442–443 Nucleic Acids Research, 2003, Vol. 31, No. 1
DOI: 10.1093/nar/gkg039

The Ribosomal Database Project (RDP-II): previewing a new autoaligner that allows regular updates and the new prokaryotic taxonomy

J. R. Cole*, B. Chai, T. L. Marsh, R. J. Farris, Q. Wang, S. A. Kulam, S. Chandra, D. M. McGarrell, T. M. Schmidt[1], G. M. Garrity[1] and J. M. Tiedje[1]

Center for Microbial Ecology, 2225A Biomedical Physical Sciences and [1]Department of Microbiology and Molecular Genetics, Biomedical Physical Sciences, Michigan State University, East Lansing, MI 48824-4320, USA

Received September 13, 2002; Accepted September 24, 2002

ABSTRACT

The Ribosomal Database Project-II (RDP-II) provides data, tools and services related to ribosomal RNA sequences to the research community. Through its website (http://rdp.cme.msu.edu), RDP-II offers aligned and annotated rRNA sequence data, analysis services, and phylogenetic inferences (trees) derived from these data. RDP-II release 8.1 contains 16 277 prokaryotic, 5201 eukaryotic, and 1503 mitochondrial small subunit rRNA sequences in aligned and annotated format. The current public beta release of 9.0 debuts a new regularly updated alignment of over 50 000 annotated (eu)bacterial sequences. New analysis services include a sequence search and selection tool (Hierarchy Browser) and a phylogenetic tree building and visualization tool (Phylip Interface). A new interactive tutorial guides users through the basics of rRNA sequence analysis. Other services include probe checking, phylogenetic placement of user sequences, screening of users' sequences for chimeric rRNA sequences, automated alignment, production of similarity matrices, and services to plan and analyze terminal restriction fragment polymorphism (T-RFLP) experiments. The RDP-II email address for questions or comments is rdpstaff@msu.edu.

DESCRIPTION

This paper describes changes since the 2001 description (1). Details about specific data and analysis functions can be found at the Ribosomal Database Project-II (RDP-II) site (http://rdp.cme.msu.edu).

Data

The RDP-II obtains rRNA sequence data from the major International Nucleotide Sequence Databases (GenBank/EMBL/DDBJ). These sequences are provided to our users in aligned and annotated format. Release 8.1 contains 16 277 archaeal and bacterial, 1503 mitochondrial, and 5201 eukaryotic small subunit rRNA sequences. The number of eukaryotic sequences has more than doubled since our previous release. In order to provide a phylogenetic context for the data, RDP-II makes available over 100 trees that span the phylogenetic breadth of life. Most of these trees are new since our last update and were created using the WEIGHBOR algorithm (2). As a service to our users, we also provide a repository for published user-submitted alignments.

In addition to the release 8.1 alignments, we are offering a release 9 preview (beta) alignment containing over 50 000 (eu)bacterial small subunit rRNA sequences. This new alignment marks a major departure from prior RDP releases. Previous alignments incorporated secondary-structure constraints through manual inspection of each sequence. The new alignment was created using a modified version of the RNACAD (3), a Stochastic Context-Free Grammar (SCFG) based rRNA alignment program that incorporates rRNA secondary structure information directly in the internal model. The SCFGs are a class of probabilistic models related to the Hidden Markov Models (HMM) widely accepted in protein alignment and classification. We feel that this more automated alignment strategy will provide our users with a broader selection of up-to-date rRNA sequences in alignments that, for many purposes, are as good as hand-tuned alignments but without any unintended bias or other artifacts that may occur as a consequence of manual modification.

In addition to aligned sequences and trees, the RDP places the sequences into a phylogenetically consistent hierarchy. This hierarchy provides order to the collection. It provides a phylogenetic framework into which to place results of the RDP analysis functions, and it provides an entry point for users looking for sequences from specific groups of organisms.

Until very recently, the naming of higher-order prokaryotic taxa was based on phenotypic features rather than phylogeny. For this reason, the RDP has made a point to use trivial or colloquial names for its phylogenetically based hierarchy. In the second edition of the *Bergey's Manual of Systematic*

*To whom correspondence should be addressed. Tel: +1 5174324998; Fax: +1 5173538957; Email: rdpstaff@msu.edu

Bacteriology, Garrity and Holt (4) presented a new hierarchical classification of the prokaryotes. Their principal objective was to devise a scheme that reflected the phylogeny of prokaryotes (based on 16S rRNA sequence analysis). We have leveraged this new taxonomy for preview release 9, placing the sequences into a hierarchy that, as much as possible, is consistent with this updated taxonomy. We are assigning rRNA sequences to this phylogenetically consistent taxonomic hierarchy using a naïve Bayes classifier trained on a set of known type sequences supplemented with sequences from lineages that are only represented by unnamed or yet-uncultivated organisms.

Analysis services

A detailed description of each analysis command can be found on the www server.

The new Hierarchy Browser is a sequence search and selection tool that displays the sequences in an expandable hierarchical framework. It has visual indicators of sequence length and quality. Sequences and related annotation can be viewed online, selected for download, or selected for inclusion in other RDP-II analysis functions. For release 9, sequences may be viewed in either the new taxonomy or in a separate hierarchy matching their assignment in the NCBI taxonomy database (5). Subsets of the sequence data can be selected in release 9, including sequences from type strains only (identified in collaboration with Bergey's Trust), and near full-length sequences of 1200 or more bases.

Phylip Interface is a new interactive analysis function that permits phylogenetic reconstruction with a combination of user and RDP sequences. This tool offers a choice of either the PHYLIP (6) neighbor-joining, or WEIGHBOR (2) weighted neighbor-joining programs for phylogenetic calculation. The results can be viewed in a new interactive tree applet that allows phenogram rooting (and re-rooting) and branch rotation. The resulting phenogram can be downloaded in newick, postscript, eps and pdf formats.

Also new is an online interactive tutorial to guide users through the basics of rRNA sequence analysis. This tutorial is suitable both for the researcher new to rRNA based phylogenetic analysis and as a teaching module for upper-level undergraduate and graduate classes. The tutorial offers an introductory module plus individual modules for five of the RDP-II analysis functions: Chimera Check, Sequence Match, Sequence Aligner, Similarity Matrix and Phylip Interface.

Other tools are provided for comparing a user submitted sequence to the RDP-II database (Sequence Match), aligning a user sequence against the nearest RDP sequence (Sequence Aligner), examining probe and primer specificity (Probe Match), testing for chimeric sequences (Chimera Check), generating a similarity matrix (Similarity Matrix), analyzing T-RFLP data (T-RFLP and TAP-TRFLP), and browsing RDP phylogenetic trees (SubTree).

FUTURE CHANGES AND ADDITIONS

We expect feedback from our user community to determine the extent to which we adopt SCFG-based alignments but anticipate the release of additional SCFG alignments for *Archaea* and eukaryotes in the coming months, with the goal of monthly or better updates for these new alignments.

RDP-II ACCESS AND CONTACT

The RDP-II data and analysis services can be found at http://rdp.cme.msu.edu/. A mirror site is available at the Laboratory for Molecular Classification in the Center for Information Biology at the National Institute of Genetics (NIG), Japan (http://rdp.genes.nig.ac.jp/).

The address for email correspondence with RDP-II staff is rdpstaff@msu.edu. Telephone support is available at (+1 5174324998). The RDP-II staff may also be contacted via fax (+1 5173538957) or regular mail.

ACKNOWLEDGEMENTS

We thank several individuals for their past contributions: Robin Gutell (and his colleagues), Bonnie Maidak, Tim Lilburn, Niels Larsen, Tom Macke, Michael J. McCaughey, Ross Overbeek, Sakti Pramanik, Scott Dawson, Mitch L. Sogin, Gary Olsen and Carl Woese. The US Department of Energy Office of Science and the State of Michigan currently support RDP-II.

REFERENCES

1. Maidak,B.L., Cole,J.R., Lilburn,T.G., Parker,C.T., Saxman,P.R., Farris,R.J., Garrity,G.M., Olsen,G.J., Schmidt,T.M. and Tiedje,J.M. (2001) The RDP-II (Ribosomal Database Project). *Nucleic Acids Res.*, **29**, 173–174.
2. Bruno,W.J., Socci,N.D. and Halpern,A.L. (2000) Weighted neighbor joining: a likelihood-based approach to distance-based phylogeny reconstruction. *Mol. Biol. Evol.*, **17**, 189–197.
3. Brown,M.P.S. (2000) Small subunit ribosomal RNA modeling using stochastic context-free grammar. In *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology (ISMB 2000)*. San Diego, California, USA, pp. 57–66.
4. Garrity,G.M., Winters,M., Kuo,A.W. and Searles,D.B. (2002) *Taxonomic Outline of the Prokaryotes. Bergey's Manual of Systematic Bacteriology*, 2nd Edn, Springer-Verlag, NY.
5. Wheeler,D.L., Chappey,C., Lash,A.E., Leipe,D.D., Madden,T.L., Schuler,G.D., Tatusova,T.A. and Rapp,B.A. (2000) Database Resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **28**, 10–14.
6. Felsenstein,J. (1989) PHYLIP—Phylogeny Inference Package (Version 3.2). *Cladistics*, **5**, 164–166.