The Ribosomal Database Project (RDP)

Bonnie L. Maidak*, Gary J. Olsen, Niels Larsen[1], Ross Overbeek[2], Michael J. McCaughey and Carl R. Woese

Department of Microbiology, University of Illinois, 131 Burrill Hall, 407 South Goodwin Avenue, Urbana, IL 61801-3704, USA, [1]Center for Microbial Ecology, Michigan State University, East Lansing, MI 48824, USA and [2]Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL 60439, 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), gopher (rdpgopher.life.uiuc.edu) and World Wide Web (WWW) (http://rdpwww.life.uiuc.edu/). The electronic mail and WWW servers provide ribosomal probe checking, screening for possible chimeric rRNA sequences, automated alignment and approximate phylogenetic placement of user-submitted sequences on an existing phylogenetic tree.

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 EBI (3)] and direct submissions to the RDP. They are organized and presented in an 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 the number of existing sequences. 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 ~140 Archaea, 2700 Bacteria (including chloroplasts and a few plant mitochondria) and 440 Eucarya (an alignment supplied by M. L. Sogin, Woods Hole Marine Biology Laboratory, MA). A representative alignment of 98 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 and eukaryotic (new this year) SSU rRNA alignments. They have been assembled from appropriately overlapping subtrees, each of which has been inferred using maximum-likelihood analysis (4,5). The current trees (and subsets of them) are 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 *et al.* (6). Also new this year is the corresponding taxonomic listing for SSU eukaryotic rRNA sequences.

As stated in last year's article, the RDP has made available this year unaligned data sets for use with the SIMILARITY_RANK, CHECK_PROBE and CHECK_CHIMERA commands (see description of these commands in Table 1). The current release contains 5500 and 1400 sequences respectively for the SSU and LSU data sets.

To facilitate access to specific rRNA aligned and unaligned sequences, the RDP offers subdirectories containing GenBank-formatted files of each sequence (directory names: aligned/sequences/[A-Z] and unaligned/sequences/[A-Z]).

Data servers

During the past year, an RDP WWW server has been developed. The initial part of its Home Page is shown in Figure 1.

Table 1 lists the commands available on the electronic mail server, which are also available on the WWW server.

Programs

The programs currently available through the RDP servers are listed in Table 2.

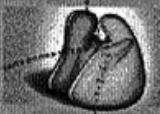* To whom correspondence should be addressed

Figure 1. The RDP home page

Table 1. Electronic mail server commands

General functions	
HELP	Obtain general instructions for using the RDP mail server, or obtain a detailed description of a specified command.
SUBSCRIBE (UNSUBSCRIBE)	Have your name put on (or taken off) the RDP electronic mailing list for notifications about new data and services.
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. Detects possible chimeric 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 in the RDP tree of a user-submitted sequence (using maximum likelihood analysis).
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 the commands SIMILARITY_RANK and CHECK_CHIMERA.
MY_LIST	Use the specified subset of available data in subsequent server commands.
MY_SEQUENCES	Provide sequence data for use in subsequent server commands.

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

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

RDP CITATION AND ACCESS

Research assisted by any RDP service should cite: the Ribosomal Database Project (RDP) at the University of Illinois in Urbana, IL; the release number and this article (i.e. Maidak *et al.*, 1996). Please state which data, programs and services were used and the method of access.

The RDP data and analysis services can be found at URL: http://rdpwww.life.uiuc.edu/.

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'). 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 WWW 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 (+1 217 333 5866), FAX (+1 217 244 6697) or regular mail.

FUTURE CHANGES AND ADDITIONS

Future plans for the RDP include improvements (i) in the display of and interaction with the phylogenetic trees, (ii) in the presentation and output options of the SUGGEST_TREE command and (iii) an improved version of the CHECK_PROBE function. Also planned is a sequence evaluation program, which assesses the quality of a user-supplied sequence, reporting back possible sequencing errors and/or idiosyncrasies, as well as a 'sequence signature' which defines the approximate taxonomic position of the sequence.

ACKNOWLEDGEMENTS

REFERENCES

1 Maidak,B.L., Larsen,N., McCaughey,M.J., Overbeek,R., Olsen,G.J., Fogel,K., Blandy,J. and Woese,C.R. (1994) *Nucleic Acids Res.*, **22**, 3485–3487.
2 Benson,D.A., Boguski,M., Lipman,D.J. and Ostell,J. (1994) *Nucleic Acids Res.*, **22**, 3441–3444.
3 Emmert,D.B., Stoehr,P.J., Stoesser,G. and Cameron,G.N. (1994) *Nucleic Acids Res.*, **21**, 3445–3449
4 Felsenstein,J. (1981) *J. Mol. Evol.*, **17**, 368–376.
5 Olsen,G.J., Matsuda,H., Hagstrom,R. and Overbeek,R. (1994) *CABIOS*, **10**, 41–48.
6 Gutell,R.R. *et al.* (1996) *Nucleic Acids Res.*, **24**, this issue.