EsMeCaTa: Estimating metabolic capabilities from taxonomic affiliations
Arnaud Belcour, Baptiste Ruiz, Clémence Frioux, Samuel Blanquart, Anne Siegel

To cite this version:
Arnaud Belcour, Baptiste Ruiz, Clémence Frioux, Samuel Blanquart, Anne Siegel. EsMeCaTa: Estimating metabolic capabilities from taxonomic affiliations. 2022. hal-03697249

HAL Id: hal-03697249
https://hal.archives-ouvertes.fr/hal-03697249
Preprint submitted on 16 Jun 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
EsMeCaTa: Estimating metabolic capabilities from taxonomic affiliations

Arnaud Belcour\textsuperscript{1}, Baptiste Ruiz\textsuperscript{1}, Clément Frioux\textsuperscript{2}, Samuel Blanquart\textsuperscript{1}\textsuperscript{*} and Anne Siegel\textsuperscript{1}\textsuperscript{*}

\textsuperscript{1}Univ Rennes, Inria, CNRS, IRISA, F-35000 Rennes, France.
\textsuperscript{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 metaproteomics 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 Taxo-
nomic affiliations) as a method permitting the estimation of functions inde-
pendently 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 taxo-
nomic 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 Pro-
tomes database find the lowest-ranking taxon for which there is at least one
reference or non-reference proteome in the database and download those pro-
tomes. 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 pro-
tomes 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

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>
<thead>
<tr>
<th>Taxon</th>
<th>Taxon rank</th>
<th>Taxon name</th>
<th>UniProt total references</th>
<th>UniProt proteomes (Busco ≥ 0.8)</th>
<th>EsMeCaTa proteomes</th>
<th>Pan-P</th>
<th>Soft-P</th>
<th>Shell-P</th>
<th>Functional annotation of clusters</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Escherichia</strong></td>
<td>Genus</td>
<td>Escherichia</td>
<td>1,506</td>
<td>3</td>
<td>3</td>
<td>5,221</td>
<td>2,421</td>
<td>3,298</td>
<td>2,153 866 1,084 673 1,096 752</td>
</tr>
<tr>
<td><strong>Citrobacter</strong></td>
<td>Genus</td>
<td>Citrobacter</td>
<td>138</td>
<td>2</td>
<td>2</td>
<td>5,674</td>
<td>2,753</td>
<td>3,128</td>
<td>2,013 772 1,835 708 2,013 772</td>
</tr>
<tr>
<td><strong>Cronobacter</strong></td>
<td>Genus</td>
<td>Cronobacter</td>
<td>15</td>
<td>0</td>
<td>15</td>
<td>9,057</td>
<td>101</td>
<td>7,986</td>
<td>970 677 0 12 600 603</td>
</tr>
<tr>
<td><strong>Lelliottia</strong></td>
<td>Genus</td>
<td>Lelliottia</td>
<td>5</td>
<td>0</td>
<td>5</td>
<td>5,252</td>
<td>2,651</td>
<td>3,245</td>
<td>1,993 756 1,784 687 1,884 718</td>
</tr>
<tr>
<td><strong>Joyeautia</strong></td>
<td>Genus</td>
<td>Joyeautia</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3,915</td>
<td>3,915</td>
<td>3,915</td>
<td>1,993 837 1,983 837 1,983 837</td>
</tr>
<tr>
<td><strong>Edaphovirga</strong></td>
<td>Genus</td>
<td>Family</td>
<td>2,435</td>
<td>42</td>
<td>42</td>
<td>25,822</td>
<td>4,155</td>
<td>2,545</td>
<td>2,253 867 514 191 1,560 595</td>
</tr>
<tr>
<td><strong>Enterobacteriales</strong></td>
<td>Order</td>
<td>Family</td>
<td>2,435</td>
<td>42</td>
<td>42</td>
<td>25,822</td>
<td>4,155</td>
<td>2,545</td>
<td>2,253 867 514 191 1,560 595</td>
</tr>
<tr>
<td><strong>Enterobacteriales</strong></td>
<td>Class</td>
<td>Family</td>
<td>3,028</td>
<td>129</td>
<td>96</td>
<td>53,617</td>
<td>375</td>
<td>2,145</td>
<td>2,475 1,020 487 175 1,063 512</td>
</tr>
<tr>
<td><strong>Gammaproteobacteria</strong></td>
<td>Class</td>
<td>Family</td>
<td>8,271</td>
<td>911</td>
<td>96</td>
<td>85,797</td>
<td>329</td>
<td>1,183</td>
<td>2,050 1,040 387 123 924 327</td>
</tr>
<tr>
<td><strong>Plistozoon</strong></td>
<td>Genus</td>
<td>Phlostozoon</td>
<td>67</td>
<td>17</td>
<td>17</td>
<td>21,287</td>
<td>1,276</td>
<td>10,981</td>
<td>1,996 229 941 104 1,103 200</td>
</tr>
<tr>
<td><strong>Leucocytozoon</strong></td>
<td>Genus</td>
<td>Haemoproteida</td>
<td>68</td>
<td>18</td>
<td>18</td>
<td>22,613</td>
<td>1,076</td>
<td>3,313</td>
<td>1,927 259 546 95 1,090 199</td>
</tr>
<tr>
<td><strong>Corallicola</strong></td>
<td>Genus</td>
<td>Class</td>
<td>30</td>
<td>10</td>
<td>10</td>
<td>46,959</td>
<td>76</td>
<td>1,326</td>
<td>1,919 530 94 14 717 121</td>
</tr>
<tr>
<td><strong>Acavomonas</strong></td>
<td>Genus</td>
<td>Class</td>
<td>124</td>
<td>48</td>
<td>48</td>
<td>248,878</td>
<td>50</td>
<td>785</td>
<td>3,746 924 42 7 418 76</td>
</tr>
</tbody>
</table>

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 (Enterobacteriales 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.
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


