OrthoInspector 2.0: Software and database updates

Benjamin Linard, Alexis Allot, Raphaël Schneider, Can Morel, Raymond Ripp, Marc Bigler, Julie D Thompson, Olivier Poch, Odile Lecompte

Application note

OrthoInspector 2.0: software and database updates

Benjamin Linard[1,2], Alexis Allot[1], Raphaël Schneider[1], Can Morel[1], Raymond Ripp[1], Marc Bigler[1], Julie D Thompson[1], Olivier Poch[1] and Odile Lecompte[1*]

[1]LBGI, Computer Science Department, ICube, UMR 7357, University of Strasbourg, CNRS, Fédération de médecine translationnelle, 4 rue Kirschleger 67085 Strasbourg, France. ,

[2]Department of Life Sciences, Natural History Museum, Cromwell Road, London SW7 5BD,

ABSTRACT

Summary: We previously developed OrthoInspector, a software suite incorporating an original algorithm for the rapid detection of orthology and inparalogy relations between different species. We have added new functional tools and considerably extended the databases of pre-computed orthology/inparalogy relationships.

Availability: Software and databases are freely available at http://lbgi.fr/orthoinspector with all major browsers supported.

Contact: odile.lecompte@unistra.fr

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

High throughput comparative analyses, functional annotations or evolutionary studies involve massive transfers of information between organisms using orthology inference. As defined by Fitch (Fitch, 1970), orthologs are homologous genes that diverged from an ancestral speciation event, while paralogs emerged from a duplication event. These definitions are based on evolutionary history , but it is widely accepted that orthologs generally share similar functions whereas paralogs can potentially evolve new functions. Numerous algorithms based on the results of Blast searches were developed to infer orthology relations (see Kristensen *et al.*, 2011; Altenhoff and Dessimoz, 2012 for reviews). We previously developed an orthology inference algorithm and implemented it in the OrthoInspector package (Linard, et al., 2011). This stand-alone software predicts large-scale orthology and inparalogy relationships between a set of proteomes, maintaining a balance between sensitivity and specificity (Linard *et al.*, 2011; Dalquen *et al.*, 2013). Here we describe the improvements that we have implemented in version 2.0 of OrthoInspector.

2 DISTINCTIVE FEATURES

2.1 Interoperability and optimization

OrthoInspector 2.0 software design focuses on large dataset handling but allows predictions on a desktop computer or small server (tutorials are available on the website). Interoperability was also a priority (figure 1A). The software can be run on any Java enabled system (Unix, MacOS, Windows…), supports the OrthoXML format (Schmitt *et al.*, 2011) and delegates data management to any SQL database chosen by the user (with an extended support for PostgreSQL and MySQL). The use of a SQL engine takes advantage of their optimized dataset manipulation capabilities. Consequently, large-scale predictions do not require any significant CPU or memory resources.

2.2 Eukaryote and prokaryote databases

We have constructed 3 orthology databases with OrthoInspector (figure 1B). The first database, named “Prokaryotes”, contains orthologs between 120 Archaea and 1568 Bacteria proteomes (Suppl. File S1). The second dataset, named “Eukaryota”, contains 259 complete proteomes and covers all main eukaryotic phyla, from unicellular organisms to plants, fungi and metazoan. The third dataset, named “Quest For Orthologs”, combines bacteria, archaea and eukaryote proteomes and corresponds to the latest version of the orthology benchmark released by the Quest for Orthologs Consortium (Dessimoz *et al.*, 2012).

2.3 Large-scale data visualization

OrthoInspector 2.0 introduces an extended graphical interface and new tools to select, compare and visualize complex orthology relationships from a few sequences up to hundreds of proteomes (figure 1C and 1D). On-going analyses can now be saved via a session system. Sequences can be batch retrieved by creating a phylogenetic profile using presence-absence criteria associated with a selection of clades. Orthology predictions can be visually reevaluated via the introduction of a novel “best-hit density graph” tool. Finally, publication quality Euler diagrams can be generated to represent global orthology relations between several complete proteomes.

3 MAIN ADDITIONS

3.1 Large-scale phylogenetic profiles

Several analyses are now supported by an interactive tree of life in the graphical interface, facilitating in particular the establishment of “phylogenetic profile” queries. Figure 1C shows a screenshot of such a query, where a selection of presence/absence criteria at different levels in the tree, allows the extraction of all Microsporidia sequences with orthologs in Basidiomycota but not in Ascomycota, in a few seconds. The number in parentheses indicates the number of species that will be selected in each tree branch for the analysis.

*To whom correspondence should be addressed.

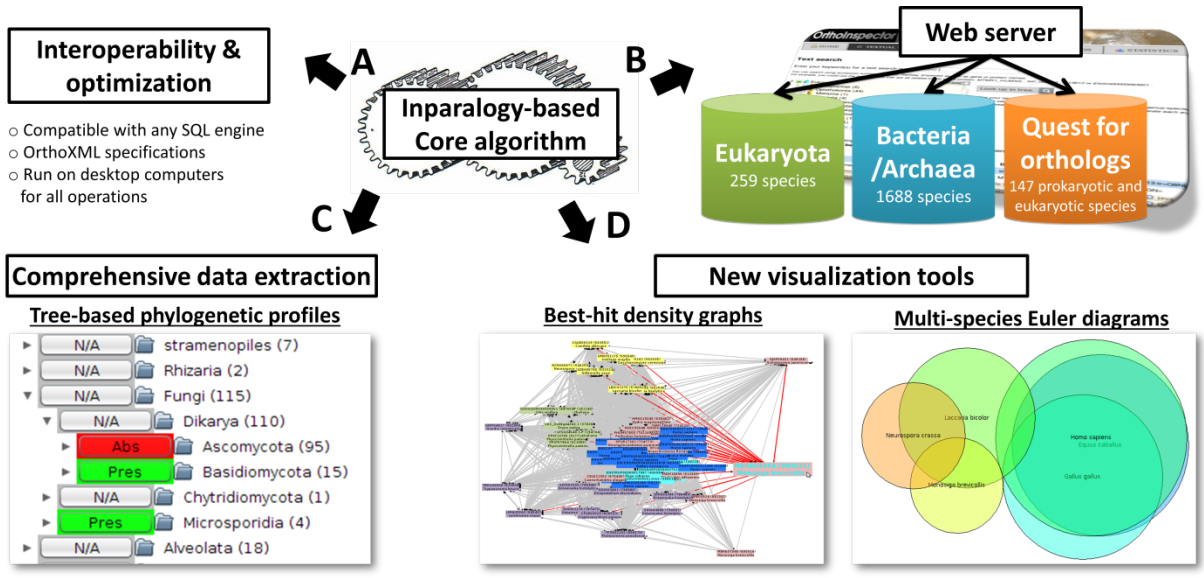

Figure 1: Overview of OrthoInspector main extensions

3.2 Orthology-based Euler diagrams

Diagrams representing sequences that overlap (via the orthology criteria) between 2 to N organisms (figure 1D) can be generated, customized and exported as images. Classical Venn diagrams (3 organisms), but also more complex Euler diagrams (>3 organisms), can be generated. Diagram overlaps are based on the VennEuler library (Wilkinson, 2012), which provides a statistical framework to estimate the best possible circle-based representation.

3.3 Best-hit density graph

The "best-hit density graph" (figure 1D) is designed to analyze the orthologous relationships linking genes in a particular family and to reveal potential sub-families. Through a dynamic graph representation of BLAST best hits linking a protein family, the user can explore conservation patterns within the set by modifying the BLAST score or E-value thresholds on the fly (Suppl. File S2). This tool can be used to adapt the delineation of subfamilies to the evolutionary rate of the family under consideration or to a given phylogenetic scope.

3.4 Web server

The 3 databases generated with OrthoIjnspector 2.0 can be accessed via a web server allowing ortholog retrieval by text or Blastp searches (Camacho *et al.*, 2009) (figure 1B). A list of organisms can be selected with an interactive species tree. Orthology relationships corresponding to the query sequence are compiled in a table format with phylum color codes in order to facilitate the analysis of phylum specific orthology distributions. The presentation of the results is designed to produce a user-friendly and intuitive overview of the evolutionary history revealed by the orthologs (see Suppl. File 3 for a case study). All results can be downloaded in Fasta format or as a list of gene or protein identifiers. Flat files of the complete databases can be downloaded in CSV or OrthoXML formats.

4 CONCLUSION

OrthoInspector is a tool dedicated to the efficient calculation and analysis of orthology data and allows a rapid and intuitive analysis of relationships associated with large clades. The OrthoInspector web server now allows the retrieval of orthology data for thousands of eukaryote and prokaryote proteomes.

ACKNOWLEDGEMENTS

Funding: This work was supported by the ANR [grant ANR-10-INSB-05-01 FRISBI, grant ANR-10-BINF-03-02 BIPBIP] and Institute funds from the CNRS, the Faculté de Médecine de Strasbourg and the Université de Strasbourg.

Conflict of Interest: none declared.

REFERENCES

Altenhoff,A.M. and Dessimoz,C. (2012) Inferring orthology and paralogy. *Methods Mol. Biol.*, **855**, 259–279.

Camacho,C. *et al.* (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.

Dalquen,D.A. *et al.* (2013) The impact of gene duplication, insertion, deletion, lateral gene transfer and sequencing error on orthology inference: a simulation study. *PLoS One*, **8**, e56925.

Dessimoz,C. *et al.* (2012) Toward community standards in the quest for orthologs. *Bioinformatics*, **28**, 900–4.

Fitch,W.M. (1970) Distinguishing homologous from analogous proteins. *Syst. Zool.*, **19**, 99–113.

Kristensen,D.M. *et al.* (2011) Computational methods for Gene Orthology inference. *Brief. Bioinform.*, **12**, 379–391.

Linard,B. *et al.* (2011) OrthoInspector: comprehensive orthology analysis and visual exploration. *BMC Bioinformatics*, **12**, 11.

Schmitt,T. *et al.* (2011) Letter to the editor: SeqXML and OrthoXML: standards for sequence and orthology information. *Brief. Bioinform.*, **12**, 485–8.

Wilkinson,L. (2012) Exact and approximate area-proportional circular Venn and Euler diagrams. *IEEE Trans. Vis. Comput. Graph.*, **18**, 321–31.