Family-Companion: analyse, visualise, browse, query and share your homology clusters

Ludovic Cottret[1], Martial Briand[2], Corinne Rancurel[3] and Sébastien Carrere[1,*]

[1]LIPM, Université de Toulouse, INRA, CNRS, Castanet-Tolosan, France

[2]IRHS, INRA, AGROCAMPUS-Ouest, Université d'Angers, SFR4207 QUASAV, 42, rue Georges Morel, 49071 Beaucouzé, France

[3]INRA, Université Côte d'Azur, CNRS, ISA, France

*To whom correspondence should be addressed. Sebastien.Carrere@inra.fr

Abstract

Identifying homology groups in predicted proteomes from different biological sources allows biologists to address questions as diverse as inferring species-specific proteins or retracing the phylogeny of gene families. Nowadays, command-line software exists to infer homology clusters. However, computing and interpreting homology groups with this software remains challenging for biologists and requires computational skills.

We propose Family-Companion, a web server dedicated to the computation, the analysis and the exploration of homology clusters. Family-Companion aims to fill the gap between analytic software and databases presenting orthologous groups based on a set of public data. It offers a user-friendly interface to launch or upload precomputed homology cluster analysis, to explore and share the results with other users. The exploration of the results is highly facilitated by interactive solutions to visualize proteome intersections via Venn diagrams, phylogenetic trees, multiple alignments, and also by querying the results by blast or by keywords.

Family-Companion is available at http://family-companion.toulouse.inra.fr with a demo dataset and a set of video tutorials. Source code and installation protocol can be found at https://framagit.org/BBRIC/family-companion/. A container-based package simplifies the installation of the web-suite.

1. Introduction

The proteome of an organism is the set of proteins encoded by its genome. Proteins that derive from a common ancestral protein are called homologous proteins. A common assumption is to consider that orthologous proteins (homologous proteins that diverged after a speciation event) have a similar biological role in different species. In contrast, paralogous proteins are homologous proteins that have diverged after a duplication event within one species. By comparing all protein sequences found in a set of organisms, it is possible to classify them in homology (orthology/paralogy) clusters. These can then be analysed to identify, for instance, proteins specific to a group of species that share a biological trait of interest (e.g. pathogenicity or symbiosis (1, 2). Besides, identification of universally conserved single copy proteins can help to infer species phylogeny (3). Finally, analysis of presence/absence and copy number of proteins in different homology groups across species allows their evolutionary history to be inferred (4).

The identification of homology groups from proteomes can be performed thanks to command-line software based on sequence comparison and clustering (5–7). However, the results, in the form of complex and non standardized text files, are hardly exploitable or interpretable without computational skills. Existing web resources provide solutions to facilitate the exploration of pre-computed homology groups. Orthomcl-db (8) and OrthoDB (9) allow complex queries to be made from several characteristics such as the inclusion or exclusion of taxa, statistics or functional annotations. EggNogg (10) proposes an online visualization of trees that show duplication and speciation events to distinguish between paralogs and orthologs. Orthomcl-db, OrthoDB and EggNogg all propose a catalogue of orthologs spanning a large scope of species. They also allow user-provided proteomes to be mapped onto pre-computed homology groups.

However, none of these web servers provide a combined interface to calculate homology groups from a user-defined set of proteomes of interest, and to explore them through online queries and visualisation tools. OrthoVenn (11) provides tools for cluster annotation and Venn diagrams from user-provided protein sequences, even if the number of proteomes is limited to 6. Spocs (12) provides a graph-based ortholog prediction method and tools to visualise predicted ortholog/paralog relationships.

In order to supplement existing offers, we propose Family-Companion (mentioned below as F-C), a rich web application that allows computation, exploration and sharing of homology groups.

2. Results

An overview of the analysis flow in F-C is illustrated in Supp Fig 1.

2.1. Homology cluster computation and automatic analyses

Inference of homology clusters

A dynamic online form enables users to upload an unlimited number of proteomes in FASTA format. Once all the proteomes of interest have been uploaded, homology groups can be inferred by automatically launching OrthoMCL with parameters that can be easily tuned (Supp. Fig. 2 and Supp. Fig. 3). If the user has already generated homology groups using OrthoMCL v1.4 or V 2, Orthofinder Synergy or any other tool generating a compatible format, results can be directly uploaded to Family Companion for exploration and visualization (Supp. Fig. 2).

Alignments and phylogenetic trees

For each homology group, a multi-alignment and a phylogenetic tree are computed to facilitate the analysis of protein families.

Generation of core, pan proteomes and species-specific proteins

The core-proteome corresponds to single-copy proteins conserved in all analysed proteomes. This set of one-to-one conserved orthologous proteins can be particularly useful to infer species phylogenies. To facilitate their reconstruction, F-C provides a super alignment built by concatening alignments of proteins involved in each one-to-one ortholog cluster. Of note, the content of the core proteome depends highly on the parameters used for protein clustering.
The pan-proteome provides a representation of the diversity of proteins present in the set of organisms of interest. It is composed of one representative of each homology group and all the species-specific single copy proteins. The user can optionally tag a ‘reference’ proteome. In this case the proteins from the reference proteome will be selected as representative in each homology group in which the species is present. Otherwise, a matrix is built by computing a distance between each sequence. The sequence with the minimal sum of distances is chosen as representative.
For each proteome, a set of specific proteins is computed. It is composed of single-copy proteome-specific proteins (i.e. all the proteins that are not present in any of the homology groups) and multi-copy proteome-specific proteins (i.e. proteins in homology groups present in a unique species).
The core-proteome super-alignment, the pan-proteome and the specific proteins are all downloadable in FASTA format (Supp. Fig. 4).

Generation of phylogenetic profiles

In order to assess the presence, expansion or reduction of protein families across species, F-C computes presence/absence and abundance matrices. In these downloadable matrices, rows correspond to the homology groups, and columns to taxa. The cells either include a Boolean value (presence / absence) or copy numbers of proteins belonging to a homology group in the corresponding taxon.

Functional annotation of homology clusters

InterPro annotations (13) can be uploaded in the main F-C online form (Supp. Fig. 3). Then, functional annotations are assigned to homology groups by concatenating the annotations of the proteins composing them.

2. 2 Interactive visualization solutions

Alignments and phylogenetic trees
For each homology group, a multi-alignment and a phylogenetic tree are displayed in an interactive way (Supp. Fig. 5). This facilitates the analysis of the evolutionary events (speciation, duplication) that link the proteins of the homology group.

Visualization of phylogenetic profiles
Each phylogenetic profile can be visualized online with an interactive heatmap (Supp. Fig. 6) where clustering trees for rows (homology groups) and columns (proteomes) are displayed.

Interactive charts and Venn diagrams
A series of interactive charts is produced at each analysis to visualize differences between the selected proteomes. For instance, one can visualize the number of proteins classified into homology groups, or specific to particular taxa (Supp. Fig. 7).
F-C provides interactive Venn diagrams to compare proteomes based on their content in homology groups (Supp. Fig. 8). By clicking on diagram intersections, links to corresponding homology clusters are displayed.

2. 3 Browse and Query

The web interface provides an immediate navigation between tables listing proteins, homology groups, alignments, trees and phylogenetic profiles (Supp. Fig. 9). In order to explore results of homology groups in more detail, F-C provides a Query Builder. A search engine returns the groups to which a particular protein belongs via its accession number or a keyword present in its annotation. An integrated Blast (14) server can also find homology clusters through sequence similarity to a query protein (Supp. Fig. 10).
Moreover, a form allows queries to be graphically constructed to extract the homology clusters present in a set of proteomes and absent in another set (Supp. Fig. 11). This can be used to investigate clade-specific functions.

2. 4 Share private analyses

By default, new analyses are private and only accessible by the authenticated user who uploaded data. However, this user has the possibility to generate a public URL in order to make the results available to others. This stable URL allows the link to be integrated into analyses in other pages or simply to share them with collaborators. The sharing of an analysis can also be cancelled by a simple click (Supp. Fig. 12).

3. Implementation

The back-end part is implemented in Perl. The front-end part is implemented in Javascript. The framework Ext Js 6.0[1] was used for the user interface. Interactive charts use the HighCharts library[2]. Venn diagrams are displayed thanks to the Jvenn library (15). Interactive heatmaps are built with the Inchlib JavaScript and python libraries (16). Alignments are displayed with the BioJs MSA viewer library (17). Phylogenetic trees are visualised thanks to the Phylocanvas javascript library[3]. New homology clusters are inferred thanks to a modified version of OrthoMcl V1.4 (5) using Diamond (18) and Parallel (19) software to speed up similarity computing between sequences. Multiple alignments are computed with Mafft (20) and cleaned with Trimal (21). Phylogenetic trees are inferred by FastTree (22).

4. Conclusion

[1] https://www.sencha.com/products/extjs

[2] https://www.highcharts.com

[3] http://phylocanvas.org/

The user-friendly interface of F-C provides powerful, interactive web solutions to explore homology clusters and address the most common questions in this kind of approach. Another interesting feature is to handle homology clusters for a set of proteomes provided by the users themselves, which complements the existing solutions. In addition, sharing private analyses facilitates collaborative work.
Future improvements can be imagined. Annotation transfer could be facilitated between the proteomes. Statistical analyses such as Principal Component Analysis could be automatically performed to describe proteomes according to the homology clusters. Finally, an incremental update of clusters and analyses could save time when new proteomes are provided.

Acknowledgements

We are grateful to the genotoul bioinformatics platform Toulouse Midi-Pyrenees (Bioinfo Genotoul) for hosting F-C. We thank Etienne Danchin, Matthieu Barret and Jérôme Gouzy for testing F-C, reading the manuscript and providing helpful comments and Clare Gough for checking language mistakes.

Funding

This work has been supported by BBRIC network (INRA/SPE).
Conflict of Interest: none declared.

References

1. Cesbron,S., Briand,M., Essakhi,S., Gironde,S., Boureau,T., Manceau,C., Fischer-Le Saux,M. and Jacques,M.-A. (2015) Comparative Genomics of Pathogenic and Nonpathogenic Strains of Xanthomonas arboricola Unveil Molecular and Evolutionary Events Linked to Pathoadaptation. *Front. Plant Sci.*, **6**, 1126.
https://doi.org/10.3389/fpls.2015.01126
http://www.ncbi.nlm.nih.gov/pubmed/26734033

2. Roux,B., Bolot,S., Guy,E., Denancé,N., Lautier,M., Jardinaud,M.-F., Fischer-Le Saux,M., Portier,P., Jacques,M.-A., Gagnevin,L., *et al.* (2015) Genomics and transcriptomics of Xanthomonas campestris species challenge the concept of core type III effectome. *BMC Genomics*, **16**, 975.
https://doi.org/10.1186/s12864-015-2190-0
http://www.ncbi.nlm.nih.gov/pubmed/26581393

3. Wu,M. and Scott,A.J. (2012) Phylogenomic analysis of bacterial and archaeal sequences with AMPHORA2. *Bioinformatics*, **28**, 1033–1034.
https://doi.org/10.1093/bioinformatics/bts079

4. Dettman,J.R., Rodrigue,N., Aaron,S.D. and Kassen,R. (2013) Evolutionary genomics of epidemic and nonepidemic strains of Pseudomonas aeruginosa. *Proc. Natl. Acad. Sci. U. S. A.*, **110**, 21065–70.
https://doi.org/10.1073/pnas.1307862110
http://www.ncbi.nlm.nih.gov/pubmed/24324153

5. Li,L., Stoeckert,C.J., Roos,D.S. and Roos,D.S. (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.*, **13**, 2178–89.
https://doi.org/10.1101/gr.1224503
http://www.ncbi.nlm.nih.gov/pubmed/12952885

6. Emms,D.M. and Kelly,S. (2015) OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome*

Biol., **16**, 157.
https://doi.org/10.1186/s13059-015-0721-2
http://www.ncbi.nlm.nih.gov/pubmed/26243257

7. Wapinski,I., Pfeffer,A., Friedman,N. and Regev,A. (2007) Automatic genome-wide reconstruction of phylogenetic gene trees. *Bioinformatics*, **23**, i549–i558.
https://doi.org/10.1093/bioinformatics/btm193

8. Chen,F., Mackey,A.J., Stoeckert,C.J., Roos,D.S. and Roos,D.S. (2006) OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res.*, **34**, D363-8.
https://doi.org/10.1093/nar/gkj123
http://www.ncbi.nlm.nih.gov/pubmed/16381887

9. Kriventseva,E. V, Tegenfeldt,F., Petty,T.J., Waterhouse,R.M., Simão,F.A., Pozdnyakov,I.A., Ioannidis,P. and Zdobnov,E.M. (2015) OrthoDB v8: update of the hierarchical catalog of orthologs and the underlying free software. *Nucleic Acids Res.*, **43**, D250-6.
https://doi.org/10.1093/nar/gku1220
http://www.ncbi.nlm.nih.gov/pubmed/25428351

10. Huerta-Cepas,J., Szklarczyk,D., Forslund,K., Cook,H., Heller,D., Walter,M.C., Rattei,T., Mende,D.R., Sunagawa,S., Kuhn,M., *et al.* (2016) eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res.*, **44**, D286-93.
https://doi.org/10.1093/nar/gkv1248
http://www.ncbi.nlm.nih.gov/pubmed/26582926

11. Wang,Y., Coleman-Derr,D., Chen,G. and Gu,Y.Q. (2015) OrthoVenn: a web server for genome wide comparison and annotation of orthologous clusters across multiple species. *Nucleic Acids Res.*, **43**, W78-84.
https://doi.org/10.1093/nar/gkv487
http://www.ncbi.nlm.nih.gov/pubmed/25964301

12. Curtis,D.S., Phillips,A.R., Callister,S.J., Conlan,S. and McCue,L.A. (2013) SPOCS: software for predicting and visualizing orthology/paralogy relationships among genomes. *Bioinformatics*, **29**, 2641–2.
https://doi.org/10.1093/bioinformatics/btt454
http://www.ncbi.nlm.nih.gov/pubmed/23956303

13. Finn,R.D., Attwood,T.K., Babbitt,P.C., Bateman,A., Bork,P., Bridge,A.J., Chang,H.-Y., Dosztányi,Z., El-Gebali,S., Fraser,M., *et al.* (2017) InterPro in 2017-beyond protein family and domain annotations. *Nucleic Acids Res.*, **45**, D190–D199.
https://doi.org/10.1093/nar/gkw1107
http://www.ncbi.nlm.nih.gov/pubmed/27899635

14. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
https://doi.org/10.1016/S0022-2836(05)80360-2
http://www.ncbi.nlm.nih.gov/pubmed/2231712

15. Bardou,P., Mariette,J., Escudié,F., Djemiel,C. and Klopp,C. (2014) jvenn: an interactive Venn diagram viewer. *BMC Bioinformatics*, **15**, 293.
https://doi.org/10.1186/1471-2105-15-293

16. Skuta,C., Bartůněk,P. and Svozil,D. (2014) InCHlib - interactive cluster heatmap for web applications. *J. Cheminform.*, **6**, 44.
https://doi.org/10.1186/s13321-014-0044-4

http://www.ncbi.nlm.nih.gov/pubmed/25264459

17. Yachdav,G., Wilzbach,S., Rauscher,B., Sheridan,R., Sillitoe,I., Procter,J., Lewis,S.E., Rost,B. and Goldberg,T. (2016) MSAViewer: interactive JavaScript visualization of multiple sequence alignments. *Bioinformatics*, **32**, 3501–3503. https://doi.org/10.1093/bioinformatics/btw474 http://www.ncbi.nlm.nih.gov/pubmed/27412096

18. Buchfink,B., Xie,C. and Huson,D.H. (2015) Fast and sensitive protein alignment using DIAMOND. *Nat. Methods*, **12**, 59–60. https://doi.org/10.1038/nmeth.3176

19. Tange,O. (2011) GNU Parallel - The Command-Line Power Tool. *;login USENIX Mag.*, **36**, 42–47.

20. Katoh,K., Misawa,K., Kuma,K. and Miyata,T. (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.*, **30**, 3059–66. http://www.ncbi.nlm.nih.gov/pubmed/12136088

21. Capella-Gutiérrez,S., Silla-Martínez,J.M. and Gabaldón,T. (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, **25**, 1972–3. https://doi.org/10.1093/bioinformatics/btp348 http://www.ncbi.nlm.nih.gov/pubmed/19505945

22. Price,M.N., Dehal,P.S. and Arkin,A.P. (2010) FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One*, **5**, e9490. https://doi.org/10.1371/journal.pone.0009490 http://www.ncbi.nlm.nih.gov/pubmed/20224823