OLESA: Operon Loci Examination and Sorting Application

Olesya I. Klimchuk

School of Bioengineering and Bioinformatics, Lomonosov Moscow State University, Moscow 119992, Russia

olesya@fbb.msu.ru

Daria V. Dibrova

Belozersky Institute of Physico-Chemical Biology, Lomonosov Moscow State University, Moscow 119992, Russia

udavdasha@belozerky.msu.ru

Armen Y. Mulkidjanian

School of Physics, University of Osnabruck, D-49069 Osnabruck, Germany, amulkid@uos.de

Belozersky Institute of Physico-Chemical Biology and

School of Bioengineering and Bioinformatics, Lomonosov Moscow State University, Moscow 119992, Russia

The genes of prokaryotic organisms can be organized in conserved operon structures. The clustering of genes, which enables coordinated regulation of functionally coupled genes [1, 2], also may result in the horizontal transfer of whole operons between prokaryotic genomes [3, 4].

Widely used tools for detection of conserved genetic neighborhoods and their inspection, for example STRING [7] and SEED [8] servers, provide a useful overview of the overall gene distribution. Here we present a program OLESA (after Operon Loci Examination and Sorting Application) that has been developed for detailed analysis of genetic neighborhoods in relation with phylogenomic analysis (the main window of the program is depicted in *Figure 1*).

In OLESA, the set of nucleotide entries for analysis (chromosomes, plasmids and contigs) could be defined by the user; in addition, the program currently incorporates a representative sample of 711 completely sequenced prokaryotic genomes previously selected for the latest release of the COG (Clusters of Orthologous Groups) database [5]. Complete proteomes were extracted from these genomes and domain architectures were assigned to each protein using the Pfam domain database [6] and the COGs database [5]. Overlapping domains are filtered in OLESA with an additional algorithm with respect to the relative size of the overlapping region and each hit score. Operon structures are determined by using criteria proposed by

Overbeek [10]; the user may also choose to view the complete neighborhood of a gene instead of restricting to the operons only. As an input, OLESA can accept a plain text file of protein accession numbers, or an output file of a phylogenetic tree produced by the PHYLIP package [9], or a sequence of a query protein in a FASTA format (in this mode a classical similarity search is launched). The program returns an ordered list of operon structures (or gene surroundings) where target genes are colored according to their domain structures. The neighboring genes of interest can be also colored on demand. The color legend with the operons (or gene surroundings) can be saved in a vector format (see *Figure 1*).

OLESA has at least two advantages as compared to other currently available tools. First, such tools routinely use the results of a BLAST search for detecting the sequence similarity, which could lead to misinterpretations when highly similar but distinct proteins are erroneously recognized as "the same" protein. We have tried to avoid such errors by assigning each protein to a particular domain architecture (set of detectable domains could be customized, but we have used either a full set of Pfam domains [6] or COGs [5]).

Second, routinely used tools search for proteins similar to a query and assign the genetic neighborhoods of the hits to the taxonomy of corresponding organisms. While preserving this useful functionality (see *Figure 1A*), OLESA is focused on depicting the genetic neighborhoods for all the proteins in a set. Therefore, a further option in OLESA is the possibility to assign the gene neighborhood information to a list of proteins ordered by a rectangular phylogenetic tree. This option helps to detect a conservative gene order within a particular tree clade. We suggest that the mapping of protein gene neighborhoods on a phylogenetic tree of a particular protein could be important for specific assignments of similar proteins to different functions. A preserved gene order could confirm reliability of a particular tree clade, while a different gene order may indicate a previously unknown operon (an example of conserved operons of phylogenetically distinct F-type and N-type rotary ATPases [3], as mapped on a phylogenetic tree, is shown in *Figure 2*).

In OLESA, an automatic mapping of genome context on a phylogenetic tree could be performed for any protein; it becomes particularly interesting in the case of large protein complexes. The current version of the main OLESA program and the supplementary scripts (written in Python 2.7) are available from the authors upon request.

Acknowledgements:

This work was supported by a Grant of the Russian Science Foundation (14-50-00029).

References:

1. R. Overbeek et al. (1999) The use of gene clusters to infer functional coupling, *Proc. Natl. Acad. Sci.*, **96:** 2896-2901.

2. M. Galperin and E. Koonin (2000) Who's your neighbor? New computational approaches for functional genomics, *Nat. biotech.*, **18:** 609-613.

3. D. Dibrova, M. Galperin and A. Mulkidjanian (2010) Characterization of the N-ATPase, a distinct, laterally transferred Na^+-translocating form of the bacterial F-type membrane ATPase, *Bioinformatics*, **26:** 1473-1476.

4. J. Xiong, K. Inoue and C. Bauer (1998) Tracking molecular evolution of photosynthesis by characterization of a major photosynthesis gene cluster from *Heliobacillus mobilis*, *Proc. Natl. Acad. Sci.*, **95:** 14851-14856.

5. M. Galperin et al. (2015) Expanded microbial genome coverage and improved protein family annotation in the COG database, *NAR*, **43:** D261–D269.

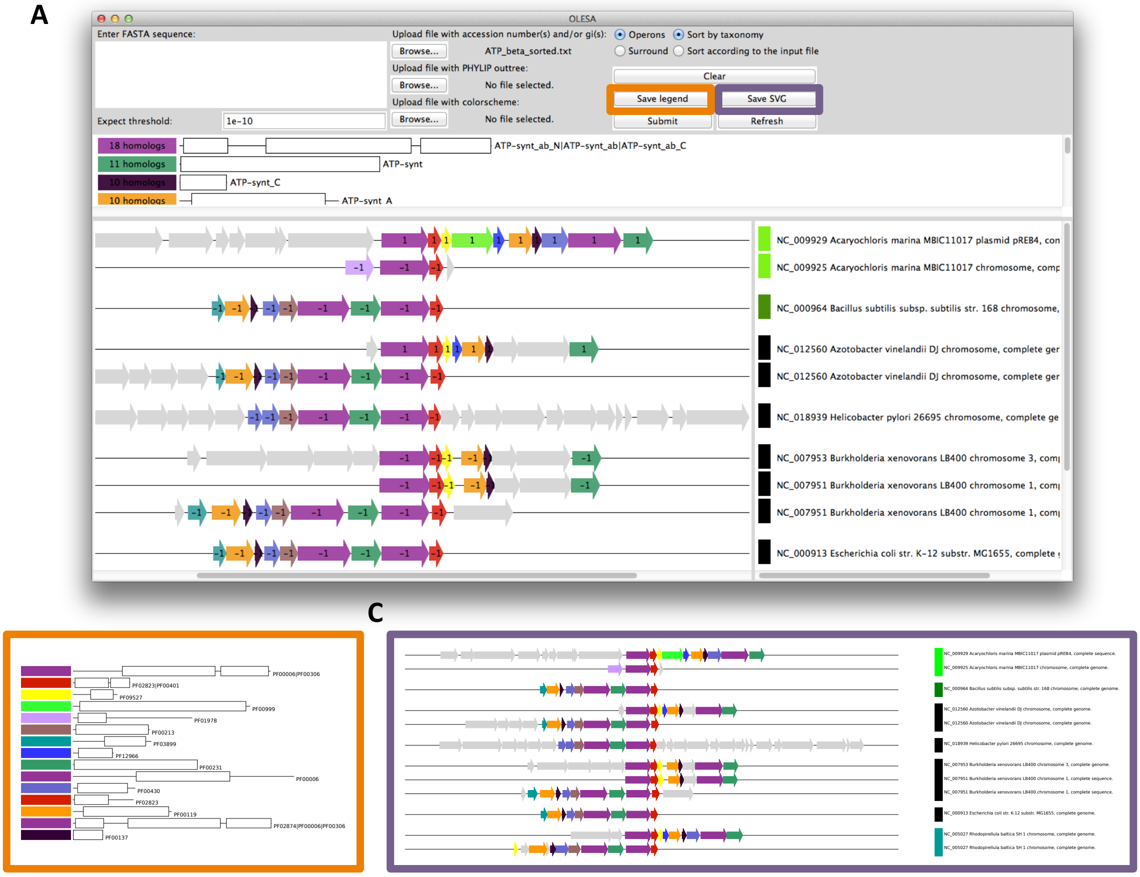
6. R. Finn et al. (2014) Pfam: the protein families database, *NAR*, **42:** D222-D230.

7. D. Szklarczyk at al. (2015) STRING v10: protein–protein interaction networks, integrated over the tree of life, *NAR*, **43:** D447–D452.

8. R. Overbeek et al. (2014) The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST), *NAR*, **42:** D206-D214.

9. D. Plotree and D. Plotgram (1989) PHYLIP-phylogeny inference package (version 3.2), *Cladistics*, **5:** 163-166.

10. R. Overbeek et al. (1999) Use of contiguity on the chromosome to predict functional coupling, *In silico biology*, **1:** 93-108.



Enter FASTA sequence:

Upload file with accession number(s) and/or gi(s): ● Operons ● Sort by taxonomy

Browse... ATP_beta_sorted.txt ○ Surround ○ Sort according to the input file

Upload file with PHYLIP outtree:

Browse... No file selected.

Upload file with colorscheme:

Browse... No file selected.

Expect threshold: 1e-10

Clear | Save legend | Save SVG | Submit | Refresh

18 homologs ATP-synt_ab_N|ATP-synt_ab|ATP-synt_ab_C

11 homologs ATP-synt

10 homologs ATP-synt_C

10 homologs ATP-synt_A

NC_009929 Acaryochloris marina MBIC11017 plasmid pREB4, con

NC_009925 Acaryochloris marina MBIC11017 chromosome, comp

NC_000964 Bacillus subtilis subsp. subtilis str. 168 chromosome,

NC_012560 Azotobacter vinelandii DJ chromosome, complete ger

NC_012560 Azotobacter vinelandii DJ chromosome, complete ger

NC_018939 Helicobacter pylori 26695 chromosome, complete ge

NC_007953 Burkholderia xenovorans LB400 chromosome 3, com

NC_007951 Burkholderia xenovorans LB400 chromosome 1, com

NC_007951 Burkholderia xenovorans LB400 chromosome 1, com

NC_000913 Escherichia coli str. K-12 substr. MG1655, complete

Figure 1. The OLESA program and some its options. (**A**) The main window of the program. The top panel manages the query input and the selection of options, the central panel shows colors as assigned to each domain architecture (these colors could be managed by the user or chosen randomly). The bottom panel shows the resulting genetic neighborhoods or operons that are sorted according to the user's choice, either by taxonomy of organisms or by the initial input order; (**B**) a saved legend; (**C**) saved operons.

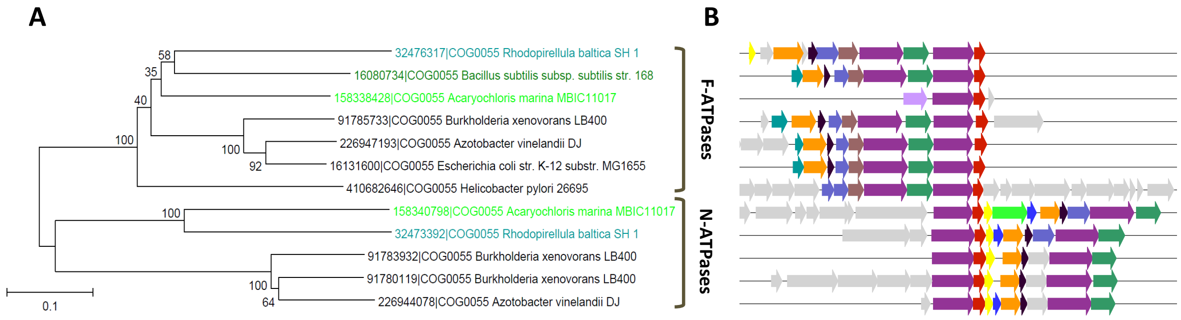

A
B
58
35
40
100
100
92
100
100
64
0.1
32476317|COG0055 Rhodopirellula baltica SH 1
16080734|COG0055 Bacillus subtilis subsp. subtilis str. 168
158338428|COG0055 Acaryochloris marina MBIC11017
91785733|COG0055 Burkholderia xenovorans LB400
226947193|COG0055 Azotobacter vinelandii DJ
16131600|COG0055 Escherichia coli str. K-12 substr. MG1655
410682646|COG0055 Helicobacter pylori 26695
158340798|COG0055 Acaryochloris marina MBIC11017
32473392|COG0055 Rhodopirellula baltica SH 1
91783932|COG0055 Burkholderia xenovorans LB400
91780119|COG0055 Burkholderia xenovorans LB400
226944078|COG0055 Azotobacter vinelandii DJ
F-ATPases
N-ATPases

Figure 2. The membrane rotary F-type and N-type ATPases form individual clades on the phylogenetic tree and have different operons. (**A**) The phylogenetic tree of the β-subunits of F-ATPases and N-ATPases; (**B**) the operons of F-ATPases and N-ATPases.