**Roundup 2.0: Enabling Comparative Genomics for over 1800 Genomes**

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Published Version</td>
<td>doi:10.1093/bioinformatics/bts006</td>
</tr>
<tr>
<td>Citable link</td>
<td><a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:8519645">http://nrs.harvard.edu/urn-3:HUL.InstRepos:8519645</a></td>
</tr>
<tr>
<td>Terms of Use</td>
<td>This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at <a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA">http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA</a></td>
</tr>
</tbody>
</table>
Roundup is an online database of gene orthologs for over 1800 genomes, including 226 Eukaryota, 1447 Bacteria, 113 Archaea and 21 Viruses. Orthologs are inferred using the Reciprocal Smallest Distance algorithm. Users may query Roundup for single-linkage clusters of orthologous genes based on any group of genomes. Annotated query results may be viewed in a variety of ways including as clusters of orthologs and as phylogenetic profiles. Genomic results may be downloaded in formats suitable for functional as well as phylogenetic analysis, including the recent OrthoXML standard. In addition, gene IDs can be retrieved using FASTA sequence search.

All source code and orthologs are freely available.

Availability: http://roundup.hms.harvard.edu

Contact: dpwall@hms.harvard.edu; todd_deluca@hms.harvard.edu

Received on August 27, 2011; revised on December 23, 2011; accepted on January 3, 2012

1 INTRODUCTION

Orthologs are genes from different organisms that descend from a single ancestral gene in the most recent common ancestor (Fitch, 1970). In comparative genomics, they are used to infer the function of novel genes from the function of well-studied ones, to construct phylogenies and explore the evolution of genes and species, and to study sequence conservation and change. They are also valuable in analyzing gene networks, studying gene gain and loss, and finding genes in model organisms that correspond to human disease genes (Altenhoff and Dessimoz, 2009). Also, we replaced ClustalW (Thompson et al., 1994) with NCBI BLAST (Altschul et al, 1990). Also, we replaced ChalstW (Thompson et al., 1994) with Kalign (Lassmann and Sonnhammer, 2005). Kalign is faster than ChalstW and produces better alignments for more distantly related sequences. This change resulted in 9% closer maximum likelihood distances between orthologs computed using PAML 4.0 (Yang, 2007), and 0.3% more orthologs on average. Since the Roundup database stores orthology for 12 combinations of divergence and E-value thresholds, RSD was modified to compute orthologs for any number of parameter combinations as quickly as for one parameter combination. This change should be of interest to researchers investigating the effect of different parameter settings and degree of global sequence similarity on ortholog inference. With the addition of other caching and file I/O changes, RSD is over six times faster than the previous version in our performance tests.

For Roundup 2.0, we changed RSD to improve its speed, stability and ortholog inference. We replaced WU-BLAST (W.Gish, personal communication) with NCBI BLAST (Altschul et al., 1990). Also, we replaced ChalstW (Thompson et al., 1994) with Kalign (Lassmann and Sonnhammer, 2005). Kalign is faster than ChalstW and produces better alignments for more distantly related sequences. This change resulted in 9% closer maximum likelihood distances between orthologs computed using PAML 4.0 (Yang, 2007), and 0.3% more orthologs on average. Since the Roundup database stores orthology for 12 combinations of divergence and E-value thresholds, RSD was modified to compute orthologs for any number of parameter combinations as quickly as for one parameter combination. This change should be of interest to researchers investigating the effect of different parameter settings and degree of global sequence similarity on ortholog inference. With the addition of other caching and file I/O changes, RSD is over six times faster than the previous version in our performance tests.

In addition to housing the orthologs inferred by RSD, Roundup builds clusters of orthologous genes, i.e. orthogroups, using deterministic single-linkage clustering. It partitions a graph into connected subgraphs by creating a cluster for every gene and then...
merging two clusters if a gene in one of the clusters is orthologous to a gene in the other one. The result is that every gene in a group is orthologous to at least one other gene in the group and to no genes in any other groups. In contrast to other orthology databases (Chen et al., 2006; Schneider et al., 2007; Tatusov et al., 2003), Roundup orthologous groups are built on the fly using genomes selected by the user. This allows users to include exactly their genomes of interest and to explore the effects of including different genomes on the grouping of orthologs.

3 GENOMES AND ORTHOLOGS

The 1807 genomes in Roundup 2.0 are from UniProtKB (Magrane et al., 2011), including 226 Eukaryota, 1447 Bacteria, 113 Archaea, 21 Viruses and Viroids. The approximately 63 CPU core-years to compute the orthologs took several weeks on our research computing cluster. Roundup used a fault-tolerant computational pipeline to compute orthologs for all 1 631 721 pairs of genomes across 12 parameter combinations selected to allow researchers access to results for a broad range of divergence and E-value threshold settings. As a result, there are over 11 billion orthologs available in Roundup. The genomes and orthologs are updated 2–4 times per year.

4 WEB INTERFACE

The Roundup website provides two ways to search for orthologs. First, the "Browse" query is a genome-centric search that retrieves all orthologs between one genome and a set of other genomes. Results can be filtered by genome name or gene identifier. To aid users in finding gene identifiers, a FASTA sequence may be used to retrieve a gene id. The second query, "Retrieve", returns all orthologs for all pairs of genomes in a set of genomes the user specifies. Query results are then clustered into groups of orthologous genes as described above. All genes in the groups are linked to UniProt and annotated with available gene names and GO Process terms provided by UniProtKB and Gene Ontology (Ashburner et al., 2000). FASTA sequences for genes in orthologous groups are also provided for further analysis.

In addition to the standard view of search results, there are summaries by GO Terms and by Gene Clusters. The orthologous groups may be downloaded in several formats: NEXUS, PHYLIP, OrthoXML, Phylogenetic Profile and Text. OrthoXML (Schmitt et al., 2011) is provided to support interoperability with other Orthology databases and the quest for orthologs (Gabaldon et al., 2009; Kuzniar et al., 2008). Query results are cached for up to 30 days and may be retrieved by using the initial URL.

To support research and replication, we make available for download from the website: FASTA sequences for genomes; orthologs in OrthoXML and text formats; and code for RSD and Roundup. Orthologs are also available through an HTTP-based API.