doi 10.18699/bgrs2024-1.1-16

COGcollator 2.0: an improved web server for analysis of distant relationships between homologous protein families

Dibrova D.V.[1, 2]*, Rykov S.Y.[3]

[1] *Belozersky Institute of Physico-Chemical Biology, Lomonosov Moscow State University, Moscow, Russia*

[2] *School of Bioengineering and Bioinformatics, Lomonosov Moscow State University, Moscow, Russia*

[3] *Institute of Philosophy, RAS, Moscow, Russia*

* *udavdasha@belozersky.msu.ru, dibrova.daria@yandex.ru*

Key words: clusters of orthologous groups; phylogenomic analysis; web server; domain structure; proteins

Motivation and Aim: Functional analysis of a protein sequence starts with similarity searches. A decent similarity points to the homology between protein sequences, but does not clarify if they are "the same" protein, and in what extent could the function be transferred from one homologous protein to another. Proteins usually have complex relationships between each other: homologous proteins could be either orthologs or paralogs, which originate through gene duplication and usually diverge more substantially during subsequent specification compared to orthologs. Genes could be lost during evolution, and the horizontal gene transfer which occurs frequently in prokaryotes contribute to gene gain process. These factors provide a sophisticated picture of protein phylogenetic relationships and occurrence in genomes. An approach of constructing Clusters of Orthologous Groups (COGs) and assigning proteins to them proved very fruitful from the beginning [1] and is widely applied until now [2]. We previously created a visual access tool to the data in the COG database, specifically to investigate relationships between different COGs [3]. Here, we provide its major update with an improved web interface which matches the last release of the COG database and has several important additions to our original tool.

Methods and Algorithms: We based our research on an assignment of proteins from 1309 representative prokaryotic genomes to COGs [2]. We constructed multiple sequence alignments of proteins belonging to each COG with Muscle 3.8.31 [4], and corrected multiple alignments when needed. Then we performed a search with the HMMer 3.1 [5] for each of profile HMMs on three sequence samples (Fig. 1): (1) proteins assigned to COGs in 1309 prokaryotic genomes, (2) proteins encoded in the same genomes which are not assigned to COGs, (3) eukaryotic proteins from the representative sample of 102 genomes. In multiple cases when genomes of organelles were missing from eukaryotic genome assembly, we searched for these genomes and added them manually (this was done for 36 mitochondrial genomes and 5 chloroplast genomes). Our tool shows two graphs for each COG, one with prokaryotic hits and another with eukaryotic hits of corresponding profile HMM. Each graph shows how the overall *hmmsearch* score decreases with an increase of hit rank. Both graphs are interactive: a user can obtain a protein sequence from each individual dot or multiple sequences from a range of dots, zoom graphs, browse additional information about hits, hide data which correspond to any COG in prokaryotic panel or any taxon in eukaryotic panel.

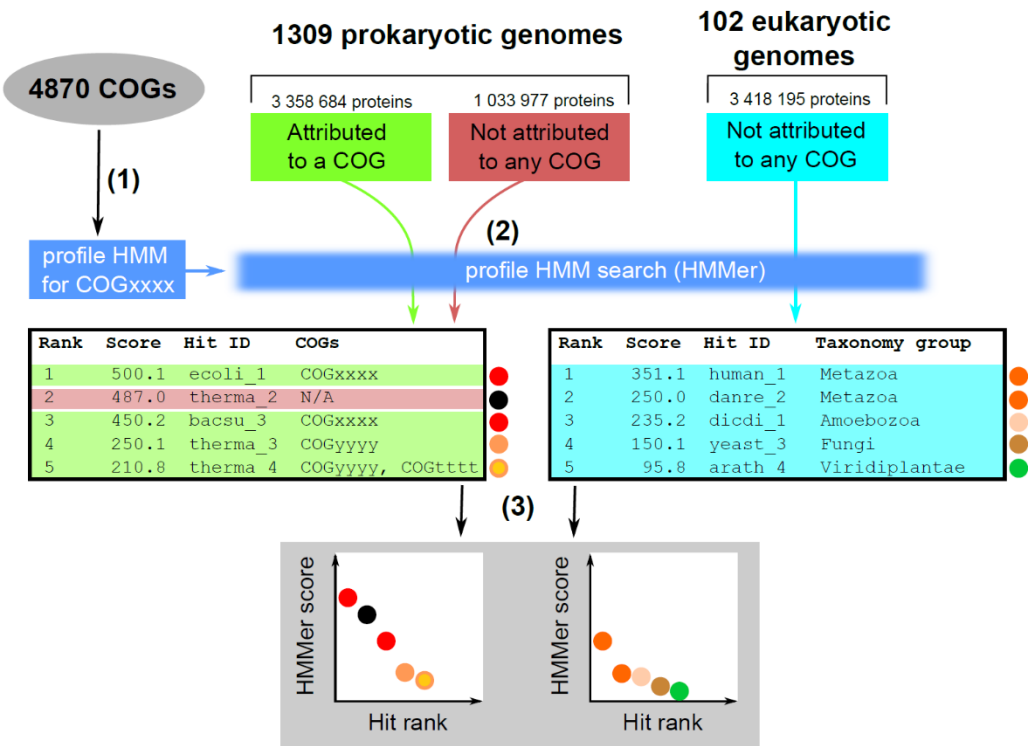

Fig. 1. Pipeline of the COGcollator 2.0. **(1)** Profile HMMs are obtained for each COG either from automatically constructed multiple alignments as described before [3], or from manually curated alignments. **(2)** Ranked hit tables for a COG are obtained for each sequence database, prokaryotic hit tables are mixed, ranked and depicted separately from the eukaryotic hit table. **(3)** Prokaryotic proteins are colored according to the COGs to which they are assigned and eukaryotic proteins are colored according to the high-level taxonomy of corresponding species

Results: We constructed profile HMM for almost all COGs from the latest COG database release (4870 out of 4877), then we calculated sensitivity of each profile and tried to improve profiles which poorly identified proteins assigned to corresponding COGs by manual inspection of alignments (126 cases it total). Detailed information on a process of each profile construction is available through the COGcollator interface. Then we performed a search with the HMMer [4] for each of profile HMMs on three sequence samples, as described above. Example of the COGcollator interface with the result for COG0055 (the catalytic β-subunit of the F_OF_1-type ATP synthase) is shown in Figure 2. It is possible to switch between two e-value thresholds for hits to be shown (1e-1 and 1e-5), thus we enabled non-strict threshold 0.1 to search for distantly related sequences and/or truncated versions of proteins. We also allowed a user to filter out COG annotations which do not intersect by protein coordinates with the hits of target COG profile HMM. This reduces number of shown COGs which are found only as fusions to the target COG.

Conclusion: The web server for COGcollator 2.0 is freely available at: http://boabio.belozersky.msu.ru/en/COGcollator. All constructed profile HMMs for 4870 COGs are united in the database and are available for download. We also provide another option to access our data: DomainAnalyser, a highly versatile sequence search tool which is based on the *hmmscan* from HMMer package and allows to visualize a domain structure of a protein set in both interactive and downloadable way.

DomainAnalyser supports our database of profile HMM for COGs as well as a widely used Pfam domain database [6]. This tool is freely available at: http://boabio.belozersky.msu.ru/en/DomainAnalyser.

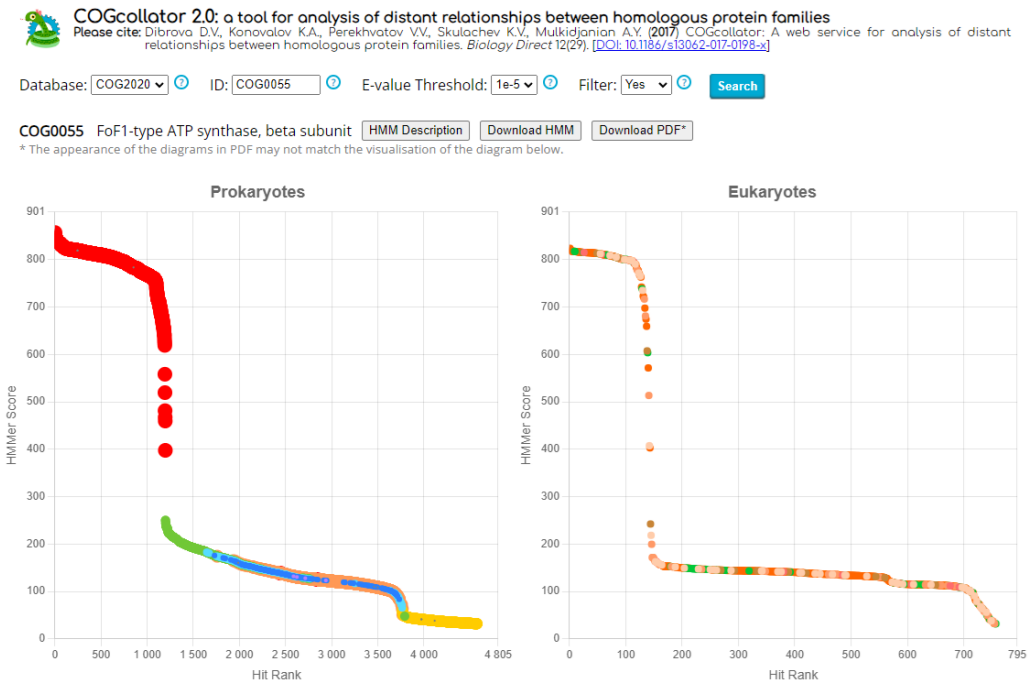

Fig. 2. Part of the COGcollator 2.0 interface: HMMer score vs. hit rank graphs for the COG0055 profile HMM at e-value threshold of 1e-5. **Left panel**, prokaryotic hits. The dots are colored according to the COG database assignment. **Right panel**, eukaryotic hits. The dots are colored according to the taxonomy of the corresponding organism

Funding: The development of original version of COGcollator and idea behind it was supported by the Dmitry Zimin "Dynasty" Foundation (D.V.D.). COGcollator 1.0 was supported by the Russian Science Foundation (RSCF), grant 14–50–00029 (D.V.D.).

References

1. Tatusov R.L., Koonin E.V., Lipman D.J. A genomic perspective on protein families. *Science.* 1997;278(5338):631-637. doi 10.1126/science.278.5338.631
2. Galperin M.Y., Wolf Y.I., Makarova K.S., Vera Alvarez R., Landsman D., Koonin E.V. COG database update: focus on microbial diversity, model organisms, and widespread pathogens. *Nucleic Acids Res.* 2021;49(D1):D274-D281. doi 10.1093/nar/gkaa1018
3. Dibrova D.V., Konovalov K.A., Perekhvatov V.V., Skulachev K.V., Mulkidjanian A.Y. COGcollator: a web server for analysis of distant relationships between homologous protein families. *Biol Direct.* 2017;12(1):29. doi 10.1186/s13062-017-0198-x
4. Edgar R.C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics.* 2004;5:113. doi 10.1186/1471-2105-5-113
5. Potter S.C., Luciani A., Eddy S.R., Park Y., Lopez R., Finn R.D. HMMER web server: 2018 update. *Nucleic Acids Res.* 2018;46(W1):W200-W204. doi 10.1093/nar/gky448
6. Mistry J., Chuguransky S., Williams L., Qureshi M., Salazar G.A., Sonnhammer E.L.L., Tosatto S.C.E., Paladin L., Raj S., Richardson L.J., Finn R.D., Bateman A. Pfam: The protein families database in 2021. *Nucleic Acids Res.* 2021;49(D1):D412-D419. doi 10.1093/nar/gkaa913