BIOINFORMATICS APPLICATIONS NOTE

Vol. 27 no. 6 2011, pages 877–878
doi:10.1093/bioinformatics/btr023

Systems biology

Advance Access publication January 19, 2011

PhyloPro: a web-based tool for the generation and visualization of phylogenetic profiles across Eukarya

Xuejian Xiong[1], Hongyan Song[1], Tuan On[1,2], Lucas Lochovsky[1], Nicholas J. Provart[3,4] and John Parkinson[1,2,4,5,*]

[1]Program in Molecular Structure and Function, Hospital for Sick Children, [2]Department of Molecular Genetics, [3]Department of Cell and Systems Biology, [4]Centre for the Analysis of Genome Evolution and Function and [5]Department of Biochemistry, University of Toronto, Toronto, Canada

Associate Editor: David Posada

ABSTRACT

Summary: With increasing numbers of eukaryotic genome sequences, phylogenetic profiles of eukaryotic genes are becoming increasingly informative. Here, we introduce a new web-tool Phylopro (http://compsysbio.org/phylopro/), which uses the 120 available eukaryotic genome sequences to visualize the evolutionary trajectories of user-defined subsets of model organism genes. Applied to pathways or complexes, PhyloPro allows the user to rapidly identify core conserved elements of biological processes together with those that may represent lineage-specific innovations. PhyloPro thus provides a valuable resource for the evolutionary and comparative studies of biological systems.

Contact: jparkin@sickkids.ca

Received on November 3, 2010; revised on December 21, 2010; accepted on January 6, 2011

In its simplest form, a phylogenetic profile is simply a pattern denoting the presence or absence of a homolog across a set of species. Previous studies have shown that genes sharing similar functions often have similar phylogenetic profiles, a phenomenon that has been exploited in various gene function prediction algorithms (Date and Marcotte, 2005; von Mering *et al*., 2007). Due to a lack of eukaryotic genomes, analyses involving phylogenetic profiles have typically been restricted to bacteria. However, the generation of increasing numbers of eukaryotic genomes is beginning to allow more meaningful comparisons of profiles across the Eukarya. In addition to function prediction, these profiles may be usefully exploited to explore the evolutionary trajectories of sets of genes that operate in related processes (Gabaldon, 2008; On *et al*., 2010; Persaud *et al*., 2009). Such system-based analyses can help identify both conserved genes that represent ancestral core elements of a pathway or complex, as well as those providing taxon-specific innovations (Hulsen *et al*., 2009). Here we introduce PhyloPro, a web-based tool that provides a user-friendly portal for the visualization of pre-computed phylogenetic profiles for sets of genes defined by the user. Through the implementation of a clustering step, genes are grouped on the basis of their phylogenetic profiles. These groupings are then visualized through a java-based interface that facilitates the ready delineation of genes restricted to specific lineages.

*To whom correspondence should be addressed.

Focusing on the six model organisms: *Arabadopsis thaliana* (plant), *Saccharomyces cerevisiae* (yeast), *Caenorhabditis elegans* (worm), *Drosophila melanogaster* (fly), *Mus musculus* (mouse) and *Homo sapiens* (human), the Inparanoid algorithm (Remm *et al*., 2001) was used to perform pairwise homology searches for each model species against 120 eukaryotes for which a complete genome sequence has been generated. These comprise 5 plant species, 30 protozoa, 26 fungi, 2 'basal' metazoa, 7 nematodes, 17 arthropods, 2 chordates, 7 vertebrates and 24 mammals. The use of Inparanoid readily facilitates the identification of so-called in-paralogs representing lineage-specific gene duplication events. Given a list of query genes from a model species (the 'query' species), for each 'target' species, we define one of five possible homology relationship: no detectable ortholog; 1:1—a single query gene has a single ortholog in the target species; 1:Many—a single query gene has two or more orthologs in the target species; Many:1—a query gene together with at least one additional paralog are orthologs of a single gene in the target species; and Many:Many—a query gene together with at least one additional paralog are orthologs of at least two orthologs in the target species genome. The collation of these relationships for each of the 120 target species defines a phylogenetic profile for each query gene which is stored in a local PostgreSQL database (http://www.postgresql.org).

PhyloPro is simple to use, and can be completed in five steps: (i) input of query genes; (ii) selection of target species; (iii) data retrieval and clustering; (iv) visualization; and (v) data download. Users begin by entering a list of sequence identifiers. These can include gene symbols (e.g. for worm 'mek-2'), entrez gene identifiers (IDs) (e.g. '171872'), protein IDs (e.g. 'CE25437'), open reading frame (ORF) names (e.g. 'Y54E10BL.6') or other organism-specific terms (e.g. 'WBGENG00003186'). Sequence identifiers can be entered directly into a text box or uploaded as a file. Next, the user selects the target species from a series of nine expandable taxonomy groups highlighted above. Finally, the user is given the option of changing two parameters associated with the clustering algorithm. After reviewing user input and options, the phylogenetic profiles of each query gene is retrieved from the database and clustered using the tool Cluster 3.0 (Eisen *et al*., 1998). Users are then presented with a visualization of their results which features an interactive heatmap of orthology relationships (Fig. 1). Within the heatmap the order of the 120 eukaryotes is predefined. The heatmap can be downloaded as a Portable Network Graphic (PNG)

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

© The Author 2011. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

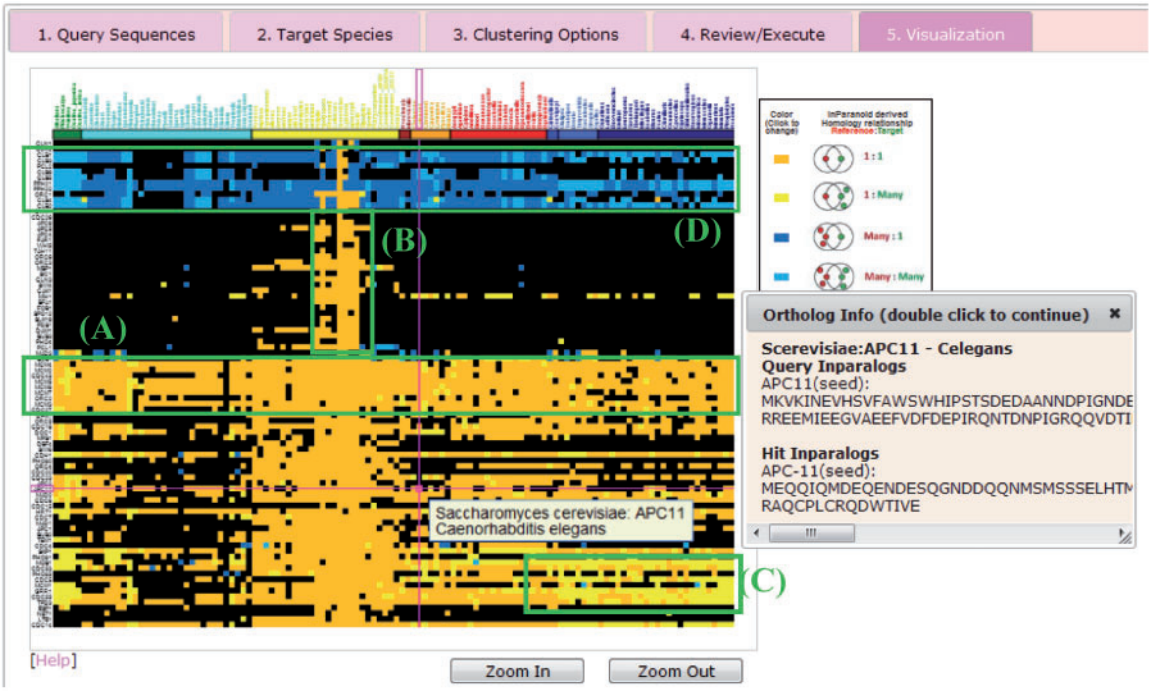

Fig. 1. Screenshot from PhyloPro showing the phylogenetic profiles of 86 genes involved in the yeast cell cycle. Colored tiles indicate the type of orthology relationship between the query gene and the target species: black, no ortholog detected; orange, 1:1 relationship; yellow, 1:M relationship; dark blue, M:1 relationship and light blue indicates M:M relationship. Mousing over tiles indicates the query gene and target species. Clicking colored tiles allows the IDs and sequences of the query and orthologous genes to be shown in a pop-up window. (**A**–**D**) Groups of genes with distinct patterns of conservation—see main text for more details.

or tab-delimited text file. Users may also download the list of all orthologs shown in the heatmap. Figure 1 shows a typical application of PhyloPro, visualizing the phylogenetic profiles of 86 yeast genes associated with the cell cycle. From the heatmap, four groups of genes with specific patterns of orthology relationship are readily delineated: (A) highly conserved across all 120 Eukarya; (B) fungal-specific innovations; (C) duplications in vertebrates; and (D) duplications in yeast. PhyloPro has successfully been applied to explore the evolutionary landscape of the chromatin modification machinery (On *et al.,* 2010), and the conservation of substrates of the Nedd4 family of ubiquitin ligases (Persaud *et al.,* 2009). There are several caveats associated with orthology detection (Ruano-Rubio *et al*., 2009). For example, in the absence of detailed phylogenetic analyses, domain gains, losses and shuffling events can significantly complicate orthology assignments. Hence, future development plans for PhyloPro include incorporating alternative orthology predictions from other graph or tree-based methods such as OrthoMCL (Li *et al*., 2003) and TreeFam (Ruan *et al*., 2008). Furthermore, the list of target species will be expanded as new genomes become available. Finally, we plan to extend the range of query species beyond the six model organisms to permit unbiased sequence-based searches. In summary, PhyloPro provides an intuitive web-based resource for performing rapid evolutionary and comparative studies of biological systems.

ACKNOWLEDGEMENTS

The authors would like to thank Dr James Wasmuth for advice during the creation of PhyloPro. Computing resources were provided by the Center for Computational Biology, Hospital for Sick Children and the SciNet HPC Consortium.

Funding: Canadian Institute for Health Research (CIHR - MOP#82940); New Investigators award from CIHR (to J.P.); Early Researchers Award from Ontario Ministry for Research and Innovation (to J.P.).

Conflict of Interest: none declared.

REFERENCES

Date,S.V. and Marcotte,E.M. (2005) Protein function prediction using the Protein Link EXplorer (PLEX). *Bioinformatics*, **21**, 2558–2559.

Eisen,M.B. *et al.* (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.

Gabaldon,T. (2008) Comparative genomics-based prediction of protein function. *Methods Mol. Biol.*, **439**, 387–401.

Hulsen,T. *et al.* (2009) PhyloPat: an updated version of the phylogenetic pattern database contains gene neighborhood. *Nucleic Acids Res.*, **37**, D731–D737.

Li,L. *et al.* (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.*, **13**, 2178–2189.

On,T. *et al.* (2010) The evolutionary landscape of the chromatin modification machinery reveals lineage specific gains, expansions, and losses. *Proteins*, **78**, 2075–2089.

Persaud,A. *et al.* (2009) Comparison of substrate specificity of the ubiquitin ligases Nedd4 and Nedd4-2 using proteome arrays. *Mol. Syst. Biol.*, **5**, 333.

Remm,M. *et al.* (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.*, **314**, 1041–1052.

Ruan,J. *et al.* (2008) TreeFam: 2008 update. *Nucleic Acids Res.*, **36**, D735–D740.

Ruano-Rubio,V. *et al.* (2009) Comparison of eukaryotic phylogenetic profiling approaches using species tree aware methods. *BMC Bioinformatics*, **10**, 383.

von Mering,C. *et al.* (2007) STRING 7–recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res.*, **35**, D358–D362.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016