The COG-or: the Improved Functional Annotation of Bacterial Genomes

P. Polakovičová¹, K. Sedlář¹

¹Department of Biomedical Engineering, Faculty of Electrical Engineering and Communication, Brno University of Technology, Czechia

E-mail: xpolak37@vut.cz, sedlar@vut.cz

Abstract—The study of bacterial organisms is given great emphasis in scientific research. By studying the bacterial genome, scientists can gain knowledge of various infectious diseases. Nowadays, it is even possible to edit genes and thus advancing the field of tailored medicine by treating many genetic diseases. Furthermore, it is becoming the norm to utilize bacteria in the industry for their ability to produce important products such as plastics, or fuels. It is therefore essential to extract the necessary information about its genes and other features. This paper focuses mainly on improving the functional annotation of bacterial genomes, classification of protein-coding sequences into clusters of orthologous groups, and visualization of the final annotated genome. For this task, the python package was developed with several functions for working with the outputs of multiple annotation tools, namely eggNOG-mapper, Operon-mapper, and Batch CD-Search. The package was tested on Clostridium diolis DSM 15410 and Clostridium beijerinckii DSM 791.

Keywords—Bacterial genome, Functional annotation, COG, eggNOG-mapper, Operon-mapper, Batch CD-Search

1. INTRODUCTION

The functional annotation of bacterial genomes reveals the role of the genetic structures encoded in the DNA sequence. A typical approach for functional annotation is applying similarity searches against several databases of protein sequences, such as Uniprot [1]. However, this approach has several disadvantages. At larger phylogenetic distances, the high similarity does not necessarily indicate the same function. In this case, identifying orthologs constitutes a fundamental task in accurate function assignment. Orthologs are genes in different species that evolved from a common ancestral gene by speciation. Several databases and tools have been developed that provide precomputed clusters of orthologous groups (COG). The most popular database of these proteins is the COG database [2] which provides manually curated groups split into 26 categories by function. Another well-known database is eggNOG [3], which includes manually curated groups from the COG database as well as automatically generated ones.

Unfortunately, a major disadvantage of COG annotation is that there is currently no suitable tool that can assign COGs to all protein-coding sequences in the genome. Also, the existing tools often differ in their assignment and their output files are confusing and hard to be visualized.

2. MATERIALS AND METHODS

To use the package COG-or, the user must firstly annotate the genome using web-based tools eggNOG-mapper [4], Operon-mapper [5], and Batch CD-Search [6], see Figure 1. The input of eggNOG-mapper and Batch CD-Search are protein-coding sequences (CDS), and the input of Operon-mapper is the genomic sequence with features coordinates in GFF3. The package then works with the decorated.gff file from eggNOG-mapper, predicted_COGs.txt, and ORFs_coordinates.txt from Operon-mapper and with hitdata.txt from Batch CD-Search.

The eggNOG-mapper is a tool for functional annotation of large sets of sequences based on fast orthology assignments using the eggNOG database. The sequence mapping is performed with DIAMOND [7] search. The Operon-mapper is a web server that is not primarily designed for COG assignments, it predicts the operons of any prokaryotic genome sequence. The homology gene
assignments are determined using hmmsearch [8] against models which represent COGs and ROGs (Remained Orthologous Groups). The last tool used is the Batch CD-Search, a web-based tool that can be used for scanning against COG database using RPS-BLAST (a variant of the PSI-BLAST algorithm). The described tools offer different files in various formats, which, moreover, are not structured very well. The designed module contains functions that process outputs from individual programs. The outputs of these functions contain sequence id, source, feature type, start, end, score, strand, frame, and attribute for each annotated feature. The module also contains functions that make it easier to work with the Batch CD-search tool because of its inability to annotate more than 4000 sequences at once. In case the bacterium contains more CDSs, the user can split the single CDS file in half and merge them again after annotation.

The three aforementioned tools use different approaches to assign COGs to individual genes. By combining them, it is possible to achieve more reliable results and a greater percentage of coverage of the whole genome. The consensus function combines all three files and subsequently adds features such as RNA genes and pseudogenes. The output file is ready to be visualized in DNAPlotter [9], which is an interactive Java application for generating circular and linear representations of genomes.

To visualize the genome with the differentiation of each COG category, the consensus function file alone is not sufficient itself. In DNAPlotter, after loading the file, it plots the genome according to the specified features. To color the COG categories and other features, the user has to import another file in the Track Manager option. This file can be generated along with the legend also by the COG-or.

The developed package was tested on two bacteria, both published by the Department of Biomedical Engineering. Clostridium diolis DSM 15410 (accession number CP043998.1) was sequenced by PacBio RSII and Illumina NextSeq, and Clostridium beijerinckii DSM 791 (accession number CP073653.1) was sequenced with Oxford Nanopore MinION and Illumina MiSeq. All necessary input files are available in the GenBank database.

3. RESULTS AND DISCUSSION
The total numbers of CDSs annotated by different tools are listed in Table I. The eggNOG-mapper was able to achieve the highest percentage of assignment of COGs for both genomes (~95%). This is not surprising as the latest version uses novel extensive reference database. On the other hand, CD-search was able to assign only 54.8% and 55.4%, respectively. The reason can be found in limited reference database used by this tool. Nevertheless, even this tool can be used to supplement eggNOG-mapper results because the COG-or package, combining various tools reached the best results, while assigning 97.02% and 97.47% CDSs in C. beijerinckii and C. diolis, respectively. Utilization of non-model organisms is advantageous to demonstrate differences between various annotation tools and improvement by the COG-or. Unfortunately, accuracy of results could not be calculated since there is no model organism with 100% COG annotation. The reason is in scarce utilization of COG system that was rediscovered in 2021 and started to be extensively used in the past few months [2].
Table I: Comparison of tools in COG assignment

<table>
<thead>
<tr>
<th>Bacterium</th>
<th>Number of CDS</th>
<th>eggNOG-mapper</th>
<th>Batch CD-Search</th>
<th>Operon-mapper</th>
<th>COG-or</th>
</tr>
</thead>
<tbody>
<tr>
<td>C. beijerinckii</td>
<td>5061</td>
<td>4823</td>
<td>2771</td>
<td>4526</td>
<td>4910</td>
</tr>
<tr>
<td>C. diolis</td>
<td>5102</td>
<td>4884</td>
<td>2825</td>
<td>4581</td>
<td>4973</td>
</tr>
</tbody>
</table>

Furthermore, the percentages in which the COG-or package changed the COG assignment were calculated, see Table II. This means that for a given CDS the COG assigned by COG-or is different from the COG assigned by particular tool. The highest percentages can be observed for Operon-mapper (exceeding 21%). This might be caused by different primary purpose of the tool, which is operon prediction and not COG annotation. However, correct COG prediction is important for operon inference as all genes in an operon should be involved in the same biological pathway, thus belong to the same COG. The erroneous annotation by Operon-mapper could harm its ability of correct operon structure inference.

Table II: The percentage of changed assignment by the COG-or package

<table>
<thead>
<tr>
<th>Bacterium</th>
<th>Changes in the COG assignment</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>eggNOG-mapper</td>
</tr>
<tr>
<td>C. beijerinckii</td>
<td>2.01%</td>
</tr>
<tr>
<td>C. diolis</td>
<td>1.92%</td>
</tr>
</tbody>
</table>

Finally, the output of an analysis performed with the COG-or was visualized as the genome map of Clostridium beijerinckii DSM 791 using DNAPlotter, see Figure 2. The various colors represent the 26 COG categories as well as RNA genes. The first and the second circles show CDSs on the forward and reverse strands. The third circle represents pseudogenes and in the fourth circle, the RNA genes are plotted. The two inner circles represent GC content and GC skew.

Figure 2: The genome map of Clostridium beijerinckii DSM 791 drawn using DNAPlotter and the COG-or package.
4. CONCLUSIONS
The main topic of this paper was to develop a tool that improves the functional annotation of bacterial genomes using clusters of orthologous groups. The presented python package, the COG-or, can work efficiently with the outputs of various tools to further improve the assignment of individual CDSs to their COG group and category. The package with its several functions was tested on annotation of two bacterial genomes, *Clostridium beijerinckii* DSM 791 and *Clostridium diolis* DSM 15410. Thanks to combination of different approaches, more than 97% of protein-coding sequences were assigned a COG category for both genomes. Subsequently, the genomes were simply plotted by DNAPlotter, and individual features were visually differentiated.

In summary, we proposed the auspicious package that can definitely be utilized in modern bioinformatics. It simplifies the work with different annotation programs, improves their results and even visualize the final data. In the future, it would be desirable to run this tool via command line using the main function that incorporates the whole package. Furthermore, it could have more customize functions for genome visualization.

The COG-or is freely available online at [https://github.com/xpolak37/the-COG-or](https://github.com/xpolak37/the-COG-or)

REFERENCES


