The ISME Journal (2013) 7, 457–460

www.nature.com/ismej

COMMENTARY

Secondary structure improves OTU assignments of 16S rRNA gene sequences

Patrick D Schloss

The ISME Journal (2013) **7,** 457–460; doi:10.1038/ismej.2012.102; published online 27 September 2012

In contrast to the commentary of Wang *et al.* (2011), I contend that the inclusion of secondary structure information provides a faster and more robust analysis when assigning sequences to operational taxonomic units (OTUs). The authors ignored the long history of molecular phylogenetics, where multiple sequence alignments are preferred to pairwise sequence alignments because multiple sequence alignments preserve positional homology across all sequences, not just pairs of sequences (Durbin *et al.*, 1998). Furthermore, profile-based alignments that incorporate the secondary structure of the 16S rRNA molecule are preferred because they provide additional biological information that strengthens the confidence that positional homology is being conserved (Keller *et al.*, 2010). The degree to which these theoretical points are important is contested. I contend that profile-based alignments using curated secondary structure models combined with hierarchical clustering algorithms should be the standard method of assigning 16S rRNA gene sequences to OTUs. I will justify my position by re-emphasizing my previous work, which the Wang commentary ignored, and illustrate several incorrect conclusions that they reached in supporting their thesis (Schloss 2009, 2010, 2011; Schloss and Westcott, 2011).

The Turnbaugh data set that was used in the Wang commentary was not curated using now state-of-the-art methods including denoising, quality trimming or chimera checking (Turnbaugh *et al.*, 2009). Instead, I have chosen data from two regions of the 16S rRNA gene (that is, V13 and V35), which were amplified and sequenced from a DNA extraction obtained from a single stool sample as part of the Human Microbiome Project (Project SRP002397 in the NCBI Short Read Archive). These data were denoised using PyroNoise (Quince *et al.*, 2011) and screened for chimeras using UChime (Edgar *et al.*, 2011) and presented in a previous study (Schloss *et al.*, 2011). In total, there were 779 398 and 560 962 sequences in the V13 and V35 data sets, respectively, after denoising and chimera curation. Once duplicate sequences were removed from the data sets, there were 12 877 and 11 270 unique sequences in the V13 and V35 data sets, respectively.

An additional limitation of the Wang commentary was their selection of alignment algorithms. They compared results from alignments generated using the Needleman–Wunsch, Infernal (inference of RNA alignment) and PARTS (probabilistic alignment for RNA joint secondary structure prediction) algorithms. The choice of PARTS is problematic because it has not been used in the microbial ecology literature and the absence of a NAST (Nearest Alignment Space Termination)-based alignment using either the SILVA or greengenes reference sets is conspicuous as it is a standard method of aligning 16S rRNA sequences (DeSantis *et al.*, 2006b; Schloss, 2009; Caporaso *et al.*, 2010). Finally, they indicate that they used ESPRIT-Tree to carry out their Needleman–Wunsch alignments; however, the ESPRIT-Tree documentation indicates that alignments are carried out using the Gotoh algorithm, which uses an affine gap penalty (http://plaza.ufl.edu/sunyijun/Paper/ESTree_User.pdf). Here, I will utilize the Needleman–Wunsch algorithm (gap opening penalty = 2), Gotoh algorithm (gap opening penalty = 10, gap extension penalty = 0.5; these are same as described in the ESPRIT-Tree documentation) to carry out pairwise alignments (Gotoh, 1982; Needleman and Wunsch, 1970). In addition, I used the mothur-implemented version of the NAST algorithm using greengenes and SILVA reference alignments (DeSantis *et al.*, 2006a; Pruesse *et al.*, 2007; Schloss, 2009) and Infernal as described on the Ribosomal.Database Project website (http://rdp.cme.msu.edu/download/RDPinfernalTraindata.zip) (Cole *et al.*, 2009; Nawrocki *et al.*, 2009).

A significant problem with the pairwise alignment approach is determining how to treat DNA sequences that do not overlap the same region of the gene. This is particularly problematic because PCR primers can amplify spurious templates. For example, I observed 11 V13 and 1 V35 sequences that did not start immediately downstream of the expected priming site. Although they are a relatively small number of sequences, pairwise alignments would not have detected these PCR artifacts, and so they would not be removed from the data sets. To simplify subsequent analyses, I removed these sequences for all comparisons. Obtaining fully overlapping sequences can also be complicated because the number of high-quality bases generated per sequence is rarely uniform. For example, the

sequence lengths in the Turnbaugh data set used in the Wang commentary varied in length between 200 and 317 bp (mean = 232; s.d. = 13). Even when trimming all flowgrams to 450 flows, as has been suggested prior to using PyroNoise (Schloss *et al.*, 2011), I observed reads that varied in length between 250 and 300 bp for the V13 and V35 data sets after removing barcodes and primers. Although all of the sequences within these data sets start at the same alignment position, they vary in where they end. Thus, one is left to determine how to treat sequences of varying length that are identical over the overlapping region. Because the 16S rRNA gene does not evolve uniformly over its length, it is necessary to trim the sequences so that every sequence overlaps the same alignment positions. In this approach, every sequence and base is treated equally. This cannot be done when using pairwise alignments because the variable number of insertions and deletions along different phylogenetic branches makes a purely length-based trimming impossible.

To evaluate the pairwise distances that are calculated with and without trimming sequences to the same alignment coordinates, I created a V13 data set that was trimmed to include *Escherichia. coli* positions 300–516 and a V35 data set that included positions 659–906. These positions were selected to balance the tradeoff between the number of sequences and their length, and resulted in median lengths of 245 and 262 bp, respectively. By determining the regression of distances calculated between non-trimmed sequence alignments as a function of the distances calculated using trimmed sequences aligned using the NAST algorithm with a SILVA-based reference alignment, I could measure the percent increase of the untrimmed alignments relative to the trimmed alignment. The SILVA-based alignment was selected as the reference based on previous observations that it does a better job of preserving Watson–Crick base pairing in variable regions compared to the Ribosomal.Database Project and greengenes reference alignments (Schloss, 2009). Distances calculated using untrimmed sequences aligned by the Needleman–Wunsch, Gotoh and NAST-SILVA algorithms were 11%–20% larger than those calculated using sequences aligned using NAST-SILVA algorithm that were trimmed to a common region in the alignment (Table 1). I also observed that when sequences were trimmed to a common alignment region and then realigned, the distances calculated using Needleman and Gotoh alignments were similar to those using NAST-SILVA and that those calculated using Infernal-Ribosomal. Database Project and NAST-greengenes were larger than the NAST-SILVA distances (Table 1). These are similar to my previous results (Schloss, 2010). These results are illustrative of the variation that can be observed with and without trimming sequences to a common set of alignment positions. As it is not possible to trim sequences to the same alignment positions in a pairwise alignment-based analysis, the remainder of my analysis will use distances calculated with untrimmed sequences for the Needleman and Gotoh alignments and trimmed sequences for the Infernal and NAST-based alignments.

Table 1 Comparison of distances calculated using various alignment algorithms when used with sequences that are trimmed or left untrimmed relative to distances calculated using sequences aligned using the NAST algorithm with the SILVA reference alignment and trimmed to common alignment coordinates

Region	*Alignment method*	*Trimmed to alignment region*	*Slope*	R^2
V13	Needleman	No	1.11	0.95
	Gotoh	No	1.16	0.90
	NAST-SILVA	No	1.13	0.95
	NAST-greengenes	Yes	1.10	0.93
	Infernal-RDP	Yes	1.05	0.92
	Needleman	Yes	0.99	0.99
	Gotoh	Yes	1.01	0.98
V35	Needleman	No	1.17	0.95
	Gotoh	No	1.18	0.94
	NAST-SILVA	No	1.20	0.96
	NAST-greengenes	Yes	1.34	0.95
	Infernal-RDP	Yes	1.01	0.97
	Needleman	Yes	0.99	0.99
	Gotoh	Yes	1.01	0.98

Abbreviations: Infernal, inference of RNA alignment; NAST, Nearest Alignment Space Termination; RDP, Ribosomal.Database Project.

As the Wang commentary correctly describes, it is difficult to assess OTU assignment accuracy. They chose to use an approach that utilizes sequence alignments and taxonomic information to map sequences to the references species. This is problematic as it is widely known that taxonomic levels do not correlate well with individual distance thresholds based on 16S rRNA gene sequences (Schloss and Westcott, 2011). I previously implemented a database-independent method of measuring the quality of sequence assignment to OTUs at a 3% distance threshold (Schloss and Westcott, 2011). In this approach the OTU assignments generated using the trimmed NAST-SILVA alignments were assumed to represent the truth because of the reasons I have outlined thus far in my commentary. I then identified pairs of sequences that had distances less than 3% and co-occurred in the same OTU (that is, true positives) or were found in different OTUs (that is, false negatives). For those pairs of sequences with distances greater than 3%, some were found in separate OTUs (that is, true negatives) and others co-occurred in the same OTU (that is, false negatives). As it is impossible to have false positives or false negatives, I used the Matthew's correlation coefficient (MCC) to synthesize these four parameters (Baldi *et al.*, 2000). As shown in Table 2, the NAST and Infernal algorithms performed better than the pairwise alignment algorithms as judged by the MCC. An alternative approach is USearch, which is a fast alignment-independent heuristic (Edgar, 2010). When I used USearch to assign sequences to OTUs as described in the user manual

Table 2 Quality of OTU assignments for various alignment methods using data collected from the V13 and V35 regions of the 16S rRNA gene when using the average neighbor clustering algorithm to assign sequences at a distance threshold of 0.03

Region	*Alignment algorithm*	*True positives*	*True negatives*	*False positives*	*False negatives*	*Matthew's correlation coefficient*	*Number of OTUs*
V13	NAST-SILVA	11 788 844	69 880 006	717 908	373 787	0.948	248
	NAST-greengenes	10 903 337	70 035 748	562 166	1 259 294	0.911	279
	Needleman-Wunsch	10 852 102	70 063 205	534 709	1 310 529	0.909	328
	Infernal-RDP	10 781 897	69 956 380	641 534	1 380 734	0.901	287
	Gotoh	10 556 961	70 140 027	457 887	1 605 670	0.898	469
	USearch	4 792 232	70 348 100	249 814	7 370 399	0.664	287
V35	NAST-SILVA	12 262 540	49 374 768	1 752 017	100 221	0.914	202
	NAST-greengenes	11 845 329	49 773 304	1 353 481	517 432	0.909	240
	Infernal-RDP	12 097 732	49 373 547	1 753 238	265 029	0.905	217
	Needleman–Wunsch	11 097 695	50 056 419	1 070 366	1 265 066	0.882	267
	Gotoh	10 985 871	50 134 673	992 112	1 376 890	0.880	345
	USearch	7 994 796	49 990 967	165 908	5 337 875	0.726	255

Abbreviations: Infernal, inference of RNA alignment; NAST, Nearest Alignment Space Termination; RDP, Ribosomal.Database Project. The algorithms are sorted in descending order by their Matthew's correlation coefficients. There were 12 866 and 11 269 V13 and V35 sequences, respectively.

(http://drive5.com/usearch/UsearchUserGuide5.1.pdf), its MCC was substantially lower than the other methods and had a substantially higher false-negative rate. Comparing the NAST-SILVA to Needleman–Wunsch OTU assignments, the NAST-SILVA assignments had more true positives (V13: 1.09-fold, V35: 1.10-fold), comparable true negatives, more false positives (V13: 1.12-fold, V35: 1.64-fold), and fewer false negatives (V13: 351-fold, V35: 1260-fold). For both the V13 and V35 regions, this translated into observing 32% more OTUs in the Needleman–Wunsch. These results contrast with those of the Wang commentary, which indicated that alignments that incorporated secondary structure information generated lower-quality OTUs relative to those from pairwise alignments.

The Wang commentary also commented that pairwise alignment algorithms were significantly faster than methods that incorporate the secondary structure of the 16S rRNA molecule. Although this is true for PARTS, it is not true for the Infernal or NAST algorithms. Making all pairwise alignments and calculating their pairwise distances requires an amount of time proportional to N^2L^2, where N is the number of sequences and L is their length. In contrast, the time requirements for generating alignments using Infernal or NAST are proportional to NL^2. The profile-based alignments that incorporate the gene's secondary structure were more than 45-fold faster than either of the pairwise alignment methods (Table 3). Even if pairwise alignments generated OTU assignments that were as good as those that incorporate the secondary structure, it is clear that they are prohibitively slower.

Finally, the Wang commentary states that 'A first and crucial step… is the binning of 16S sequences'. In fact, the first step should involve sequence trimming, denoising and chimera removal. I have found that a significant component to the perceived difficulty in assigning sequences to OTUs is the increased sequencing error rates at the distal end of sequence reads (Schloss *et al.*, 2011). One might assume that since a sequencing platform provides 317 base calls for a sequence that all 317 of those base calls are equally good. In fact, the distal end of every sequence has diminished quality. The end result is that these sequencing errors create a scenario where there are an artificially inflated number of unique sequences. This then increases the amount of time required to perform alignments, distance calculations and OTU assignments. Once sequences are trimmed, denoised and evaluated for chimeras, the number of unique sequences and spurious OTUs are dramatically decreased (Schloss *et al.*, 2011).

Table 3 Number of minutes required to align sequences and calculate distances between sequences using various algorithms

Alignment algorithm	*V13*	*V35*
NAST-greengenes	17.4	10.3
NAST-SILVA	19.2	15.7
Infernal-RDP	27.6	24.4
Needleman–Wunsch	1321.1	1081.5
Gotoh	1516.1	1221.2

Abbreviations: Infernal, inference of RNA alignment; NAST, Nearest Alignment Space Termination; RDP, Ribosomal.Database Project. All times are based on using a single processor although all algorithms can be parallelized. The alignment algorithms are sorted according to the number of minutes required.

It should be noted that the various algorithms will perform differently when used with different data sets, and different regions will perform differently than the human stool sample described here. Therefore, on the theoretical grounds of preserving positional homology, it does not make sense to

preferentially choose one of the alternative alignment algorithms instead. Yet even if the outputs of these various algorithms were the same, the profile-based algorithms that incorporate the secondary structure of the 16S rRNA gene are significantly faster than those that do not.

Acknowledgements

PDS is supported by grants from the National Institutes for Health (R01HG005975 and P30DK034933).

PD Schloss is at Department of Microbiology and Immunology, University of Michigan, Ann Arbor, MI, USA
E-mail: pschloss@umich.edu

References

Baldi P, Brunak S, Chauvin Y, Andersen CA, Nielsen H. (2000). Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics* **16**: 412–424.

Caporaso JG, Bittinger K, Bushman FD, DeSantis TZ, Andersen GL, Knight R. (2010). PyNAST: a flexible tool for aligning sequences to a template alignment. *Bioinformatics* **26**: 266–267.

Cole JR, Wang Q, Cardenas E, Fish J, Chai B, Farris RJ *et al.* (2009). The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res* **37**: D141–D145.

DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K *et al.* (2006a). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol* **72**: 5069–5072.

DeSantis TZ Jr., Hugenholtz P, Keller K, Brodie EL, Larsen N, Piceno YM *et al.* (2006b). NAST: a multiple sequence alignment server for comparative analysis of 16S rRNA genes. *Nucleic Acids Res* **34**: W394–W399.

Durbin R, Eddy SR, Krogh A, Mitchison G. (1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press: Cambridge, UK, New York.

Edgar RC. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**: 2460–2461.

Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R. (2011). UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* **27**: 2194–2200.

Gotoh O. (1982). An improved algorithm for matching biological sequences. *J Mol Biol* **162**: 705–708.

Keller A, Forster F, Muller T, Dandekar T, Schultz J, Wolf M. (2010). Including RNA secondary structures improves accuracy and robustness in reconstruction of phylogenetic trees. *Biol Direct* **5**: 4.

Nawrocki EP, Kolbe DL, Eddy SR. (2009). Infernal 1.0: inference of RNA alignments. *Bioinformatics* **25**: 1335–1337.

Needleman SB, Wunsch CD. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* **48**: 443–453.

Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J *et al.* (2007). SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res* **35**: 7188–7196.

Schloss PD. (2009). A high-throughput DNA sequence aligner for microbial ecology studies. *PLoS ONE* **4**: e8230.

Schloss PD. (2010). The effects of alignment quality, distance calculation method, sequence filtering, and region on the analysis of 16S rRNA gene-based studies. *PLoS Comput Biol* **6**: e1000844.

Schloss PD, Gevers D, Westcott SL. (2011). Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. *PLoS One* **6**: e27310.

Schloss PD, Westcott SL. (2011). Assessing and improving methods used in operational taxonomic unit-based approaches for 16S rRNA gene sequence analysis. *Appl Environ Microbiol* **77**: 3219–3226.

Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE *et al.* (2009). A core gut microbiome in obese and lean twins. *Nature* **457**: 480–484.

Quince C, Lanzen A, Davenport RJ, Turnbaugh PJ. (2011). Removing noise from pyrosequenced amplicons. *BMC Bioinformatics* **12**: 38.

Wang X, Cai Y, Sun Y, Knight R, Mai V. (2011). Secondary structure information does not improve OTU assignment for partial 16s rRNA sequences. *ISME J* **6**: 1277–1280.