa a python package predicting shared proteins from taxonomic affiliations. EsMeCaTa relies on the UniProt database to retrieve the public proteomes associated with a taxon and then uses MMseqs2 in order to compute the set of proteins shared in the taxon. Finally, EsMeCaTa extracts the functional annotations of these proteins to provide an accurate estimate of the functional potential associated to taxonomic affiliations.
Availability: EsMeCaTa is available at: https://github.com/AuReMe/esmecata under the GPL-3 license.

1 Introduction

Sequencing of gene markers, such as 16S rRNA gene, is commonly applied to characterize the diversity of organisms in environmental samples. From these data, taxonomic affiliations, such as Operational Taxonomic Units (OTUs), can then be used to estimate the potential functions in the environment.

Various tools have been developed to estimate functional profiles associated with an OTU, such as PICRUSt2 (Douglas et al., 2020), Paprica (Bowman and Ducklow, 2015), Tax4fun2 (Wemheuer et al., 2020). These tools have been developed mainly to proceed 16S rRNAs, although other gene markers can be considered (*e.g.* Ogier et al. (2019)). Apart from marker genes, taxonomic

affiliations can be obtained from shallow whole genome sequencing data, but the association of function with these data remains uneasy.

We developed EsMeCaTa (Estimating Metabolic Capabilities from Taxonomic affiliations) as a method permitting the estimation of functions independently of the taxonomic assignment method used. EsMeCaTa is a python package that relies on the UniProt database to estimate the shared proteins associated with prokaryotic or eukaryotic taxa considered as input, providing insights into their putative metabolic capabilities.

2 Approach

EsMeCaTa takes as input a tabulated file containing two columns. The first column is an identifier and the second contains a taxonomic affiliation (starting with the highest taxonomic rank, such as kingdom, to the lowest taxonomic rank, such as species) as defined by the NCBI Taxonomy database (Schoch et al., 2020). The outputs of the workflow are, for each taxon (1) fasta files of all proteomes selected by EsMeCaTa, (2) a fasta file of the shared proteins clustered by MMseqs2 from these proteomes, and (3) a tabulated file containing the functional annotations associated with these proteins (Gene Ontology Terms (GO), Enzyme Commission (EC)).

The first part of the workflow selects the lowest taxonomic rank of each input taxonomic affiliation for which the UniProt Proteomes database (The UniProt Consortium, 2021) contains at least one proteome exhibiting a BUSCO score higher than 80% and considered "non-redundant" and "not-excluded" by UniProt (column 'Proteomes selection' in Table 1). More precisely, the taxonomic affiliation is processed using the ete3 python package (Huerta-Cepas et al., 2016) in order to associate a taxon ID (from the NCBI taxonomy database) to each taxon from the affiliation. Using this ID, queries against the UniProt Proteomes database find the lowest-ranking taxon for which there is at least one reference or non-reference proteome in the database and download those proteomes. If the number of proteomes is greater than a threshold (100 by default), only a subsample is downloaded.

The second part of the workflow aims at identifying proteins shared by proteomes associated with a taxon. With the downloaded proteomes, EsMeCaTa performs protein clustering using MMseqs2 (Steinegger and Söding, 2017) and selects clusters such that proteins are shared by at least X% of the proteomes (see the column 'Protein clusters (MMseqs2)' in Table 1). The threshold $X = 0$ corresponds to the case where all protein clusters are selected (called 'Pan-proteome', abbreviated Pan-P, in reference to pan-genome). A second threshold at $X = 0.95$ retains only the clusters containing a protein originating from at least 95% of the proteomes (called 'Soft core proteome', abbreviated Soft-P). A third threshold at $X = 0.5$ retains cluster containing proteins occurring in at least 50% of the proteomes (called 'Shell core proteome', abbreviated Shell-P). For each protein cluster, EsMeCaTa selects the representative protein (first sequence in the alignment made by MMseqs2) to represent the cluster. The

selected sequences of the representative proteins are stored in a fasta file using biopython (Cock et al., 2009).

The final step of the workflow annotates the protein clusters by querying the UniProt database (GO, EC, see the columns 'Functional annotation of clusters' in the Table 1).

3 Results

Input		**Taxons selected by EsMeCaTa**		**Proteomes selection (Busco ≥ 0.8)**			**Protein clusters (MMseqs2)**			**Functional annotation of clusters**					
Lowest taxon name	Taxon rank	Taxon rank used	Taxon name used	UniProt total	UniProt references	EsMeCaTa proteomes	Pan-P	Soft-P	Shell-P	Pan-P		Soft-P		Shell-P	
										GO	EC	GO	EC	GO	EC
Escherichia	Genus	Genus	*Escherichia*	1,506	3	3	5,821	2,421	3,298	2,183	866	1,661	679	1,906	792
Citrobacter	Genus	Genus	*Citrobacter*	138	2	2	5,674	2,753	5,674	2,013	772	1,835	708	2,013	772
Cronobacter	Genus	Genus	*Cronobacter*	15	0	15	9,057	101	3,128	970	677	0	12	600	603
Lelliottia	Genus	Genus	*Lelliottia*	5	0	5	5,252	2,651	3,245	1,993	756	1,784	687	1,884	718
Jejubacter	Genus	Genus	*Jejubacter*	1	1	1	3,915	3,915	3,915	1,983	837	1,983	837	1,983	837
Edaphovirga	Genus	**Family**	Enterobacteriaceae	2,435	42	42	25,822	415	2,581	2,253	867	514	193	1,560	595
Enterobacteriaceae	Family	Family	Enterobacteriaceae	2,435	42	42	25,822	415	2,581	2,253	867	514	193	1,560	595
Enterobacterales	Order	Order	Enterobacterales	3,028	129	96	53,617	375	2,145	2,475	1,010	487	175	1,383	512
Gammaproteobacteria	Class	Class	Gammaproteobacteria	8,271	911	96	85,797	329	1,183	2,650	1,040	387	123	924	327
Plasmodium	Genus	Genus	*Plasmodium*	67	17	17	21,287	1,276	4,263	1,305	225	611	104	1,103	200
Leucocytozoon	Genus	**Order**	Haemosporida	68	18	18	22,813	1,076	4,313	1,327	259	546	95	1,090	199
Corallicola	Genus	**Class**	Conoidasida	30	10	10	46,959	76	1,326	1,919	530	94	14	717	121
Acavomonas	Genus	**Clade**	Alveolata	124	48	48	248,878	50	785	3,746	924	42	7	418	76

Table 1: **Result of EsMeCaTa on 13 taxonomic affiliations** (described by names and ranks). Soft-P: Soft core proteome. Shell-P: Shell core proteome. Pan-P: Pan-proteome.

Table 1 shows the application of EsMeCaTa (with UniProt 2021_03 and MMseqs2 13.45111) to 13 different taxonomic affiliations (the lowest taxon in these affiliations is indicated in the 'Lowest taxon name' column) selected to cover both prokaryotic (Gammaproteobacteria) and eukaryotic (Alveolata) taxa with different taxonomic ranks (from class to genus), in order to illustrate the uncertainty in the input taxonomic affiliation, the available knowledge and the biases toward most documented clades.

For 9 taxonomic affiliations, proteomes were selected using the lowest taxonomic rank, whereas for the 4 other affiliations, EsMeCaTa selected a higher taxonomic rank (bold in the column 'Taxon rank used'). For the 13 selected taxa, UniProt contained from 1 to 8,271 proteomes. In two cases (*Cronobacter* and *Lelliottia*), no reference proteome was found and EsMeCaTa returned non-reference proteomes of Uniprot. In 9 cases (such as *Escherichia*), EsMeCaTa returned the reference proteomes found in Uniprot (from 1 to 48). In the last two cases (*Enterobacterales* and *Gammaproteobacteria*), more than 99 reference proteomes were found, of which 96 proteomes were selected by the subsampling procedure.

We observe that in general (in the column 'Protein clusters (MMseqs2)' of the table 1), the size of the Pan-P increases with the number of selected proteomes, while the size of the Soft-P decreases. The size of the Shell-P appears to be much more stable. The numbers of GOs and ECs recovered follow the same trends and are systematically lower than the Pan-P, Shell-P and Soft-P sizes

(column 'Functional annotation of clusters' of Table 1).

To test EsMeCaTa on environmental samples, we analysed the taxonomic affiliations contained in the 16S rRNA (413 OTUs) and rpoB (309 OTUs) datasets provided in Ogier et al. (2019). Run of EsMeCaTa on the 722 OTUs took 2 days and 10 hours on a 20 CPU cluster. For the 16S and rpoB taxonomic affiliations respectively, means of 31 and 22 proteomes were recovered. Soft-P contained respectively 1,335 and 1,669 proteins clusters in average, associated with respectively 815 and 1,014 GOs, and 288 and 367 ECs in average.

This suggests that EsMeCaTa can be used for the analysis of environmental samples by predicting proteins and functions. This paves the way to study the metabolic capabilities of taxa present in the sample.

Financial Support: This work was supported by Deep Impact ANR-20-PCPA-0004.

Conflict of Interest: none declared.

Acknowledgements: We acknowledge the GenOuest bioinformatics core facility at https://www.genouest.org.

References

Jeff S. Bowman and Hugh W. Ducklow. Microbial communities can be described by metabolic structure: A general framework and application to a seasonally variable, depth-stratified microbial community from the coastal west antarctic peninsula. *PLOS ONE*, 10(8):e0135868, 2015. ISSN 1932-6203. doi: 10.1371/journal.pone.0135868. URL `https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0135868`.

Peter J. A. Cock, Tiago Antao, Jeffrey T. Chang, Brad A. Chapman, Cymon J. Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, and Michiel J. L. de Hoon. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, 2009. ISSN 1367-4803. doi: 10.1093/bioinformatics/btp163. URL `https://doi.org/10.1093/bioinformatics/btp163`.

Gavin M. Douglas, Vincent J. Maffei, Jesse R. Zaneveld, Svetlana N. Yurgel, James R. Brown, Christopher M. Taylor, Curtis Huttenhower, and Morgan G. I. Langille. PICRUSt2 for prediction of metagenome functions. *Nature Biotechnology*, 38(6):685–688, 2020. ISSN 1546-1696. doi: 10.1038/s41587-020-0548-6. URL `https://www.nature.com/articles/s41587-020-0548-6`.

Jaime Huerta-Cepas, François Serra, and Peer Bork. ETE 3: Reconstruction, analysis, and visualization of phylogenomic data. *Molecular Biology and Evolution*, 33(6):1635–1638, 2016. ISSN 0737-4038. doi: 10.1093/molbev/msw046. URL `https://doi.org/10.1093/molbev/msw046`.

Jean-Claude Ogier, Sylvie Pagès, Maxime Galan, Matthieu Barret, and Sophie Gaudriault. rpoB, a promising marker for analyzing the diversity of bacterial communities by amplicon sequencing. *BMC Microbiology*, 19(1):171, 2019. ISSN 1471-2180. doi: 10.1186/s12866-019-1546-z. URL `https://doi.org/10.1186/s12866-019-1546-z`.

Conrad L. Schoch, Stacy Ciufo, Mikhail Domrachev, Carol L. Hotton, Sivakumar Kannan, Rogneda Khovanskaya, Detlef Leipe, Richard Mcveigh, Kathleen O'Neill, Barbara Robbertse, Shobha Sharma, Vladimir Soussov, John P. Sullivan, Lu Sun, Seán Turner, and Ilene Karsch-Mizrachi. NCBI taxonomy: a comprehensive update on curation, resources and tools. *Database: The Journal of Biological Databases and Curation*, 2020:baaa062, 2020. ISSN 1758-0463. doi: 10.1093/database/baaa062.

Martin Steinegger and Johannes Söding. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, 35(11):1026–1028, 2017. ISSN 1546-1696. doi: 10.1038/nbt.3988. URL `https://www.nature.com/articles/nbt.3988`.

The UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research*, 49:D480–D489, 2021. ISSN 0305-1048. doi: 10.1093/nar/gkaa1100. URL `https://doi.org/10.1093/nar/gkaa1100`.

Franziska Wemheuer, Jessica A. Taylor, Rolf Daniel, Emma Johnston, Peter Meinicke, Torsten Thomas, and Bernd Wemheuer. Tax4fun2: prediction of habitat-specific functional profiles and functional redundancy based on 16s rRNA gene sequences. *Environmental Microbiome*, 15(1):11, 2020. ISSN 2524-6372. doi: 10.1186/s40793-020-00358-7. URL `https://doi.org/10.1186/s40793-020-00358-7`.