The Ribosomal Database Project

Gary J.Olsen, Ross Overbeek[1], Niels Larsen, Terry L.Marsh, Michael J.McCaughey, Michael A.Maciukenas, Wen-Min Kuan, Thomas J.Macke[2], Yuqing Xing 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 Molecular Biology, MB1, Scripps Clinic, 10666 North Torrey Pines Road, La Jolla, CA 90237, USA

ABSTRACT

The Ribosomal Database Project (RDP) compiles ribosomal sequences and related data, and redistributes them in aligned and phylogenetically ordered form to its user community. It also offers various software packages for handling, analyzing and displaying sequences. In addition, the RDP offers (or will offer) certain analytic services. At present the project is in an intermediate stage of development.

DESCRIPTION

The sequences in the RDP alignments are drawn from various previously available rRNA collections (1–3), from major sequence repositories [GenBank (4) and EMBL (5)] and from individuals who have kindly deposited their own laboratory's rRNA sequence collection with the RDP. Although sequences are now manually aligned [on the basis of both primary and higher-order structural similarity (6)], automated alignment programs are being developed.

At present the RDP's sequence offerings include an aligned, phylogenetically ordered set of prokaryotic small subunit rRNAs, both bacterial and archaeal, a comparable alignment of eukaryotic small subunit rRNA sequences (kindly supplied by M.L.Sogin, Woods Hole Marine Biology Laboratory), and a large subunit rRNA alignment comprising prokaryotic, chloroplast and eukaryotic sequences. Users may choose which sequences they wish to receive from any alignment, as well as the specific portions of those sequences. In addition, a representative collection, comprising a small but phylogenetically diverse selection of sequences is available for prokaryotic small subunit rRNA alignment. The sequences are available in the user's choice of formats [GenBank and EMBL (with inserted alignment gaps), AE2 editor format, PAUP (7), PHYLIP (8), a printable text file, or others when these can be readily accommodated]. Sequence alignments are routinely available by electronic mail or anonymous ftp, or by tape or diskette if special arrangement is made.

The current data are available via anonymous ftp to info.mcs.anl.gov (currently 140.221.10.1). Once you are logged into that machine (using a user-id of 'anonymous' and your electronic mail address for password), the database can be reached with the command 'cd pub/RDP'. You should get and examine the README file and then access whichever files you find appropriate.

For automated electronic mail access, use the address rdp@mcs.anl.gov. To obtain an overview of what data and services are currently available, send a mail message with the word 'help' as the body of the message. The mail server will reply by sending you a description of the server functions. Table 1 provides a summary of the server functions.

Sequences are distributed in a phylogenetically ordered form. The underlying phylogenetic tree (inferred by a maximum-likelihood method) is available from the electronic mail server

Table 1. Mail Server Commands[1]

HELP	Obtain either general instructions for using the RDP mail server, or a detailed description of a specified command
INFORMATION	Obtain a description of the data in a specified RDP directory
DIRECTORY	Obtain a listing of the files in an RDP directory or directory hierarchy
GET	Obtain a copy of a file from an RDP directory
NAMES	Obtain a list of the names of the sequences represented in a specified alignment or tree
MY_LIST	Define a subset of the sequences in an alignment or tree for use by subsequent server commands
MY_SEQUENCES	Specify sequence data for use by subsequent server commands
FULL_ALIGNMENT	Obtain a copy of a sequence alignment in a requested format
SUBALIGNMENT	Obtain an alignment containing a specified subset of the sequences and/or positions from a larger alignment
TREE	Obtain a copy of a phylogenetic tree in a requested format
SUBTREE	Obtain a tree containing a specified subset of the sequences from a larger tree
CHECK_PROBE	Analyze the occurrences of a specified 'probe' sequence in an alignment
SUBSCRIBE	Have your name put on the RDP electronic mailing list for notifications about new data and services
UNSUBSCRIBE	Have your name removed from the RDP electronic mailing list

[1]Mail messages utilizing these commands should be sent to rdp@mcs.anl.gov

* To whom correspondence should be addressed

in a machine-readable Newick format and either 'printer plot' or PostScript representations of a phenogram-style tree drawing. The user may also specify the subset of the sequences that he/she wishes included in returned tree.

The RDP also distributes the collection of rRNA secondary structure diagrams created by Gutell *et al.* (2).

The RDP's software offerings now include three sequence 'editors': the Olsen VAX/VMS-based editor; the Genetic Data Environment (GDE) X-window-based package (designed by S. Smith, initially as an RDP project); and AE2, a UNIX- based alignment editor developed by T.J.M. (and recently updated by him in collaboration with the RDP), which runs without modification on Sun workstations. As it becomes available, the RDP will also distribute the Macintosh-based multiple sequence editor and analysis environment written by D. Gilbert (Indiana University). Another new addition to the RDP server is TreeTool, an X-windows-based program (M.A.M., unpublished); it reads Newick tree format [as written by PAUP (7), PHYLIP (8), and MacClade (9)] and allows ready manipulation of phylogenetic trees, including reordering branches, moving portions of a tree, and rerooting trees.

The RDP is also providing and/or working toward several analytic services (via its electronic mail server):

1. Maximum-likelihood tree inference for nucleotide sequences. The user-supplied alignment will be subjected to an analysis similar to that performed by Felsensteins's PHYLIP version 3.3 dnaml program (8). The output will include a description of the analysis performed, a diagram of the inferred tree, and a Newick format description of the unrooted tree.

2. A probe design service. At present the only aspect of this package available is a very simple 'probe checker', which screens the user-supplied probe sequence against the RDP's full sequence database, returning a listing of the sequences in which the (complementary) match is above a specified threshold, including the actual matches and the positions in those sequences where the matches occur.

3. A 'sequence assessment' system (in the planning and design stage). The intent is to provide an alignment of a user-supplied sequence relative to a prealigned reference set, and report salient sequence characteristics such as idiosyncrasies, group diagnostic features, and possible sequencing errors.

If you wish to be notified via electronic mail when new data and services become available, use the mail server SUBSCRIBE command noted in Table 1.

The electronic mail address for RDP correspondence (as opposed to automated mail server functions) is rdp@origin.life.uiuc.edu. Telephone contact is through Terry Davis at 217-333-1142.

ACKNOWLEDGMENTS

The RDP is largely supported by the National Science Foundation, Division of Instrumentation and Resources. G.J.O. is the recipient of a National Science Foundation Presidential Young Investigator Award. N.L. is supported by grant 11-8804 from the Danish Natural Sciences Research Council. T.A.M. is the recipient of a National Institutes of Health Senior Research Fellowship.

REFERENCES

1. Neefs,J.-M., Van de Peer,Y., Hendriks,L. and De Wachter,R. (1990) *Nucleic Acids Res.*, **18**, 2237–2317.
2. Gutell,R.R., Schnare,M.N. and Gray,M.W. (1992) *Nucleic Acids Res.*, [in this volume].
3. Specht,T., Wolters,J. and Erdmann,V.A. (1990) *Nucleic Acids Res.*, **18**, 2215–2235.
4. Burks,C., Cassidy,M., Cinkosy,M.J., Cumella,K.E., Gilna,P., Hayden,J.E.-D., Keen,G.M., Kelley,T.A., Kelley,M., Kristofferson,D. and Ryals,J. (1991) *Nucleic Acids Res.*, **19**, 2221–2225.
5. Stoehr,P.J. and Cameron,G.N. (1991) *Nucleic Acids Res.*, **19**, 2227–2230.
6. Woese,C.R., Gutell,R.R., Gupta,R. and Noller,H.F. (1983) *Microbiol. Rev.*, **47**, 621–669.
7. Swofford,D.L. (1990) *PAUP: Phylogenetic Analysis Using Parsimony, Version 3.0*, Illinois Natural History Survey, Champaign, IL.
8. Felsenstein,J. (1990) *PHYLIP Manual Version 3.3*. University Herbarium, University of California, Berkeley, CA.
9. Maddison,W.P. (1989) *Folia Primatol.*, **53**, 190–202.