BIOINFORMATICS APPLICATIONS NOTE Vol. 22 no. 16 2006, pages 2044–2046
doi:10.1093/bioinformatics/btl286

Genome analysis

Roundup: a multi-genome repository of orthologs and evolutionary distances

Todd F. DeLuca, I-Hsien Wu, Jian Pu, Thomas Monaghan, Leonid Peshkin, Saurav Singh and Dennis P. Wall*

The Center for Biomedical Informatics & Department of Systems Biology, Harvard Medical School, Boston, MA 02115, USA

Received on March 1, 2006; accepted on April 30, 2006
Advance Access publication June 15, 2006
Associate Editor: Christos Ouzounis

ABSTRACT

SUMMARY: We have created a tool for ortholog and phylogenetic profile retrieval called Roundup. Roundup is backed by a massive repository of orthologs and associated evolutionary distances that was built using the reciprocal smallest distance algorithm, an approach that has been shown to improve upon alternative approaches of ortholog detection, such as reciprocal blast. Presently, the Roundup repository contains all possible pair-wise comparisons for over 250 genomes, including 32 Eukaryotes, more than doubling the coverage of any similar resource. The orthologs are accessible through an intuitive web interface that allows searches by genome or gene identifier, presenting results as phylogenetic profiles together with gene and molecular function annotations. Results may be downloaded as phylogenetic matrices for subsequent analysis, including the construction of whole-genome phylogenies based on gene-content data.

Availability: http://rodeo.med.harvard.edu/tools/roundup

Contact: dpwall@hms.harvard.edu

1 INTRODUCTION

The number of fully sequenced genomes is growing at an unprecedented rate. Presently there are 364 completed genomes and 2025 in various stages of construction, including 607 Eukaryotes (Liolios *et al*., 2006). The sheer number of newly sequenced genomes brings exciting opportunities and challenges to the field of comparative genomics. Paramount to this field is ensuring that comparative genomics tools keep pace with the rate of genome sequencing.

The ability to detect orthologs is a minimum starting requirement for any comparative genomics study. Thus, it is critical that ortholog detection tools that are accurate, fast and comprehensive be readily available. Orthologs have already played important roles in numerous biological research questions, including studies of variables influencing the rate of protein evolution (Wall *et al*., 2005), functional prediction (Pellegrini *et al*., 1999), studies of proteins implicated in cancer (Lopez-Bigas and Ouzounis, 2004), and in gene-content based phylogeny reconstruction (Tekaia and Yeramian, 2005). The number of roles will only increase as more genomes become available. In order to increase the scale of comparative genomics studies to match the number of genomes, our ability to study orthologs must also grow in scale.

2 ALGORITHM

As an answer to this challenge, we have created Roundup, a publicly accessible system for investigating orthologs among every fully sequenced genome presently available from public sources. Roundup is at once a repository, holding precompiled results of orthologs, and an algorithm, determining new orthologs when prompted. The algorithm used to detect orthologs is the reciprocal smallest distance algorithm (RSD) (Wall *et al*., 2003). RSD has been shown to improve upon approaches that are based on reciprocal best blast hits (Remm *et al*., 2001; Tatusov *et al*., 2000) because it uses global rather than local sequence alignments and evolutionary estimates of distance between sequences rather than blast probability scores, an approach that can often be misleading when trying to determine the functional equivalence of two proteins (Koski and Golding, 2001). The evolutionary distances generated are maximum likelihood estimates of the number of amino acid substitutions separating any two protein sequences given an empirical amino acid substitution matrix. These values are stored together with the orthologs so that the extent of divergence among sets of orthologs may be ascertained.

At the time of writing, we have used RSD to precompute orthologs for 250 fully sequenced genomes, including 32 Eukaryotes, more than doubling the coverage of any similar resource (e.g. Chen *et al*., 2006). And, because finding the 'true' set of orthologs between two lineages depends on many parameters, including the date of divergence between the lineages, rates of gene duplication and the intensity of selection, we have run RSD on all pairs of the 250 genomes using variable settings of two parameters, blast *E*-value and percent sequence divergence. Specifically, we used four increasingly stringent blast *E*-value thresholds, 1e−5, 1e−10,1e−15 and 1e−20, and three increasingly stringent divergence thresholds, 0.8, 0.5 and 0.2. Therefore, for every pair of genomes, 12 possibly different sets of orthologs exist in the Roundup results archive. We chose these settings to encompass enough of the sequence divergence space to provide sufficient

*To whom correspondence should be addressed.

exploratory power to the user without risking the inclusion of paralogs. In most of the cases that we have examined, results using more conservative parameter settings are a subset of the result set from less conservative settings.

We have built Roundup to be a community evolvable resource. A user may request that new genomes be added if they are missing from our selection. Also, if newly added genomes have not yet been compared, a user may initiate that comparison, receiving an email when it is complete. Furthermore, Roundup runs procedures to periodically download new and updated genomes. For these, new RSD processes are automatically initiated. Thus, Roundup is constantly growing in size to keep pace with the rates of genome updates and sequencing.

3 INTERFACE

The web interface currently has three main points of access to the ortholog repository. The 'Retrieve Phylogenetic Profiles' (RPP) query allows the user to build phylogenetic profiles for any set of genomes. Phylogenetic profiles (sensu Pellegrini *et al*., 1999) are subgraphs in which the nodes are proteins from different organisms and the edges are undirected orthologous relationships established by the RSD algorithm. At present, two types of queries are allowed. The first finds any orthologs that exist among the genomes selected, and the second reports only those orthologs that are found in all genomes, enforcing that the orthologous relationships of the proteins be transitively closed among the genomes under study. The first is less stringent and will report the presence of a gene in a genome even if an ortholog for the gene can be found in only one of the other genomes. The second is of value if a user is particularly interested in genes that have remained highly conserved since the genomes last shared a common ancestor. These queries may be made more or less stringent by adjusting the *E*-value and divergence parameter settings and may be filtered to report only those orthologs that conform to a particular range of distance values. All queries will return a table of orthologs listing unique accession numbers (e.g. Ensembl and GenBank accession numbers), and if selected, a listing of gene names and GO molecular function terms.

'Browse' is the second access point into Roundup. Here, a user is allowed to search for orthologs to a single protein or set of proteins from one primary genome to any number of secondary genomes in the repository. Because it is a directed query and requires that an orthologous relationship exist between the primary and at least one secondary genome for an ortholog to be reported, it is less inclusive than the RPP query. However it can be a powerful tool if the user is interested only in finding direct orthologs to a model organism for a subset of genes belonging to a particular pathway. Also, given the reduced search space, it requires significantly less time to compute results than the RPP query. Again, searches may be constrained by altering the values of the blast and divergence parameters, and may be filtered on values of evolutionary distance. Finally, the 'Download Raw Data' view allows users to download the complete set of data generated by RSD for each pair of genomes and any of the 12 combinations of parameters.

4 USE AND APPLICATIONS

Any result from either the Browse or RPP queries may be downloaded as an array of phylogenetic profiles [sensu (Makarova *et al*., 2003; Pellegrini *et al*., 1999; Tekaia and Yeramian, 2005)]. The array contains, for every gene, a binary vector representing the presence or absence of that gene in the genomes being investigated. A gene is present if an orthologous relationship was found between any two genomes, in the case of the RPP query, or between the query genome and one of the subject genomes in the case of the Browse query. Such an array can be a powerful, whole-genomic means of investigating biological events, such as the propensity for gene loss (Krylov *et al*., 2003) and rates of horizontal transfer (Wolf *et al*., 2001), as well as for understanding the composition and relationships of protein families. The array may also be transformed into a matrix of binary characters for reconstructing whole genome phylogenies using the pattern of gene composition. Roundup allows the user to download phylogenetic matrices in either Phylip or Nexus format, conferring flexibility to choose among alternative phylogenetic analysis packages, optimality criteria and character weighting schemes. Previous attempts to build gene-content based phylogenies have provided new support for phylogenetic branch points in the tree of life that otherwise have been difficult to resolve (Tekaia and Yeramian, 2005). In a pilot study using unweighted maximum parsimony, we learned that Roundup correctly resolves well supported phylogenetic relationships among Eukaryotes, a testament to the internal consistency of the ortholog data. It remains to be seen what hypotheses about the phylogenetic relationships of major lineages of life Roundup will generate or resolve, but in general we expect the tool to have an important impact on this and related areas of study.

In summary, Roundup has applications to numerous fields of biology both as a lookup tool for quick ortholog retrieval and as an interrogation engine to discover novel patterns either among genes or genomes. It may be accessed at http://rodeo.med.harvard.edu/tools/roundup, where a revised version of the RSD algorithm is also available for download.

ACKNOWLEDGEMENTS

The authors would like to thank the numerous beta testers who helped to refine the tool. The manuscript and tool were greatly improved by the comments of three anonymous reviewers. Funding to pay the Open Access publication charges was provided by the Center for Biomedical Informatics, Harvard Medical School.

Conflict of Interest: none declared.

REFERENCES

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

Koski,L.B. and Golding,G.B. (2001) The closest BLAST hit is often not the nearest neighbor. *J. Mol. Evol*., **52**, 540–542.

Krylov,D.M. *et al*. (2003) Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. *Genome Res*., **13**, 2229–2235.

Liolios,K. *et al*. (2006) The Genomes On Line Database (GOLD) v.2: a monitor of genome projects worldwide. *Nucleic Acids Res*., **34**, D332–D334.

Lopez-Bigas,N. and Ouzounis,C.A. (2004) Genome-wide identification of genes likely to be involved in human genetic disease. *Nucleic Acids Res*., **32**, 3108–3114.

Makarova,K.S. *et al*. (2003) Potential genomic determinants of hyperthermophily. *Trends Genet*., **19**, 172–176.

Pellegrini,M. *et al*. (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl Acad Sci. USA*, **96**, 4285–4288.

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

Tatusov,R.L. *et al*. (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.*, **28**, 33–36.

Tekaia,F. and Yeramian,E. (2005) Genome Trees from Conservation Profiles. *PLoS Comput. Biol.*, **1**, e75.

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

Wall,D.P. *et al*. (2005) Functional genomic analysis of the rates of protein evolution. *Proc. Natl Acad. Sci. USA*, **102**, 5483–5488.

Wolf,Y.I. *et al*. (2001) Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context. *Genome Res.*, **11**, 356–372.