© 1994 Oxford University Press Nucleic Acids Research, 1994, Vol. 22, No. 17 3485–3487

The Ribosomal Database project

Bonnie L.Maidak*, Niels Larsen, Michael J.McCaughey, Ross Overbeek[1], Gary J.Olsen, Karl Fogel, James Blandy[2] and Carl R.Woese
Department of Microbiology, University of Illinois, 131 Burrill Hall, 407 South Goodwin Avenue, Urbana, IL 61801, [1]Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL 60439 and [2]Department of Biology, Indiana University, Jordan Hall 142, Bloomington, IN 47405, USA

ABSTRACT

The Ribosomal Database Project (RDP) is a curated database that offers ribosome-related data, analysis services, and associated computer programs. The offerings include phylogenetically ordered alignments of ribosomal RNA (rRNA) sequences, derived phylogenetic trees, rRNA secondary structure diagrams, and various software for handling, analyzing and displaying alignments and trees. The data are available via anonymous ftp (rdp.life.uiuc.edu), electronic mail (server@rdp.life.uiuc.edu) and gopher (rdpgopher.life.uiuc.edu). The electronic mail server also provides ribosomal probe checking, approximate phylogenetic placement of user-submitted sequences, screening for chimeric nature of newly sequenced rRNAs, and automated alignment.

DESCRIPTION

The Ribosomal Database Project (RDP) provides data, programs and services related to the ribosome. In this paper we summarize these offerings, the changes that have been introduced since last year's description (1), and some future features.

Data

The ribosomal RNA sequences in the RDP alignments are drawn from major sequence repositories [GenBank (2) and EMBL (3)] and direct submissions to the RDP. They are organized and presented in aligned and phylogenetically ordered form. Each sequence is annotated with its organismal source (for cultured organisms: the genus, species, culture collection numbers, etc.), cellular compartment, origin of sequence data (usually a literature citation), and other relevant information. If multiple versions of a given sequence exist, the RDP attempts to select by a variety of criteria which include the frequency of putative sequence errors and completeness only one of the versions for release. As a consequence, the number of released sequences is lower than if such selection was not performed. The RDP staff also examines the original publications and updates annotations, strain designations and organism names. Submitters and/or the public sequence databases are notified of possible errors.

The small subunit (SSU) rRNA alignments currently comprise sequences from approximately 120 Archaea, 2100 Bacteria (including chloroplasts and a few plant mitochondria) and 440 Eucarya (an alignment supplied by M. L. Sogin, Woods Hole Marine Biology Laboratory). A representative alignment of 76 prokaryotic small subunit rRNA sequences is also available. The number of large subunit (LSU) rRNA sequences remains at 150.

A phylogenetic tree is available for the sequences in the posted prokaryotic SSU rRNA alignment. It has been assembled from appropriately overlapping subtrees, each of which has been inferred using maximum-likelihood analysis (4,5). The current tree (and subsets of it) is available in printable text, PostScript, and Newick formats. The RDP also offers a collection of SSU and LSU rRNA secondary structure diagrams in PostScript format generated and supplied by R. Gutell and his collaborators (6).

Electronic mail server

Table 1 lists the available server commands. Changes in the past year have focused on improved analysis options and the new SUGGEST__TREE command.

Two options have been added to the ALIGN__SEQUENCE command. The COMPLETENESS__MINIMUM option prevents alignment of a submitted sequence against a partial sequence. The MOST__SIMILAR option allows the user to define the number of most similar sequences to be returned in aligned form.

The MY__DATABASE option enables treatment of submitted sequences as if they were part of an RDP data set. All user-supplied sequences remain private, and are not retained by the RDP or available to other users.

The new SUGGEST__TREE command performs a preliminary placement of one or more user-submitted sequences on the RDP phylogenetic tree, without changing the existing tree. The submitted sequence is first aligned using the algorithm underlying the ALIGN__SEQUENCE command. The unambiguously aligned positions are then submitted to fastDNAml (see Table 2 for description) which evaluates alternative placements on the RDP tree.

Programs

The programs currently available through the RDP servers are listed in Table 2. Programs added in the last year include DNArates and two data and graphics conversion programs

* To whom correspondence should be addressed

Table 1. Electronic mail server commands. Mail messages utilizing these commands should be sent to server@rdp.life.uiuc.edu

General functions	
HELP	Obtain general instructions for using the RDP mail server, or obtain a detailed description of a specified command.
SUBSCRIBE	Add your name to the RDP electronic mailing list for notifications about new data and services.
UNSUBSCRIBE	Remove your name from the RDP electronic mailing list.
Directory and file functions	
DIRECTORY	Obtain a listing of the files in an RDP directory or directory hierarchy.
INFORMATION	Obtain a description of the data in a specified RDP directory.
GET	Obtain a copy of a specified file.
Data retrieval	
FULL_ALIGNMENT	Obtain a copy of a complete sequence alignment. Options allow selection of the format.
SUBALIGNMENT	Obtain a subalignment containing specified sequences and/or positions from a larger alignment. Options allow selection of the format.
FULL_TREE	Obtain a copy of a phylogenetic tree in a requested format (printable text, PostScript, or Newick).
SUBTREE	Obtain a tree containing specified sequences from a larger tree in a requested format.
NAMES	Obtain a list of the names of the sequences represented in a specified alignment or tree.
Analytic functions	
ALIGN_SEQUENCE	Align a user-supplied sequence on the most similar sequence from the RDP. An option allows the user to avoid short matches with partial sequences.
CHECK_CHIMERA	Analyze a user-supplied sequence for evidence of chimeric structure. Options allow the user to add their own sequences to the database used in the analysis and to ignore short matches with partial sequences.
CHECK_PROBE	Analyze the occurrences of a specified 'probe' sequence in a set of sequences.
SIMILARITY_RANK	Obtain a list of the sequences most similar to that submitted. Options allow the user to add their own sequences to the database used in the analysis and to ignore short matches with partial sequences.
SUGGEST_TREE	Obtain an approximate placement of a user-submitted sequence in the RDP tree.
Defining the data to be used in analyses	
RDP_LIST	Use all available data in subsequent server commands.
REP_LIST	Use a standard representative subset of the available data in subsequent server commands.
MY_DATABASE	Add the user-provided sequences to the database used in SIMILARITY_RANK and CHECK_CHIMERA commands.
MY_LIST	Use the specified subset of available data in subsequent server commands.
MY_SEQUENCES	Provide sequence data for use in subsequent server commands.

obtained from other archival sites: EPSFilter and GraphicsConverter. Most of the other programs have been updated since their previous description (1).

RDP ACCESS AND CITATION

The RDP data can be accessed via anonymous ftp to rdp.life.uiuc.edu. Once you are logged in (using a user-id of 'anonymous' and your electronic mail address for password), examine the 00README files, which describe the organization of the data and programs.

The address of the automated electronic mail server is server@rdp.life.uiuc.edu. To obtain an overview of what data and services are currently available, send a mail message with the phrase 'help' as the body of the message. (Full command descriptions can be obtained by sending 'help complete' or 'help <command_name>'.) If your electronic mail address is unknown to the e-mail server, you will also receive a registration form. After returning the completed registration form, you will be automatically notified when new data or services become available.

The RDP gopher host name is rdpgopher.life.uiuc.edu. Gopher access to RDP data through the World Wide Web is also available (URL: gopher://rdpgopher.life.uiuc.edu/).

Electronic mail correspondence with RDP staff should be addressed to rdp@phylo.life.uiuc.edu. Those without access to electronic mail may contact the RDP curator (B.L.M.) via telephone (217-33-5866), fax (217-244-6697), or regular mail.

Research assisted by any RDP service should cite: the Ribosomal Database Project (RDP) at the University of Illinois in Urbana, Illinois; the release number; and this article (i.e., Maidak *et al.*, 1994). Please state which data, programs and services were used and the method of access.

Table 2. Programs available through the RDP servers

Convert_aln	A sequence alignment format conversion program for UNIX and VAX/VMS systems.
DNArates	A maximum likelihood method to estimate site-specific rates of nucleotide substitution from a sequence alignment and a user-defined phylogenetic tree. Data formats are similar to those used in J. Felsenstein's PHYLIP package. Compatible with a wide variety of computers.
Editor_AE2	An alignment editor and analysis program written by T. Macke for UNIX systems.
Editor_GDE	The Genetic Data Environment sequence alignment editing and analysis package written by S. Smith. Posted version is for Sun Microsystems computers.
EPSFilter	Macintosh program for working with Encapsulated PostScript (EPS) files written by B. Fowler.
fastDNAml	A maximum likelihood tree inference program based on version 3.3 of J. Felsenstein's DNAML. It has features to facilitate analysis of a larger number of taxa. Compatible with a wide variety of computers.
GraphicConverter	Macintosh program for conversion between graphics formats written by T. Lemke.
Readseq	A suite of sequence format conversion programs written by D. Gilbert. Compatible with a wide variety of computers.
SeqEdit	An alignment editor and analysis program for VAX/VMS systems.
Subalign	A program to extract specified rows and columns from an alignment. For UNIX and VAX/VMS systems.
TreeTool	A X-windows-based phylogenetic tree manipulation program for Sun Microsystems computers.

FUTURE CHANGES AND ADDITIONS

In addition to the curated data sets, unaligned rRNA sequences will be made available for some analyses (SIMILARITY_RANK, CHECK_PROBE, and CHECK_CHIMERA) by the mail server. This collection will be updated more frequently than the curated alignments and will enable these RDP analytical services to be performed on all rRNA sequences available from the public sequence databases at any time.

We are in the process of implementing a full interface to the RDP on the World Wide Web. Among other things, this will provide an alternative mode of access (beyond electronic mail) to the analysis services of the RDP.

A number of new programs and services are also planned. An alignment editor based on GNU Emacs v19 (7) is in development. The new editor will be mouse- and menu-driven and will feature extreme ease of customization. A neural network procedure for rRNA classification is being developed by Dr Cathy Wu, University of Texas at Tyler, in collaboration with the RDP. A faster version of CHECK_PROBE, which includes more options, is being tested. The SUGGEST_PROBE command, which will answer the question 'which probe(s) is most specific for my set of sequences?', remains in development.

ACKNOWLEDGEMENTS

We thank R.Gutell (and his colleagues) and M.L.Sogin for providing their data collections. We also thank Brad Strader, Tammy Nelson, and Michelle Vorwald for their valuable assistance. We thank the Free Software Foundation for excellent software. The RDP is largely supported by the National Science Foundation, Biological Instrumentation and Resources Division. G.J.O. is the recipient of a National Science Foundation Presidential Young Investigator Award.

REFERENCES

1. Larsen,N., Olsen,G.J., Maidak,B.L., McCaughey,M.J., Overbeek,R., Macke,T.J., Marsh,T.L., and Woese,C.R. (1993) *Nucleic Acids Res.*, **21**, 3021–3023.
2. Benson,D., Lipman,D.J. and Ostell,J. (1993) *Nucleic Acids Res.*, **21**, 2963–2965.
3. Rice,C.M., Fuchs,R., Higgins,D.G., Stoehr,P.J. and Cameron,G.N. (1993) *Nucleic Acids Res.*, **21**, 2967–2971.
4. Felsenstein,J. (1981) J. *Mol. Evol.*, **17**, 368–376.
5. Olsen,G.J., Matsuda,H., Hagstrom,R., and Overbeek,R. (1994) *Comput. Appl. Biosci.*, **10**, 41–48.
6. Gutell,R.R. *et al.* (1994) this issue.
7. Free Software Foundation, 675 Massachusetts Avenue, Cambridge, MA 02139, USA. Anonymous ftp access: prep.ai.mit.edu, cd pub/gnu.