Journal of Microbiological Methods 75 (2008) 572–575

Contents lists available at ScienceDirect

Journal of Microbiological Methods

journal homepage: www.elsevier.com/locate/jmicmeth

Note

Supplemental programs for enhanced recovery of data from the DOTUR application

Sergio E. Morales [a], Theodore Cosart [b,c], Jesse V. Johnson [b,c], William E. Holben [a,c,*]

[a] Microbial Ecology Program, Division of Biological Sciences, The University of Montana, Missoula, MT, USA
[b] Department of Computer Science, The University of Montana, Missoula, MT, USA
[c] Montana-Ecology of Infectious Diseases Program, The University of Montana, Missoula, MT, USA

ARTICLE INFO

Article history:
Received 16 April 2008
Received in revised form 4 July 2008
Accepted 21 July 2008
Available online 25 July 2008

Keywords:
16S rRNA gene
ARB
DOTUR
Phylogenetic analysis

ABSTRACT

In order to retrieve phylogenetic information from distance matrices generated from large-scale clone libraries, and to explore OTU distribution among them, we have developed downstream applications for use with the already available DOTUR program. These programs enhance and ease data extraction, providing phylogeny to the already generated distance data.

The past ten years have seen an explosion in microbial ecology research based on 16S rRNA gene sequences. Concurrent with that movement has come the development of tools for analyzing ever-increasing data sets of rRNA gene sequences (Cole et al., 2003, 2005, 2007; Desantis et al., 2006; Lozupone and Knight 2005; Wang et al., 2007). Two of the most commonly used tools are DOTUR (Schloss and Handelsman 2005) and ARB (Ludwig et al., 2004). ARB properly aligns sequence data and generates distance matrices that can be transformed by DOTUR into operational taxonomic unit (OTU) composition data: groups of sequences binned together under given similarity parameters. OTU groups are then used to make collector's and rarefaction curves for sampling coverage estimations, richness estimators, and diversity indices. Although the two programs represent powerful tools for data analysis, some information is not easily extracted from DOTUR output files. Rank abundance and OTU distribution files are generated, but these are not linked to sequences, and are not ordered by abundance (in the case of OTU distribution). Sequence identity is left out of the analysis unless individual identification tags (names) from DOTUR are manually linked to their corresponding DNA sequence prior to using some phylogenetic assignment tool. This is a time-intensive task which is often skipped, but which provides critical information regarding OTU bin composition. The ability to use alternative methods to validate group phylogeny depends on being able to track specific OTUs to their sequences. Although manual searching and matching is feasible for small libraries, new studies analyzing thousands of sequences make this task intractable.

Another concept not easily explored is the cohesive organization of OTU placements at different phylogenetic levels and how they relate to tree topology in well-annotated trees. Although programs like ARB allow users to align their own sequences to reference sequences and insert them into annotated trees, the DOTUR program bins them into OTUs using arbitrary cutoffs not linked to any validated hierarchy (Schloss and Handelsman, 2005). To date, multiple 16S rRNA gene based studies have used this approach without rigorously assessing its appropriateness or validity.

To address both of these concerns, we developed two simple programs that are freely available from the authors at: http://dbs.umt.edu/research_labs/holbenlab/links.php. The DAM (*D*OTUR — *A*RB *M*atching) application matches a list of sequence IDs to bins as given by the DOTUR program for some range of DOTUR distance values (Fig. 1). This allows the user to identify the phylogenetic distance at which all sequences within the provided list are grouped as a single OTU. Input for the program (Fig. 1A) includes: (i) A DOTUR list file comprised of rows of space or tab delimited entries, (ii) A list file with one sequence ID on each line, and (iii) A configuration file (not shown). The program was created to match to DOTUR bins a list of sequences that represent a cluster gathered from the output of the ARB program, but any list of sequence IDs can be matched so long as each is present in all the DOTUR list lines and they are in a file with the proper format.

Having stored the bin information in the DOTUR list file, the list of target sequence IDs (TIDs) to be matched to the bins, and a user-specified range of distance values for each distance value in range, DAM first keys each TID to a bin in the DOTUR list by finding it's match in the DOTUR bin information. Each bin that contains at least one of

* Corresponding author. Mailing address: Microbial Ecology Program, Division of Biological Sciences, The University of Montana, Missoula, Montana, 59812-1006, USA. Tel.: +1 406 243 6163; fax: +1 406 243 4184.
E-mail address: bill.holben@mso.umt.edu (W.E. Holben).

0167-7012/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.mimet.2008.07.017

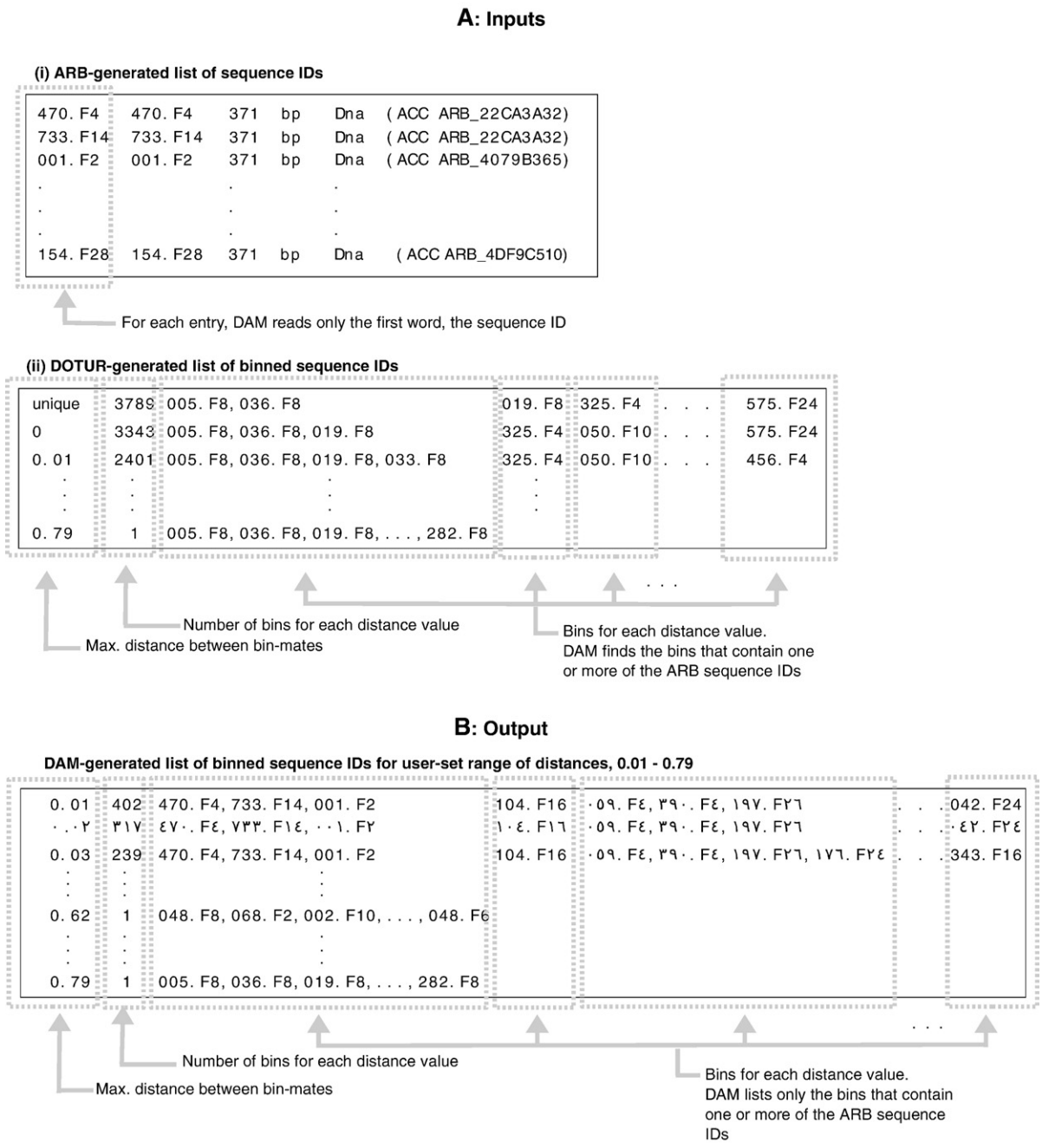

Fig. 1. Examples of the format of input files and the output file from the DAM program. The data shown comes from a 16S rRNA gene sequence survey from an agricultural soil in Michigan. DAM also reads a configuration file (not shown) listing the user-specified settings, an example of which is given with the program's help file.

the TIDs is then copied from the DOTUR file and added to a list of bins for the given distance. The output, then, is a filtered version of the original DOTUR list file, in which bins are listed only for the specified range of distance values, and for each distance only bins containing at least one of the TIDs in the sequence ID list file.

DAM output is a file formatted similarly to a DOTUR list file (Fig. 1B). For each DOTUR distance value, there is a line in the file giving the distance value itself and the set of bins found in the DOTUR file that account for all of the sequences given by the list file, followed by a list of bins and their contents.

The second program, DotMan (*DOT*UR *Man*ipulation Program), creates FASTA files from DOTUR bins (Fig. 2). For a list of DOTUR distance values, it makes one FASTA file for each of the k largest bins listed at each distance. Inputs for the program (Fig. 2A) include: (i) A DOTUR list file comprised of rows of space- or tab-delimited entries, (ii) A file of sequences and their identification tags in FASTA format matching those contained in the DOTUR list, and (iii) A configuration file that provides the program with the names of the above data files, a base name for the FASTA files, a range of distance values over which to create FASTA files, and a list of distance values, each of which is the basis for a series of FASTA files (not shown). Dotman reads, in order, the DOTUR list file, the FASTA file, the user-specified *k* value and then the list of distance values. Then, for each specified distance value, the program first orders the DOTUR bins by population from largest to smallest. For each of the *k* largest bins in the list, it matches each sequence ID in the bin to an entry in the FASTA file, assembling one FASTA file for each bin. In each output file, The ID entry is the sequence ID as given in the DOTUR bin, and the sequence itself is formatted as

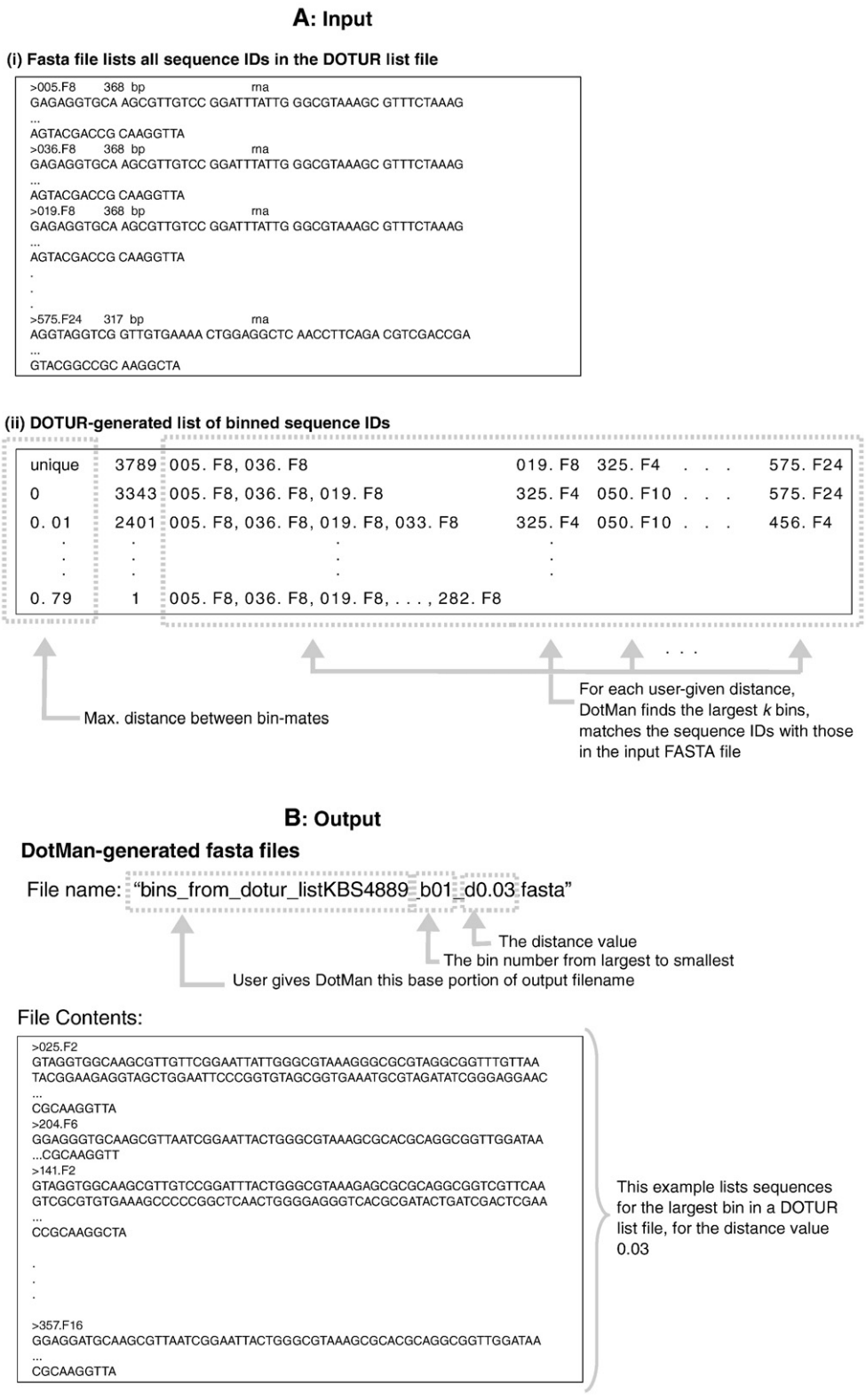

Fig. 2. Examples of the format of DotMan input files and an example FASTA file from its output. These examples come from the same soil study noted in Fig. 1. As with the DAM program, DotMan also reads a configuration file (not shown) with user-specified settings. An example of the configuration file is given in the program's help file.

given in the original FASTA file (Fig. 2B). If the number of bins *n* in the DOTUR list for a given distance value is less than *k,* DotMan writes one file for each of the *n* bins, ordered from largest to smallest. When selecting the i*th* bin of the k largest bins, DOTUR bin order is not necessarily preserved when selecting among bins of equal size. This means, for example, that when requesting the 5 largest bins from a set of 10 bins of equal size, the FASTA files produced will not necessarily represent the 5 leftmost bins as ordered in the DOTUR list file.

Both programs are written in C++ as command-line programs. They are available either as Windows-executables or source-code

packages for compiling on Linux or the MacIntosh using OSX. Recent studies in our lab have used both programs, allowing us to identify specific groups of sequences found to be numerically dominant within our clone library (Morales et al., 2008)and demonstrating that "universal" cutoff values for binning at the phylum level were not observed within the studied site . A sample output file from that study is used in Fig. 1. The figure shows the distance score (similarity score is calculated as 1-distance score) needed to bin all Acidobacteria sequences into a single OTU (62% distance or 38% sequence similarity). These tools should be useful to anyone interested in identifying specific subgroups of OTUs at given taxonomic levels of resolution, or wanting to corroborate OTU bins by way of tools like the RDP Classifier (Wang et al., 2007) or BLAST (Altschul et al., 1990; Tatusova and Madden 1999).

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. Journal of Molecular Biology 215 (3), 403–410.

Cole, J.R., Chai, B., Farris, R.J., Wang, Q., Kulam-Syed-Mohideen, A.S., Mcgarrell, D.M., Bandela, A.M., Cardenas, E., Garrity, G.M., Tiedje, J.M., 2007. The Ribosomal Database Project (RDP-II): Introducing MyRDP Space And Quality Controlled Public Data. Nucleic Acids Research 35 (Suppl_1), D169–D172.

Cole, J.R., Chai, B., Farris, R.J., Wang, Q., Kulam, S.A., Mcgarrell, D.M., Garrity, G.M., Tiedje, J.M., 2005. The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis. Nucleic Acids Research 33 (Suppl_1), D294–D296.

Cole, J.R., Chai, B., Marsh, T.L., Farris, R.J., Wang, Q., Kulam, S.A., Chandra, S., Mcgarrel, D.M., Schmidt, T.M., Garrity, G.M., Tiedje, J.M., 2003. The Ribosomal Database Project (RDP-II): previewing a new autoaligner that allows regular updates and the new prokaryotic taxonomy. Nucleic Acids Research 31 (1), 442–443.

Desantis, T.Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E.L., Keller, K., Huber, T., Dalevi, D., Hu, P., Andersen, G.L., 2006. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Applied And Environmental Microbiology 72 (7), 5069–5072.

Lozupone, C., Knight, R., 2005. Unifrac: a new phylogenetic method for comparing microbial communities. Applied And Environmental Microbiology 71 (12), 8228–8235.

Ludwig, W., Strunk, O., Westram, R., Richter, L., Meier, H., Kumar, Y., Buchner, A., Lai, T., Steppi, S., Jobb, G., Förster, W., Brettske, I., Gerber, S., Ginhart, A.W., Gross, O., Grumann, S., Hermann, S., Jost, R., König, A., Liss, T., Lüßmann, R., May, M., Nonhoff, B., Reichel, B., Strehlow, R., Stamatakis, A., Stuckmann, N., Vilbig, A., Lenke, M., Ludwig, T., Bode, A., Schleifer, K.H., 2004. ARB: A Software Environment For Sequence Data. Nucleic Acids Research 32 (4), 1363–1371.

Morales, S.E., Cosart, T.F., Johnson, J.V., & Holben, W.E., 2008, "Phylogenetic Analysis Of A Complex Bacterial Community Reveals Effects Of Amplicon Length, Degree Of Coverage And DNA Fractionation", In Review.

Schloss, P.D., Handelsman, J., 2005. Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. Applied And Environmental Microbiology 71 (3), 1501–1506.

Tatusova, T.A., Madden, T.L., 1999. BLAST 2 sequences, a new tool for comparing protein and nucleotide sequences. FEMS Microbiology Letters 174, 247–250.

Wang, Q., Garrity, G.M., Tiedje, J.M., Cole, J.R., 2007. Naive Bayesian Classifier For Rapid Assignment of rRNA Sequences Into The New Bacterial Taxonomy. Applied And Environmental Microbiology 73 (16), 5261–5267.