BIOINFORMATICS APPLICATIONS NOTE
Vol. 24 no. 23 2008, pages 2778–2779
doi:10.1093/bioinformatics/btn533

Phylogenetics

POPE—a tool to aid high-throughput phylogenetic analysis

Thorhildur Juliusdottir*, Fredrik Pettersson and Richard R. Copley
Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford, OX3 7BN, UK

Received on July 18, 2008; revised and accepted on October 8, 2008
Advance Access publication October 10, 2008
Associate Editor: Martin Bishop

ABSTRACT

Summary: POPE (Phylogeny, Ortholog and Paralog Extractor) provides an integrated platform for automatic ortholog identification. Intermediate steps can be visualized, modified and analyzed in order to assess and improve the underlying quality of orthology and paralogy assignments.

Availability: POPE is available for download from the website: http://www.well.ox.ac.uk/~tota/pope.

Contact: tota@well.ox.ac.uk

1 INTRODUCTION

The correct identification of orthologs and paralogs is central to understanding gene function and studies of genome evolution. Orthologs are defined as genes related by a speciation event, whereas paralogs are related via an inter-genome duplication event. The formally correct way of identifying orthologs is therefore to infer a phylogenetic tree and then parse that tree for species relationships between genes. This approach is used by online databases, such as TreeFam (Li *et al.*, 2006) and RIO/Forester (Zmasek and Eddy, 2002), both of which compare a master species tree to individual gene trees to infer orthologs and paralogs from phylogenies. Tools such as Phylogenie (Frickey and Lupas, 2004) automatically identify orthologs of a starting sequence, via database searching, sequence alignment, alignment processing and tree reconstruction. These intermediate steps, however, are hard to automate reliably in this context although their success can have a major impact on the quality of the resulting phylogeny. Such issues can often be dealt with by manually reviewing the underlying data, and modifying analysis techniques where necessary.

2 DESCRIPTION

POPE (Phylogeny, Ortholog and Paralog Extractor) is a Java and Perl program, developed in order to aid in phylogenetic data analysis using a large number of proteins. POPE includes routines to sort phylogenetic trees based on the presence or absence of orthologs of the target gene, at the same time as graphically displaying alignments, phylogenies and domain locations within the multiple alignments.

The user can either apply POPE to previously generated phylogenies or use POPE to drive phylogeny generation. POPE provides an interface to fully automate traditional phylogenetic analysis using multiple query sequences (belonging to one species). Currently BLAST (Altschul *et al.*, 1990) is used to search for homologs, MUSCLE (Edgar, 2004) to align the homologs and PHYML (Guindon and Gascuel, 2003) is applied for tree reconstruction, although in principle any equivalent tools could be used.

POPE identifies orthologs of the query sequence by parsing the local phylogenetic tree structure for genes or groups of species-specific paralogs related to the query by speciation events. Phylogenies displaying an orthologous relationship between the query species sequence and for example human, worm and fly can be extracted. The results are graphically displayed in tables in POPE, as shown in Figure 1A and B. Two different layouts are used to display the results, called the *Overview and Selection View* and the *Table View*, respectively. The *Overview and Selection View* (Fig. 1A) gives a numerical overview of the results and a schematic view of the presence/absence of any given gene in a particular species. For example, the results displayed in Figure 1A, were obtained when 87 trees (including eight species) were read by POPE in the search for human, worm and fly genes that were orthologous to *Nematostella vectensis* genes. Out of the 87 trees, 29 fulfilled the requested orthology criterion according to the tree pattern search algorithm. The query sequences in these 29 trees, are represented as rows in the table in Figure 1A (one row for each *Nematostella* sequence), along with its potential orthologs according to the tree structure. Colored rectangles indicate presence of a species whereas white rectangles represent its absence, i.e. the first row in the table in Figure 1A shows that the associated *Nematostella* sequence has potential orthologs in seven out of the eight possible species. The rectangles are colored according to the bootstrap value supporting the branch of orthologs and whether the orthologs extracted from the phylogenies are also the top blast hits (Fig. 1A). Phylogenies that are 'in agreement' with the blast results as well as having their group of orthologs supported by a bootstrap value above a given threshold are automatically selected as 'good phylogenies/results'. The user can regulate which trees are automatically selected by altering the bootstrap value supplied by default. By clicking on an entry in the overview table, or by selecting the results tab, the results can be viewed in the more detailed *Table View* (Fig. 1C).

The *Table View* is used to view, further analyze and sort the results. It has three tables associated with it: *trees to sort*, *good trees* and *bad trees* (Fig. 1C). Table entries (rows) can be moved between all three tables by selecting them and clicking on the 'move to' button. Grouping the results into the three tables, enables the user to spend most of his/her time viewing and modifying the data in the *trees to sort* table and gradually move the data into either the 'good' or 'bad' categories. Analyses are accessible through the tables in order

*To whom correspondence should be addressed.

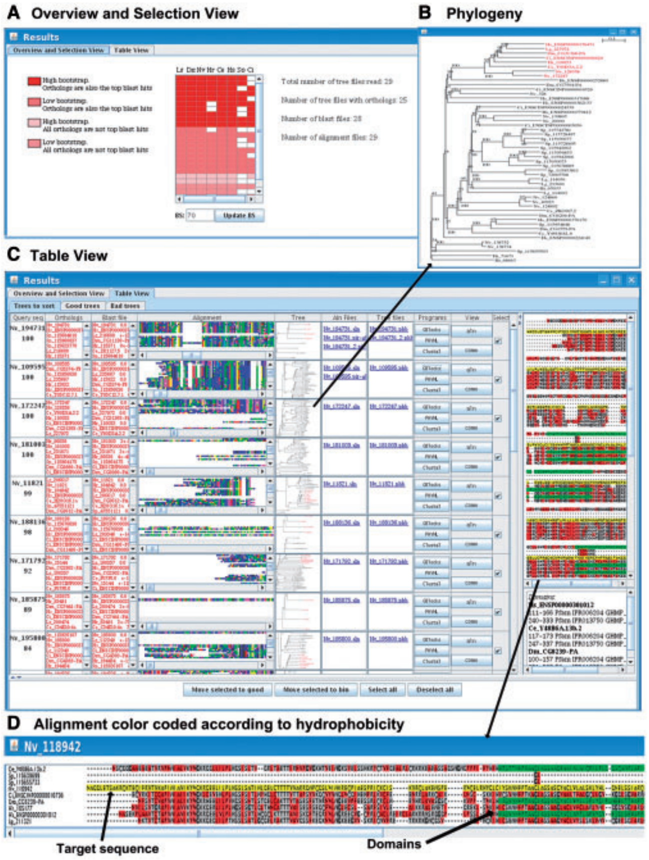
A Overview and Selection View
Results
Overview and Selection View
Table View
High bootstrap. Orthologs are also the top blast hits
Low bootstrap. Orthologs are also the top blast hits
High bootstrap. All orthologs are not top blast hits
Low bootstrap. All orthologs are not top blast hits
Lr Dm Nv Hr Ce Hs Sp Ci
Total number of tree files read: 29
Number of tree files with orthologs: 25
Number of blast files: 28
Number of alignment files: 29
BS: 70
Update BS
B Phylogeny
C Table View
Results
Trees to sort
Good trees
Bad trees
Query seq
Orthologs
Blast file
Alignment
Tree
Aln files
Tree files
Programs
View
Select
Move selected to good
Move selected to bin
Select all
Deselect all
D Alignment color coded according to hydrophobicity
Nv_118942
Target sequence
Domains

Fig. 1. POPE's results viewer. (**A**) The *Overview and Selection view,* where each query sequence and its potential orthologs are represented in the table, color coded based on bootstrap and agreement with the blast results. (**B**) An image of one of the generated phylogenies. The group of orthologs is colored in red. (**C**) The *Table view*, illustrates the query sequence along with its potential orthologs and top blast hits. Graphical representations of all generated alignments and phylogenies are accessible through the interface. (**D**) The alignment panel shows the multiple alignment associated with the selected row, colored according to hydrophobicity. The target sequence is colored in yellow, and extracted domains are distinguished by a green color.

to aid the selection of orthologs. GBlocks (Castresana, 2000) and PHYML can be run to observe whether removing less conserved regions within the alignment will affect the arrangement of potential orthologs within the tree.

POPE keeps track of all the files generated for each query sequence, by listing them in the table. Each generated alignment file can then be viewed individually or all alignment files (along with the original files) can be viewed simultaneously for comparison (Fig. 1B). Phylogenies are displayed through POPE using Njplot (Perriere and Gouy, 1996). Alternatively, an image of the original phylogeny where the orthologous group has been colored in red can be viewed. The potential orthologs for each sequence are also listed in the table, where orthologs that are also the top blast hits are colored in red. The list of orthologs can be modified and saved by the user. Trees with less strong support might be clarified by best BLAST hits, so we highlight the top blast hits that are also among the group of orthologs. The entire BLAST file can also be examined through POPE's tables.

Information on domains retrieved from Ensembl (Birney *et al.*, 2006), can also be extracted and displayed when available. This information is displayed on the right-hand side of the Table View in the *Domain* panel. A *Tree* panel and an *Alignment* panel are also displayed to the right in the Tree view. The tree panel shows the phylogeny of the selected table entry. The *Alignment* panel, shows the alignment belonging to the selected entry, color coded according to the hydrophobicity of the amino acids included in the alignment. When domains have been extracted using the domain panel, they are displayed in green within their respective sequence in the group panel (Fig. 1D). This shows the user whether the available domains have been correctly aligned, and will assist in visually evaluating the quality of the multiple alignment.

SUMMARY

In summary, POPE is a tool that automates phylogenetic analysis and offers an interactive environment to view, sort and further analyse the derived alignments and phylogenies. It provides the user with a graphical overview of the data, and makes it easy to apply different methods to the data and compare the outcome to the original results. By using POPE for large-scale phylogenetic analysis, a time-consuming and often tedious procedure due to the multiple steps involved, can be performed in an organized and efficient manner.

SYSTEM REQUIREMENTS

POPE requires Java Runtime Environment version 1.5 or higher and BioPerl.

Funding: Wellcome Trust (to R.R.C. and F.P.); Marie Curie FP6 RTN ZOONET (to T.J.).

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al*. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Birney,E. *et al.* (2006) Ensembl 2006. *Nucleic Acids Res.*, **34**, D556–D561.

Castresana,J. (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.*, **17**, 540–552.

Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.

Frickey,T. and Lupas,A.N. (2004) PhyloGenie: automated phylome generation and analysis. *Nucleic Acids Res.*, **32**, 5231–5238.

Guindon,S. and Gascuel,O. (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.*, **52**, 696–704.

Li,H. *et al.* (2006) TreeFam: a curated database of phylogenetic trees of animal gene families. *Nucleic Acids Res.*, **34**, D572–D580.

Perriere,G. and Gouy,M. (1996) WWW-query: an on-line retrieval system for biological sequence banks. *Biochimie*, **78**, 364–369.

Zmasek,C.M. and Eddy,S.R. (2002) RIO: analyzing proteomes by automated phylogenomics using resampled inference of orthologs. *BMC Bioinformatics*, **3**, 14.