Published online 7 November 2006 *Nucleic Acids Research, 2007, Vol. 35, Database issue* ***D169–D172***
doi:10.1093/nar/gkl889

The ribosomal database project (RDP-II): introducing *myRDP* space and quality controlled public data

J. R. Cole[1,*], B. Chai[1], R. J. Farris[1], Q. Wang[1], A. S. Kulam-Syed-Mohideen[1], D. M. McGarrell[1], A. M. Bandela[1], E. Cardenas[1], G. M. Garrity[1,2] and J. M. Tiedje[1,2]

[1]Center for Microbial Ecology and [2]Department of Microbiology and Molecular Genetics, Biomedical Physical Sciences, Michigan State University, East Lansing, MI 48824-4320, USA

Received September 15, 2006; Accepted October 9, 2006

ABSTRACT

Substantial new features have been implemented at the Ribosomal Database Project in response to the increased importance of high-throughput rRNA sequence analysis in microbial ecology and related disciplines. The most important changes include quality analysis, including chimera detection, for all available rRNA sequences and the introduction of *myRDP* Space, a new web component designed to help researchers place their own data in context with the RDP's data. In addition, new video tutorials describe how to use RDP features. Details about RDP data and analytical functions can be found at the RDP-II website (http://rdp.cme.msu.edu/).

DESCRIPTION

As of September 2006 (Release 9.42), the RDP maintained 262 030 aligned and annotated public rRNA sequences. Of these, 84 442 were from cultivated bacterial strains, while 177 588 were derived from environmental samples. A total of 101 877 sequences were near-full-length (≥1200 bases) and 5543 sequences were from bacterial type strains; these sequences are of special importance as they help to link taxonomy and phylogeny. As described in detail in our previous update (1), the collection is updated monthly from the International Nucleotide Sequence Database Collaboration (INSDC; DDBJ, EMBL and GenBank). New sequences are automatically aligned with the RDP's modified version of RNACAD (2), a stochastic context-free grammar based aligner trained with the secondary structure model of Robin Gutell and colleagues (3). Paired and unpaired positions that occur in >95% of bacterial species, accounting for ~97% of 16S rRNA residues in the average sequence, are modeled by the program as alignable positions. Each month, changes in annotation of existing INSDC records are propagated to the corresponding RDP record, while new updates to INSDC sequence data trigger a retirement and replacement of the corresponding RDP record.

As a major quality improvement, all sequences are now tested for sequence anomalies, including chimeric sequence anomalies, using Pintail from the Cardiff Bioinformatics Toolkit (4). Using Pintail on a subset of the RDP public sequences, those authors reported that at least 5% of rRNA records contain some type of anomaly. We employ a similar strategy to detect anomalous sequences. Each sequence is compared with at least two sequences from different publications and those reported as anomalous in both comparisons are marked as suspect. (For a small percentage of sequences, results of the first two tests are not consistent and additional comparisons are necessary to establish a pattern.) Of the 262 030 sequences in release 9.42, 21 771 are deemed anomalous by this criterion. When the sequences are subdivided based on source (isolate versus environmental) and short versus long, we find the anomalies are greatest in the environmental and short sequences (Table 1).

myRDP space

This major upgrade allows users to maintain their own private sequence collection on the RDP servers aligned in sync with the RDP public alignment. With a *myRDP* account, researchers upload private rRNA sequences in *Sequence Groups* which can range from a single sequence to thousands of sequences. Normally a group would consist of a set of sequences from a single project, but can be grouped according to the end-user's needs. Researchers may designate other *myRDP* users as *Research Buddies* and grant them access to individual Sequence Groups or grant a Research Buddy blanket access to all the researcher's data. This latter feature makes it particularly simple to share data with a mentor, a supervisor or collaborators.

After upload, *myRDP* sequences are automatically placed into the bacterial taxonomy using the RDP Classifier and aligned to match the RDP public alignment using RNACAD, the same context-free grammar based, secondary structure aware aligner that is used to produce the public alignment. Since the aligner is cpu intensive, alignment jobs are passed to a small cluster of compute servers to minimize impact on other RDP services. Alignments are completed in a few minutes to a few hours depending on load and are ready for

*To whom correspondence should be addressed. Tel: +1 517 353 3842; Fax: +1 517 353 8957; Email: colej@msu.edu

© 2006 The Author(s).
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from nar.oxfordjournals.org by guest on February 23, 2011

Table 1. Percent of low or suspect sequences in the RDP database

	Near full-length (≥1200 bp)		Short	
	Total	% Suspect[a]	Total	% Suspect
Environmental	54 412	6.9	123 176	10.1
Isolate	47 465	3.6	36 977	10.5
Total	101 877	5.4	160 153	10.2

[a]Percent of sequences flagged as containing substantial anomalies by Pintail in comparisons with trusted sequences from at least two different publications.

use the next time the researcher enters his or her *myRDP* Space. Since the alignment remains in sync with the RDP public alignment, there is no need for the alignment compromises necessary to maintain compatibility between physically separated alignments.

When entering their *myRDP* Space researchers are presented with a list of available Sequence Groups. New Sequence Groups may be uploaded at any time. Individual Sequence Groups can be browsed in either list format or in a hierarchical taxonomic representation and any combination of *myRDP* sequences and RDP public sequences can be selected for download or further analysis. Sequences can be downloaded in formats ready for input to a wide variety of third-party phylogenetic and ecological tools.

myRDP Pipeline. The new *myRDP* release incorporates a high-throughput sequence processing pipeline tailored to the requirements of single read environmental sequence projects. It provides the researcher with a simple path from sequencer output to quality-controlled, aligned sequences and analysis. The *myRDP* Pipeline consists of an integrated suite of publicly available and in-house developed programs. Users enter information about one or more sequencing *Library Runs* consisting of a set of sequencing reactions each from a single representative from one environmental library. Raw sequence trace files are then uploaded and processed into trimmed sequences and quality control data.

In the first stage, the raw sequence trace files are translated into trimmed quality sequence data and entered into the *myRDP* database. The researcher enters information about a Library Run, including cloning vector name and primer sequence. If the researcher wishes, she can include information about the layout of the sequencing reactions in 96 well microtiter plates. A directory of raw sequence trace files in .zip or .tar format is then uploaded to the *myRDP* Pipeline server. These files are converted into base calls and quality control information using Phred (5,6). Trimming and vector removal are then done using TIGR's Lucy (7). We have found it necessary to tune the base call and trim parameters in these initial sequence processing steps to optimize Phred and Lucy for single read environmental sequences as compared to the standard parameters used for reads destined for high-coverage sequence assembly.

The base calls, trim points and base quality information are stored on the *myRDP* Pipeline server in the user's account. These unaligned sequences can be downloaded for custom processing or troubleshooting. If microtiter plate information was loaded when the Library Run was created, the individual plates can be examined in a schematic plate representation depicting good and failed sequencing reactions. Horizontal or vertical patterns can indicate systematic failures in the sequencing or pre-sequencing mechanics.

Library Run sequences with Phred quality scores above Q20 for at least 200 bases are bundled as a Sequence Group and further processed through *myRDP* after masking bases with quality scores below Q20 to 'N'. These new Sequence Groups are listed with all others on the researcher's *myRDP* overview page and can be treated as any other Sequence Group, shared with Research Buddies and selected for download or analysis. The original Library run information is always available on the *myRDP* Pipeline pages and can be accessed at any time if quality questions should arise.

Analysis services

Several RDP analysis services have been modified to provide extra features with aligned *myRDP* sequences. Theses additional features are noted below, along with analysis functions new or modified since our 2005 report.

Hierarchy Browser. A greatly enhanced search feature allows both free text and field-specific searches, as well as allowing Boolean logic in search queries. Search terms can be directly entered in the search box. Simple instructions are provided to allow end-users to formulate more complex queries. Queries can be used to limit the search to specific annotation fields (RDP sequence identifier, INSDC accession, modification date, definition line, source, organism, references or feature table), implement Boolean logic (AND, OR, NOT), to formulate range searches (for example, all INSDC accessions from DQ179017 to DQ179020) and to formulate proximity searches (multiple terms in the same phrase). A new sequence quality filter, in addition to the species type strain, environmental and near-full-length filters, has been added to the filtering process based on the Pintail results.

A new Publication View feature allows users to display sequences from individual publications. This view displays publications (including unpublished INSDC submission citations) as a list ordered by the count of sequences referencing the publication. By default, the list displays all publications referenced by 100 or more rRNA sequences, but this limit is user adjustable. For those publications abstracted by PubMed, the PMID is provided as a hyperlink to the publication abstract. Clicking on the sequence count for a citation will link back to the hierarchy view displaying only sequences from that publication. The researcher can then browse all sequences from the publication and select all or any portion of the sequences based on taxonomic assignment (e.g. all Firmicutes).

Sequence Carts are lists of selected RDP sequence ids (both public and *myRDP*) that can be downloaded for storage on the researcher's computer and subsequently uploaded to automatically select desired sets of sequences. (If the uploaded cart contains *myRDP* sequence identifiers, only *myRDP* sequences for which the researcher has valid access rights, either directly or as a Research Buddy, will be retrieved.) Sequence Carts can also be created as simple lists of INSDC accessions to accommodate researchers with such lists from third-party tools. (For INSDC accessions containing more than one 16S gene, such as genome records, all corresponding genes will be selected upon upload.) Sequence Carts save the user from repeating complex

Downloaded from nar.oxfordjournals.org by guest on February 23, 2011

selections for subsequent re-analysis and can simplify the inclusion of the same set of landmark sequences in multiple analyses.

Sequence Match. This tool has been modified to report the standard pairwise percent identity suitable for publication when used with *myRDP* or public RDP sequences as queries. Unlike BLAST (8), which reports only local identity over a portion of the sequence, this program uses alignment information to calculate the percent identity over all comparable positions (aligned positions with a base in both sequences). Changing filter options on the fly makes it easy for the researcher to find in succession the closest pairwise species type strains, bacterial isolates, environmental sequences and/or short partial sequences. And, as shown previously, Sequence Match is more reliable than BLAST at retrieving the closest pairwise matches (1). In addition, a new filter has been added to allow limiting matches based on Pintail results, as for the Hierarchy Browser.

RDP Classifier. When used with *myRDP* and/or public RDP sequences, the RDP Classifier uses the stored classification results to more rapidly display results in a taxonomic hierarchy representation.

New! Library Compare. Microbial community comparison based on 16S rRNA gene sequence libraries has become commonplace in microbial ecology. However, most comparison methods, whether from traditional macroecology, such as Sorensen's index (9), or designed specifically for sequence library data, such as LIBSHUFF (10) or Martin's *P*-test (11), only provide summary information about the degree of difference between communities. These methods fail to put differences in a taxonomic context. The RDP Sample Comparison Tool combines the RDP Classifier with a statistical test to detect significant differences in the taxonomic composition between two samples. Results are presented in an interactive display that allows the selection of any taxon and presents the user with a graphical representation of the immediate subtaxa along with statistical significance information about differences between samples. (The statistic is not corrected for multiple tests, so caution in interpretation is warranted.)

Tree Tool. This upgraded tool replaces a similar tool offered in older versions of RDP. Selections containing both *myRDP* and public RDP sequences are used to create a phylogenetic tree using the Weighbor weighted neighbor-joining tree building algorithm (12). The results are presented in an interactive Java applet that allows users to re-root the tree, rearrange nodes and make other cosmetic changes. The modified tree can be downloaded in Newick format or as a PostScript file suitable for embellishment in drawing programs such as Adobe Illustrator.

Download formats. Downloads can include any mixture of public RDP and *myRDP* sequences. FASTA and PHYLIP alignment formats can be imported into many common molecular phylogeny programs, including ARB (13), PHYLIP (http://evolution.genetics.washington.edu/phylip.html), fastDNAml (14,15), MEGA (16), PAUP* (http://paup.csit.fsu.edu/paupfaq/faq.html) and others. In addition, a pairwise distance matrix suitable for the DOTUR (17) molecular macroecology package as well as many phylogenetic packages such as Weighbor and PHYLIP can be created and downloaded with either uncorrected or Jukes-Cantor-corrected distances. For users of the ARB package, a 'navigation tree' for the selected sequences can be downloaded and imported directly, along with the sequences, into that package. An alignment mask pseudo-sequence that can be used to 'mask out' unalignable variable regions from further analysis and a canonical base pair structure pseudo-sequence are available for inclusion with alignment downloads from *myRDP*. A special tool produces input appropriate for the EstimateS macroecology program (http://viceroy.eeb.uconn.edu/EstimateS).

Video tutorials and documentation. In addition to greatly expanded help files, new short video tutorials demonstrate some of the more complex tasks, including use of the *myRDP* Pipeline. These tutorials average 3 min in length. They capture the screen as the tasks are performed while the narrator explains the tasks and the choices available to the user.

RDP-II ACCESS AND CONTACT

The RDP-II data and analysis services can be found at http://rdp.cme.msu.edu/. The RDP's mission includes user support. Support questions can be emailed to rdpstaff@msu.edu. Telephone support is available (+1 517 432 4998). The RDP-II staff may also be contacted via fax (+1 517 353 8957 Attn:RDP) or regular mail.

ACKNOWLEDGEMENTS

We thank Vincent Young, Terrence Marsh, Thomas Schmidt, Bradley Stevenson and Heather Wood for help developing the *myRDP* Pipeline. We thank several individuals for their past contributions: Robin Gutell (and his colleagues), Chuck Parker, Paul Saxman, 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. This research was supported by the Office of Science (BER), U.S. Department of Energy, Grant No. DE-FG02-99ER62848 and the National Science Foundation, Grant No. DBI-0328255. Additional support for the development of the *myRDP* Pipeline was provided by the Michigan State University Center for Microbial Pathogenesis. Funding to pay the Open Access publication charges for this article was provided by the Office of Science (BER), U.S. Department of Energy.

Conflict of interest statement. None declared.

REFERENCES

1. Cole,J.R., Chai,B., Farris,R.J., Wang,Q., Kulam,S.A., McGarrell,D.M., Garrity,G.M. and Tiedje,J.M. (2005) The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis. *Nucleic Acids Res*., **33**, D294–D296.
2. Brown,M.P.S. (2000) Small subunit ribosomal RNA modeling using stochastic context-free grammars. In *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology (ISMB 2000)*, San Diego, CA. pp. 57–66.
3. Cannone,J.J., Subramanian,S., Schnare,M.N., Collett,J.R., D'Souza,L.M., Du,Y., Feng,B., Lin,N., Madabusi,L.V., Muller,K.M.

Downloaded from nar.oxfordjournals.org by guest on February 23, 2011

et al. (2002) The Comparative RNA Web (CRW) Site: An Online Database of Comparative Sequence and Structure Information for Ribosomal, Intron and other RNAs. *BMC Bioinformatics*, **3**, 2.
4. Ashelford,K.E., Chuzhanova,N., Fry,J.C., Jones,A.J. and Weightman,A.J. (2005) At least one in twenty 16S rRNA sequence records currently held in public repositories is estimated to contain substantial anomalies. *Appl. Environ. Microbiol.*, **71**, 7724–7736.
5. Ewing,B., Hillier,L., Wendl,M. and Green,P. (1998) Base-calling of automated sequencer accuracy assessment. *Genome Res.*, **8**, 175–185.
6. Ewing,B. and Green,P. (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.*, **8**, 186–194.
7. Chou,H.-H. and Holmes,M.H. (2001) DNA sequence quality trimming and vector removal. *Bioinformatics*, **17**, 1093–1104.
8. McGinnis,S. and Madden,T.L. (2004) BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res.*, **32**, W20–W25.
9. Sorenson,T. (1948) A method of establishing groups of equal amplitude in a plant based on similarity of species content and its applications to analysis of vegetation on Danish commons. *Biol. Skr.*, **5**, 1–34.
10. Singleton,D.R., Furlong,M.A., Rathburn,S.L. and Whitman,W.B. (2001) Quantitative comparisons of 16S rRNA gene sequence libraries from environmental samples. *Appl. Environ. Microbiol.*, **67**, 4374–4376.
11. Martin,A.P. (2002) Phylogenetic approaches for describing and comparing the diversity of microbial communities. *Appl. Environ. Microbiol.*, **68**, 3673–3682.
12. 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.
13. Ludwig,W., Strunk,O., Westram,R., Richter,L., Meier,H., Yadhukumar, Buchner,A., Lai,T., Steppi,S., Jobb,G. *et al*. (2004) ARB: a software environment for sequence data. *Nucleic Acids Res.*, **32**, 1363–1371.
14. Olsen,G.J., Matsuda,H., Hagstrom,R. and Overbeek,R. (1994) fastDNAml: a tool for construction of phylogenetic trees of DNA sequences using maximum likelihood. *Comput. Appl. Biosci.*, **10**, 41–48.
15. Felsenstein,J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.*, **17**, 368–376.
16. Kumar,S., Tamura,K. and Nei,M. (2004) MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment. *Brief. Bioinform.*, **5**, 150–163.
17. Schloss,P.D. and Handelsman,J. (2005) Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. *App. Environ. Microbiol.*, **71**, 1501–1506.

Downloaded from nar.oxfordjournals.org by guest on February 23, 2011