BIOINFORMATICS APPLICATIONS NOTE Vol. 18 no. 9 2002
Pages 1272–1273

# OrthoGUI: graphical presentation of Orthostrapper results

*Volker Hollich, Christian E. V. Storm and Erik L. L. Sonnhammer**

*Center for Genomics and Bioinformatics, Karolinska Institutet, SE-171 77 Stockholm, Sweden*

Received on February 22, 2002; accepted on April 25, 2002

**ABSTRACT**

**Summary:** Orthostrapper is a program that calculates orthology support values for pairs of sequences in a multiple alignment (Storm and Sonnhammer, *Bioinformatics*, **18**, 92–99, 2002). Here we present OrthoGUI, a web interface and display tool for Orthostrapper analysis. OrthoGUI visualizes the Orthostrapper output in both tabular and tree representations, and can also apply a clustering algorithm to identify groups of multiple orthologs, which are indicated by colour coding.

**Availability:** http://www.cgb.ki.se/OrthoGUI

**Contact:** erik.sonnhammer@cgb.ki.se

When predicting function and biological role for newly sequenced proteins, analysis of orthologous relationships is a good starting point, as orthologous proteins in different organisms are likely to share same function. Orthologs are often found by comparing sequence and species trees (Page, 1998; Yuan *et al.*, 1998). A potential drawback of this method is that the topology of a reconstructed phylogenetic tree is not always reliable, which can lead to incorrect ortholog assignments.

In order to assess the reliability of ortholog assignments, an algorithm was developed to extract pairwise orthologs from a tree using the bootstrap method (Efron, 1979) to calculate ortholog support values. This method was implemented in a Java program called Orthostrapper (Storm and Sonnhammer, 2002). We here present OrthoGUI, a graphical interface to Orthostrapper that combines easy access through the web combined with a graphical presentation of Orthostrapper results.

Orthostrapper analyses bootstrap trees to calculate the frequency of pairwise orthology assignments. These results can be interpreted as support values for the orthology of any two sequences in the multiple alignment. Orthostrapper was designed to work on two species or lineages. The sequences to be examined either belong to one of the two lineages, the outgroup, or the 'blank' group. The outgroup sequences should be distantly related to the sequences in the two lineages under examination. Sequences belonging to the 'blank' group are not considered during the analysis. Orthostrapper uses the program Belvu (Sonnhammer, unpublished) to calculate the bootstrap trees from the original multiple alignment. By default, Belvu uses standard multiple alignment bootstrapping (sampling with replacement of all columns) and calculates neighbour joining trees from uncorrected difference distances. In principle, any program that reports bootstrap trees could be used.

The bootstrap trees are then inspected for orthologs between the two lineages. All sequence pairs between lineage 1 and 2 are considered. Orthology between a pair is assigned if an internal node exists that joins two pure subtrees from the two lineages that include the sequences in question. If such a node does not exist, either another sequence is orthologous to one of the sequences, or an outgroup sequence is closer to one of them. This is repeated for all bootstrap trees, and the fraction of trees that produced an ortholog assignment gives the ortholog bootstrap value.

On the OrthoGUI homepage, multiple alignments are expected as input in FASTA format. To specify the lineage of each sequence, the format is enhanced by appending '&1' to the sequence names in the first lineage, '&2' in the second lineage, or '&O' in the outgroup. Sequences without an extension will be assigned to the 'blank' group. OrthoGUI was developed as a Java 1.1 applet to support a wide range of systems. As the applet starts, Orthostrapper and Belvu are queried to calculate the orthology matrix and a phylogenetic tree for the submitted sequences.

The upper OrthoGUI frame (see Figure 1) shows the matrix calculated by Orthostrapper. The matrix contains confidence values for orthology of the respective sequences. The lower frame presents the phylogenetic tree of the original alignment according to Belvu. Clicking on the tree opens a context menu which offers zoom in/out. The frames for the table and the tree can also be split into separate windows. As bootstrapping is a heuristic approach, a second query with the same data may deliver slightly different results. The number of bootstrap samples can also be chosen; by default OrthoGUI uses 100 bootstraps.

*To whom correspondence should be addressed.

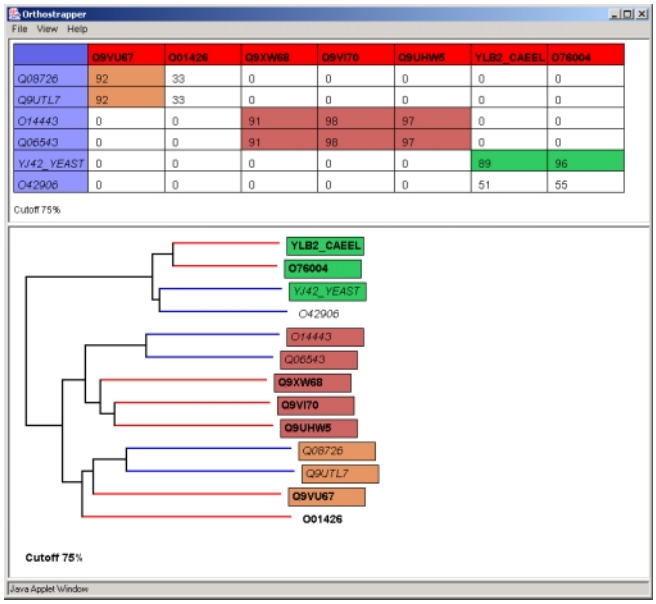


**Fig. 1.** Example of an OrthoGUI view, showing selected sequences from the ATP-bind family in Pfam (PF03029). Sequences from the fungal lineage (in rows, italics font) are 'Orthostrapped' against sequences from the metazoan lineage (in columns, bold font). The ortholog bootstrap values for each sequence pair between the two lineages are shown as percentages in the table. OrthoGUI identifies three groups of orthologs with an average ortholog bootstrap value above the used cutoff 75%. These groups are marked up by group-specific background colors in the table cells and tree leaves that correspond to the group members. The sequences are shown as SWISSPROT 40 or TREMBL 18 identifiers. The species are as follows. O76004, Q9UHW5: *H. sapiens*; Q9VI70, Q9VU67: *D. melanogaster*; YLB2_CAEEL, O01426, Q9XW68: *C. elegans*; YJ42_YEAST, Q06543, Q08726: *S. cerevisiae*; O42906, O14443, Q9UTL7: *S. pombe*.

## CLUSTERING ORTHOLOGS

OrthoGUI implements a method to find clusters of multiple orthologs. In many cases, particularly when comparing distant species, several genes in one species will be orthologous to one or more genes in the other species. When examining lineages, it is even more likely to find multiple orthologs.

Based on a selected clustering cutoff value provided in the options, OrthoGUI can cluster sequences in the pairwise Orthostrapper results. The clustering is based on the 'average linkage' method. The algorithm picks the sequence pair with the highest ortholog bootstrap value (above the cutoff) as the seed group. Then the average ortholog bootstrap value (score) to the seed group is calculated for all remaining sequences. Sequences with averages higher than the cutoff are added to the group starting with the highest. The scores for remaining sequences are recalculated. Hence, sequences are subsequently added to the group until there are only sequences left with scores below the cutoff. The group resulting from this procedure is saved. If in the remaining, unassigned sequences a pair is found with a score above or equal to the cutoff, the clustering process is run again with all unclustered sequences and this pair as a seed. This process is repeated until all unclustered sequences have scores below the cutoff.

Matrix cells and sequence names in the tree are then coloured according to the found clusters, using distinct colours for each group. The clustering algorithm may group a sequence into more than one cluster. In this case, a warning is displayed in both the matrix and the tree frame, and the sequence is coloured with both clusters' colours. The found ortholog groups can be exported with the menu option 'Display groups'. As Java applets reside in a 'sandbox' and cannot make disk access, OrthoGUI simply opens a new page in the browser with this information, which can be saved by the user.

Because Orthostrapper samples pairwise sequence relationships, there is no guarantee that all sequences in a subtree are found to be co-orthologs at a given cutoff. Therefore the output of Orthostrapper can not always be represented as a consistent speciation/duplication tree, e.g. ATV (Zmasek and Eddy, 2001), that is based on the original tree. Orthostrapper and OrthoGUI instead give a representation of the pairwise relationships found in the bootstrap trees.

## REFERENCES

Efron,B. (1979) Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, **7**, 1–26.

Page,R.D. (1998) GeneTree: comparing gene and species phylogenies using reconciled trees. *Bioinformatics*, **14**, 819–820.

Storm,C.E.V. and Sonnhammer,E.L.L. (2002) Automated ortholog inference from phylogenetic trees and calculation of orthology reliability. *Bioinformatics*, **18**, 92–99.

Yuan,Y.P., Eulenstein,O., Vingron,M. and Bork,P. (1998) Towards detection of orthologues in sequence databases. *Bioinformatics*, **14**, 285–289.

Zmasek,C.M. and Eddy,S.R. (2001) A simple algorithm to infer gene duplication and speciation events on a gene tree. *Bioinformatics*, **17**, 821–828.