Bioinformatics, 34(17), 2018, 3041–3043
doi: 10.1093/bioinformatics/bty225
Advance Access Publication Date: 6 April 2018
Applications Note

OXFORD

Phylogenetics

PhyloProfile: dynamic visualization and exploration of multi-layered phylogenetic profiles

Ngoc-Vinh Tran[1,*], Bastian Greshake Tzovaras[1] and Ingo Ebersberger[1,2,*]

[1]Department for Applied Bioinformatics, Institute of Cell Biology and Neuroscience, Goethe University, 60438 Frankfurt am Main, Germany and [2]Senckenberg Biodiversity and Climate Research Centre (BiK-F), 60325 Frankfurt am Main, Germany

*To whom correspondence should be addressed.
Associate Editor: Russell Schwartz

Received on October 10, 2017; revised on February 28, 2018; editorial decision on April 4, 2018; accepted on April 5, 2018

Abstract

Summary: Phylogenetic profiles form the basis for tracing proteins and their functions across species and through time. Novel genome sequences nowadays often represent species from the remotest corner of the tree of life. Thus, phylogenetic profiling becomes increasingly important for functionally annotating this data and to integrate it into a comprehensive view on organismal evolution. To strengthen the link between the sharing of a gene across species and of the corresponding function, it is meanwhile common to complement phylogenetic profiles with additional information, such as domain architecture similarities between orthologs, or pairwise similarities of other protein features. However, there are few visualization tools that facilitate an intuitive integration of these various information layers. Here, we present PhyloProfile, an R-based tool to visualize, explore and analyze multi-layered phylogenetic profiles.

Availability and implementation: PhyloProfile is available as open source code under the MIT license at https://github.com/BIONF/phyloprofile. An online version for testing PhyloProfile and for small to medium-scale analyses is available at http://applbio.biologie.uni-frankfurt.de/phyloprofile.

Contact: tran@bio.uni-frankfurt.de or ebersberger@bio.uni-frankfurt.de

1 Introduction

Phylogenetic profiles capture the presence–absence pattern of genes across species (Pellegrini *et al.*, 1999). The presence of an ortholog in a given species is often taken as evidence that also the corresponding function is represented (Lee *et al.*, 2007). Moreover, if two genes agree in their phylogenetic profile, it can suggest that they functionally interact (Pellegrini *et al.*, 1999). Phylogenetic profiles are therefore commonly used for tracing functional protein clusters or metabolic networks across species and through time. However, orthology inference is not error-free (Altenhoff *et al.*, 2016), and orthology does not guarantee functional equivalence for two genes (Studer and Robinson-Rechavi, 2009). Therefore, phylogenetic profiles are often integrated with accessory information layers, such as sequence similarity, domain architecture similarity or semantic similarity of Gene Ontology-term descriptions. Various approaches exist to visualize such enriched phylogenetic profiles. For example, public ortholog databases often provide the domain architectures of the identified orthologs (e.g. Altenhoff *et al.*, 2015), DoMosaics (Moore *et al.*, 2014) or the ETE3 tool kit (Huerta-Cepas *et al.*, 2016) facilitate a display of domain architectures at the leafs of a gene tree and recently Aquerium was developed to display domain-based protein occurrences on taxonomically clustered genome trees (Adebali and Zhulin, 2017). However, there is still a shortage of tools that provide a comprehensive set of functions for the display, filtering and analysis of multi-layered phylogenetic profiles comprising hundreds of genes and taxa. PhyloProfile serves to close this methodological gap.

2 Features and capabilities

2.1 Input

PhyloProfile expects as a main input the phylogenetic distribution of orthologs or more generally of homologs. This information can be complemented with domain architecture annotation and data for up to two additional annotation layers. The tool accepts tab delimited text and sequences in FASTA format as input. The stand-alone version additionally supports orthoXML (Schmitt *et al.*, 2011). To ease the generation of custom input, we provide several example datasets and a number of helper scripts, e.g. to extract phylogenetic profiles directly from the OMA database (Altenhoff *et al.*, 2015). The WIKI accompanying PhyloProfile gives a comprehensive guide of how to format input data and additionally informs about performance and scaling of run time and memory usage.

2.2 Interactive visualization and dynamic exploration of phylogenetic profiles

PhyloProfile is implemented with an interactive visualization using the Shiny package for *R* (https://CRAN.R-project.org/package=shiny). Species are automatically linked to the NCBI taxonomy and are ordered in increasing taxonomic distance from a user-specified reference taxon. Alternatively, a custom phylogeny can be uploaded for this purpose. Input taxa can be collapsed at higher order systematic ranks to rapidly change the resolution from the comparative analysis of proteins in individual species, to that across classes, phyla or entire kingdoms.

The phylogenetic profile is represented by a dot matrix (Fig. 1). Cell color, as well as dot size and dot color can accommodate further information about the shared genes. Plotting takes about 10 s for 200 genes and 200 species and scales linearly with size of the data matrix. The protein sequences together with complementary information can be accessed upon a click on the dot.

PhyloProfile is able to represent the entire data matrix or to visualize only a subset of genes and taxa for a detailed inspection, without the need of modifying the input data. Furthermore, the software provides various options to dynamically filter the data. For example, increasing the fraction of species in a systematic group that must harbor an ortholog before the gene is considered present in this group reduces the impact of spurious ortholog identification on evolutionary interpretations. Likewise, filtering genes based on the similarity of their domain architectures—if given as an information layer—can either highlight or blend out orthologs that are suspicious of having changed their function.

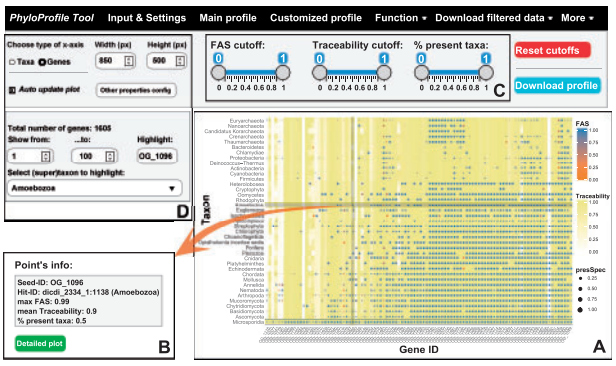

Fig. 1. Main profile plot of PhyloProfile. (**A**) Phylogenetic profile showing the presence-absence patterns of genes across taxa supplemented with two annotation layers [feature architecture similarity (Koestler *et al.*, 2010), traceability (Jain *et al.*, 2018)]. (**B**) Detailed information of a selected cell of the profile. (**C**) Dynamic cut-offs for filtering the data. (**D**) Options for modifying the appearance of the profile

2.3 Analysis functions

PhyloProfile provides several functions for dynamically analyzing phylogenetic profiles.

Profile clustering: The identification of proteins with similar phylogenetic profiles is a crucial step in the identification and characterization of novel functional protein interaction networks (Pellegrini, 2012). PhyloProfile offers the option to cluster genes according to the distance of their phylogenetic profiles.

Gene age estimation: PhyloProfile can estimate the evolutionary age of a gene from the phylogenetic profiles using an Last Common Ancestor (LCA) algorithm (Capra *et al.*, 2013). Specifically, the last common ancestor of the two most distantly related species displaying a given gene serves as the minimal gene age. Age estimates are dynamically updated upon filtering of the data.

Core gene identification: Phylogenomic reconstructions are typically based on a collection of core genes (Daubin *et al.*, 2002), i.e. genes that are shared among all genomes in a taxon collection. PhyloProfile enables users to select a set of taxa and returns their core genes.

Distribution analysis: The interpretation of phylogenetic profiles and the result of downstream analyses can change substantially upon filtering the data. To help users to decide on reasonable filtering thresholds, PhyloProfile provides a function to plot the distributions of the values incurred by the integrated information layers.

2.4 Interoperable output

Filtered data and corresponding protein sequences can be exported for downstream analysis, such as phylogenomic tree reconstruction or metabolic pathway analysis. All graphics generated by PhyloProfile can be downloaded as ready-for-publish PDF files.

Acknowledgement

The authors thank Arpit Jain for valuable discussion.

Funding

This work was supported by the Deutsche Forschungsgemeinschaft [DFG FOR 2251; Project Grant EB 285/2-1].

Conflict of Interest: none declared.

References

Adebali,O. and Zhulin,I.B. (2017) Aquerium: a web application for comparative exploration of domain-based protein occurrences on the taxonomically clustered genome tree. *Proteins*, **85**, 72–77.

Altenhoff,A.M. *et al.* (2015) The OMA orthology database in 2015: function predictions, better plant support, synteny view and other improvements. *Nucleic Acids Res.*, **43**, D240–D249.

Altenhoff,A.M. *et al.* (2016) Standardized benchmarking in the quest for orthologs. *Nat. Methods*, **13**, 425–430.

Capra,J.A. *et al.* (2013) How old is my gene? *Trends Genet.*, **29**, 659–668.

Daubin,V. *et al.* (2002) A phylogenomic approach to bacterial phylogeny: evidence of a core of genes sharing a common history. *Genome Res.*, **12**, 1080–1090.

Huerta-Cepas,J. *et al.* (2016) ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Mol. Biol. Evol.*, **33**, 1635–1638.

Jain,A. *et al.* (2018) The evolutionary traceability of proteins. *bioRxiv*, [Preprint, doi: 10.1101/302109].

Koestler,T. *et al.* (2010) FACT: functional annotation transfer between proteins with similar feature architectures. *BMC Bioinformatics*, **11**, 417.

Lee,D. *et al.* (2007) Predicting protein function from sequence and structure. *Nat. Rev. Mol. Cell. Biol.*, **8**, 995–1005.

Moore,A.D. *et al.* (2014) DoMosaics: software for domain arrangement visualization and domain-centric analysis of proteins. *Bioinformatics*, **30**, 282–283.

Pellegrini,M. (2012) Using phylogenetic profiles to predict functional relationships. *Methods Mol. Biol.*, **804**, 167–177.

Pellegrini,M. *et al.* (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci. U. S. A.*, **96**, 4285–4288.

Schmitt,T. *et al.* (2011) Letter to the editor: seqXML and OrthoXML: standards for sequence and orthology information. *Brief. Bioinformatics*, **12**, 485–488.

Studer,R.A. and Robinson-Rechavi,M. (2009) How confident can we be that orthologs are similar, but paralogs differ? *Trends Genet.*, **25**, 210–216.