Nucleic Acids Research, 1993, Vol. 21, No. 13 3021–3023

The ribosomal database project

Niels Larsen, Gary J.Olsen, Bonnie L.Maidak*, Michael J.McCaughey, Ross Overbeek[1], Thomas J.Macke[2], Terry L.Marsh 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) is a curated database that offers ribosome data along with related programs and services. The offerings include phylogenetically ordered alignments of ribosomal RNA (rRNA) sequences, derived phylogenetic trees, rRNA secondary structure diagrams and various software packages for handling, analyzing and displaying alignments and trees. The data are available via ftp and electronic mail. Certain analytic services are also provided by the electronic mail server.

DESCRIPTION

The ribosomal RNA sequences in the RDP alignments are drawn from major sequence repositories [GenBank (1) and EMBL (2)], other public or private rRNA sequence collections (e.g., 3), and direct submissions to RDP. The updating process has been largely automated, including alignment of new sequences. We currently release new versions at a rate of four per year. All offerings can be accessed by anonymous ftp, and more flexibly by electronic mail. The mail server includes some analytic functions as well.

Data

The sequence data are organized and presented in aligned and phylogenetically ordered form. When multiple versions of a given sequence exist, only the one deemed 'best' (usually the most complete) is posted, although all are archived. Each sequence is also annotated with its organismal source (for cultured organisms: the genus, species, culture collection numbers, etc.), origin of sequence data (usually a literature citation), cellular compartment, and other relevant and useful information. The small subunit (SSU) rRNA collection currently comprises sequences from approximately 100 Archaea, 1400 Bacteria (and chloroplasts) and 350 Eucarya (an alignment supplied by M. L. Sogin, Woods Hole Marine Biology Laboratory). An alignment containing a smaller number of representative sequences is available as well. The large subunit (LSU) rRNA alignment comprises about 150 sequences spanning all three phylogenetic domains. A phylogenetic tree for the posted SSU rRNA sequences is provided. It is an assembly of partially overlapping subtrees that were inferred using maximum-likelihood analyses (4). The RDP offers a collection of SSU and LSU secondary structure diagrams in PostScript format, generated and supplied by R.Gutell and his collaborators (5).

Programs

The RDP offers three different sequence alignment editors: AE2 (Alignment Editor 2, UNIX-based; T.J.M.), SeqEdit (VAX/VMS-based; G.J.O.), and GDE (Genetic Data Environment for UNIX and X-windows; written by S.Smith). The program ReadSeq (runs on multiple platforms, contributed by D.Gilbert) converts between multiple sequence and alignment formats. For phylogenetic tree inference we offer fastDNAml (G.J.O., R.Overbeek, H.Matsuda and R.Hagstrom, unpublished). This program is an adaptation of J.Felsenstein's DNAML program (4) applicable to large numbers of DNA/RNA sequences (over 60 on a desktop workstation). It runs on many different single and multiprocessor computer platforms. The interactive program TreeTool (requires UNIX and X-windows, written by M.Maciukenas, unpublished) allows ready manipulation of phylogenetic trees, including the reordering of branches, cutting and pasting of subtrees, and rerooting of trees.

Electronic mail server functions

The RDP electronic mail server (developed by M.J.M) offers flexible data access. The server reads and interprets incoming mail, executes the requested operations, and returns the results to the sender. The following are some of the server commands (Table 1 provides a complete list).

The SUBALIGNMENT command returns sections of an RDP alignment in a user-requested format [GenBank and EMBL (with inserted gaps), AE2 editor, PAUP (6), PHYLIP (7), or printable text]. The SUBTREE command extracts user-selected portions of an RDP tree and formats them as requested [printable text, PostScript, or 'Newick 8:45,' which is suitable for programs and packages such as PAUP, PHYLIP, and MacClade (8)].

The electronic mail server also provides analytic services. The command SIMILARITY_RANK (written by N.L.) lists the sequences in a specified data set that are most similar to a user-submitted sequence by comparing oligomers (currently octamers) in the submitted sequence to those of each sequence in the specified data set. It can be used for unaligned rRNA fragments

* To whom correspondence should be addressed

Table 1. Mail Server Commands[1]

ALIGN_SEQUENCE	Returns submitted sequence(s) in aligned form
CHECK_CHIMERA	Detects possibly chimeric sequences
CHECK_PROBE	Analyze the occurrences of a specified 'probe' sequence in an alignment
DIRECTORY	Obtain a listing of the files in an RDP directory or directory hierarchy
FULL_ALIGNMENT	Obtain a copy of a sequence alignment in a requested format
FULL_TREE	Obtain a copy of a phylogenetic tree in a requested format
GET	Obtain a copy of a file from an RDP directory
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
MY_LIST	Δefine 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
NAMES	Obtain a list of the names of the sequences represented in a specified alignment or tree
RDP_LIST	Use full set of available data
REP_LIST	Use a standard representative subset of the available data
SIMILARITY_RANK	Return a list of most similar sequences to the submitted.
SUBALIGNMENT	Obtain an alignment containing a specified subset of the sequences and/or positions from a larger alignment
SUBSCRIBE	Have your name put on the RDP electronic mailing list for notifications about new data and services
SUBTREE	Obtain a tree containing a specified subset of the sequences from a larger tree
UNSUBSCRIBE	Have your name removed from the RDP electronic mailing list

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

as short as 150–200 nucleotides, which allows the rapid screening of large collections of short rRNA sequences (e.g., those produced by environmental characterizations) in order to identify those worthy of more complete sequence determination. Similarly, the CHECK_CHIMERA command (by N.L.) provides a quick screen for chimeric sequences (e.g., resulting from PCR amplification). This is done by sliding a window along a given sequence and determining a similarity ranking at each position. Chimeras will show different sets of 'most similar' sequences as the window processes down the sequence. The program returns a summary of these sets as ordered lists together with the approximate point of junction between the chimeric fragments. CHECK_PROBE (written by N.L.) assesses a submitted probe/primer sequence as to its uniqueness and 'stability of annealing' within a specified sequence collection. The program also estimates a Td (dissociation temperature) for perfectly matching probes based upon an empirical formula involving G+C content and cation concentration. An alignment program (command ALIGN_SEQUENCE, written by N.L.) can quickly align a user submitted sequence against its most similar sequence in the specified RDP dataset. The algorithm is a mixture of the known 'linked lists' approach and a unique method that takes local sequence composition and random expectation in account. The returned output can be either a human readable listing or a computer readable GenBank flat file, plus an additional entry identifying ambiguously aligned positions.

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

AVAILABILITY

The RDP servers have recently been moved to the University of Illinois, and thus have new Internet addresses.

The RDP data can be accessed via anonymous ftp to rdp.life.uiuc.edu (currently 128.174.86.14). Once you are logged into that machine (using a user-id of 'anonymous' and your electronic mail address for password), examine the 00README files, which describe the organization of the data.

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 word 'help' as the body of the message. The mail server will reply by sending you a description of the server functions.

The electronic mail address for RDP correspondence (as opposed to automated mail server functions) is rdp@phylo.life.uiuc.edu. Those without access to electronic mail may contact the RDP curator (B.L.M.) via telephone (217-333-5866), fax (217-244-6697), or regular mail.

FUTURE CHANGES AND ADDITIONS

Two additional analytic services are close to completion. TREE_SEQUENCE will add a user-supplied sequence to a copy of an RDP alignment, assess the regions of alignment certainty, and then use fastDNAml to insert the new sequence into a copy of an RDP tree. The inferred sequence alignment and phylogenetic placement (relative to a selected subset of sequences) will be returned in user-requested formats. SUGGEST_PROBE (N.L., in development) is a probe design service that answers the question 'which rRNA probe sequence(s) would be most effective for a given set of organisms?' The program will take into account experimental (hybridization) conditions.

We will also have a 'sequence assessment' function that reports salient sequence characteristics such as idiosyncrasies, group diagnostic features, and possible sequencing errors.

ACKNOWLEDGEMENTS

We thank R.Gutell (and his colleagues) and M.L.Sogin for providing their data collections. We also thank Brad Strader and Tammy Nelson for their valuable assistance. We thank Free Software Foundation for excellent software. 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. was supported by grant 11-8804 from the Danish Natural Sciences Research Council. T.L.M. is the recipient of a National Institutes of Health Senior Research Fellowship.

REFERENCES

1. Burks,C., Cinkosky,M.J., Fischer,W.M., Gilna,P., Hayden,J.E.-D., Keen,G.M., Kelly,M., Kristofferson,D. and Lawrence,J. (1992) *Nucleic Acids Res.*, **20**, 2065–2069.
2. Higgins,D.G., Fuchs,R., Stoehr,P.J. and Cameron,G.N. (1992) *Nucleic Acids Res.*, **20**, 2071–2074.
3. De Rijk,P., Neefs,J.-M., Van de Peer,Y., and De Wachter,R. (1992) *Nucleic Acids Res.*, **20**, 2075–2089.
4. Felsenstein,J. (1981) *J. Mol. Evol.*, **17**, 368–376.
5. Gutell,R.R. et al. (1993) this issue.
6. Swofford,D.L. (1990) PAUP: Phylogenetic Analysis Using Parsimony, Version 3.0, Illinois Natural History Survey, Champaign, IL.
7. Felsenstein,J. (1989) *Cladistics*, **5**: 164–166.
8. Maddison,W.P. (1989) *Folia Primatol.*, **53**, 190–202.