BIOINFORMATICS APPLICATIONS NOTE

Vol. 28 no. 5 2012, pages 715–716
doi:10.1093/bioinformatics/bts006

Genome analysis

Advance Access publication January 13, 2012

Roundup 2.0: enabling comparative genomics for over 1800 genomes

Todd F. DeLuca[1,*], Jike Cui[1], Jae-Yoon Jung[1], Kristian Che St. Gabriel[2] and Dennis P. Wall[1,2,*]

[1]The Center for Biomedical Informatics, Harvard Medical School and [2]Department of Pathology, Beth Israel Deaconess Medical Center, Boston MA 02115, USA

Associate Editor: Alfonso Valencia

ABSTRACT

Summary: 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; Gabaldon *et al*., 2009; Kristensen *et al*., 2011).

Advances in high-throughput genomic sequencing have made it possible to produce many datasets in a relatively short time period. For example, from 2006 to 2011, the number of complete proteomes listed in UniProtKB (Magrane and Consortium, 2011), a repository of annotated protein sequences, has increased from around 300 to over 2500. To overcome the engineering challenges of computing and publishing orthologs for such a large number of genomes, we redesigned the comparative genomics tool, Roundup (DeLuca *et al*., 2006), to scale with the rate of genome sequencing and to enable increasingly more sophisticated comparative genomics analyses. Roundup 2.0 contains orthology data for over 1800 genomes, providing one of the largest diversities among similar orthology databases (Chen *et al*., 2006; Datta *et al*., 2009; Huerta-Cepas *et al*., 2011; Kristensen *et al*., 2011; Li *et al*., 2006; Linard *et al*., 2011; Ostlund *et al*., 2010; Rouard *et al*., 2011; Schneider *et al*., 2007; Tatusov *et al*., 2003). Roundup compares well to other major databases, with recent studies showing similar ortholog composition for model organisms (Altenhoff and Dessimoz, 2009; Chen *et al*., 2007). The data in Roundup include clusters of orthologs for a wide range of sequence conservation, allowing searches for distant orthologs, and also phylogenetic profiles that enable functional investigation, phylogenetic analysis and prediction of network organization (Cui *et al*., 2011).

2 ALGORITHMS

We used the reciprocal smallest distance (RSD) (Wall *et al*., 2003) algorithm to infer orthologs. RSD improves the sensitivity of reciprocal best blast hits by considering global alignment and maximum likelihood evolutionary distance between sequences. As a pairwise orthology algorithm, RSD scales quadratically with the number of genomes in Roundup. Altenhoff *et al*. assessed 10 ortholog inference projects and methods, confirming the reliable performance of RSD over a wide array of genomes from the tree of life (Altenhoff and Dessimoz, 2009).

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 ClustalW (Thompson *et al*., 1994) with Kalign (Lassmann and Sonnhammer, 2005). Kalign is faster than ClustalW 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 orthologs 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. ortholog groups, using deterministic single-linkage clustering. It partitions a graph into connected subgraphs by creating a cluster for every gene and then

*To whom correspondence should be addressed.

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 and Consortium, 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 gene 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.

Roundup 2.0 is an important step forward towards keeping pace with the rate of genome sequencing. The features and flexibility of Roundup 2.0, coupled with the wide coverage of genomes, enables increasingly large-scale comparative genomics analyses that address key questions in phylogeny, genome evolution and systems biology.

ACKNOWLEDGEMENTS

Computations were run on the Orchestra cluster supported by the Harvard Medical School Research Information Technology Group.

Funding: National Science Foundation (0543480 and 0640809, to D.P.W); the National Institutes of Health (LM009261, to D.P.W).

Conflict of Interest: none declared.

REFERENCES

Altenhoff,A.M. and Dessimoz,C. (2009) Phylogenetic and functional assessment of orthologs inference projects and methods. *PLoS Comput. Biol.*, **5**, e1000262.

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

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

Chen,F. *et al.* (2006) OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res.*, **34**, D363–D368.

Chen,F. *et al.* (2007) Assessing performance of orthology detection strategies applied to eukaryotic genomes. *PLoS One*, **2**, e383.

Cui,J. *et al.* (2011) Phylogenetically informed logic relationships improve detection of biological network organization. *BMC Bioinformatics*, **12**, 476.

Datta,R.S. *et al.* (2009) Berkeley PHOG: PhyloFacts orthology group prediction web server, *Nucleic Acids Res.*, **37**, W84–W89.

DeLuca,T.F. *et al.* (2006) Roundup: a multi-genome repository of orthologs and evolutionary distances. *Bioinformatics*, **22**, 2044–2046.

Fitch,W.M. (1970) Distinguishing homologous from analogous proteins. *Syst. Zool.*, **19**, 99–113.

Gabaldon,T. *et al.* (2009) Joining forces in the quest for orthologs. *Genome Biol.*, **10**, 403.

Huerta-Cepas,J. *et al.* (2011) PhylomeDB v3.0: an expanding repository of genome-wide collections of trees, alignments and phylogeny-based orthology and paralogy predictions. *Nucleic Acids Res.*, **39**, D556–D560.

Kristensen,D.M. *et al.* (2011) Computational methods for Gene Orthology inference. *Brief Bioinform.*, **12**, 379–391.

Kuzniar,A. *et al.* (2008) The quest for orthologs: finding the corresponding gene across genomes. *Trends Genet.*, **24**, 539–551.

Lassmann,T. and Sonnhammer,E.L. (2005) Kalign—an accurate and fast multiple sequence alignment algorithm. *BMC Bioinformatics*, **6**, 298.

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

Linard,B. *et al.* (2011) OrthoInspector: comprehensive orthology analysis and visual exploration. *BMC Bioinformatics*, **12**, 11.

Magrane,M. and Consortium,U. (2011) UniProt Knowledgebase: a hub of integrated protein data. *Database*, **2011**, bar009.

Ostlund,G. *et al.* (2010) InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Res.*, **38**, D196–D203.

Rouard,M. *et al.* (2011) GreenPhylDB v2.0: comparative and functional genomics in plants. *Nucleic Acids Res.*, **39**, D1095–D1102.

Schmitt,T. *et al.* (2011) SeqXML and OrthoXML: standards for sequence and orthology information. *Brief Bioinform.*, **12**, 485–488.

Schneider,A. *et al.* (2007) OMA Browser—exploring orthologous relations across 352 complete genomes. *Bioinformatics*, **23**, 2180–2182.

Tatusov,R.L. *et al.* (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**, 41.

Thompson,J.D. *et al.* (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.

Wall,D.P. *et al.* (2003) Detecting putative orthologs. *Bioinformatics*, **19**, 1710–1711.

Yang,Z. (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.*, **24**, 1586–1591.