The ISME Journal (2012) 6, 1277–1280

www.nature.com/ismej

COMMENTARY

Secondary structure information does not improve OTU assignment for partial 16s rRNA sequences

Xiaoyu Wang, Yunpeng Cai, Yijun Sun, Rob Knight and Volker Mai

The ISME Journal (2012) **6,** 1277–1280; doi:10.1038/ismej.2011.187; published online 15 December 2011

We would like to refute the suggestion that incorporating secondary structure information improves taxonomy-independent (TI; see Table 1 for list of abbreviations) 16S rRNA binning methods for partial 16S rRNA sequence reads. Although 16S rRNA secondary structure is crucial for its proper function, it is not clear that nucleotide differences at positions important for secondary structure contain unique phylogenetic information that requires different weighting than nucleotide differences at other positions. We support our position, and show the large computational expense required to include secondary structure information, using data from a simulation study. Our findings are important for microbial ecologists, as they confirm that an existing algorithm, ESPRIT-Tree (Cai and Sun, 2011), performs equally or better without consideration of secondary structure. Consequently, microbial ecologists who are already struggling with computational limitations analyzing vast data sets need not suspect that they are missing important improvements in operational taxonomic units (OTU) binning that could be achieved only at much greater computational cost, and can analyze data sets with many millions of sequences with confidence.

Table 1 Abbreviations used in the text

Abbreviation	*Definition*
DOTUR	A software package, mainly used for clustering sequences
ESPRIT	A hierarchical clustering algorithm, mainly used for clustering sequences
ESPRIT-Tree	The fast version of ESPRIT
INFERNAL	A sequence alignment algorithm considering secondary structures
MOTHUR	An improved version of DOTUR for analyzing sequences
MSA	Multiple sequence alignment
NMI	Normalized mutual information. A metric to evaluate clustering performances
NW	Needleman–Wunsch algorithm, a widely used global sequence alignment algorithm
OTU	Operational taxonomic units
PARTS	A sequence alignment algorithm considering secondary structures
PSA	Pairwise sequence alignment
TaxCollector	A set of Python scripts that attaches taxonomic information to sequences in a database
TD	Taxonomy-dependent (clustering method)
TI	Taxonomy-independent (clustering method)

Recent advances in high-throughput sequencing technologies have contributed to an explosion in sequence data from studies of the microbial diversity in various environments. The analysis of complex microbial communities frequently includes large-scale sequencing of 16S rRNA, to estimate species composition and diversity in many environments. Although sequencing technologies allow ever deeper interrogation of microbial communities, the availability of efficient bioinformatics tools that handle such vast data sets has become a bottleneck.

A first and crucial step towards microbial community analyses is the binning of 16S sequences into groups that contain sequences with a predetermined degree of similarity. In taxonomy-dependent methods, query sequences are compared at a predetermined similarity level with known sequences deposited in an annotated database (e.g., RDP and Greengenes). In contrast, TI methods assign query sequences into a set of OTUs by applying clustering algorithms to pairwise distances of sequences using specified distance thresholds (Sun *et al.*, 2011; Schloss and Westcott, 2011). A crucial advantage of TI methods is their independence from the completeness of existing databases, which allows for the inclusion of novel sequences that often represent a significant proportion of sequence data.

TI methods generally consist of two major components: (1) aligning sequences to build a pairwise distance matrix and (2) performing clustering analysis to group sequences into OTUs. Currently used methods include hierarchical clustering algorithms such as DOTUR, MOTHUR, ESPRIT, ESPRIT-Tree and greedy algorithms (Sun *et al.*, 2011; Schloss and Westcott, 2011).

Multiple sequence alignment (MSA) and pairwise sequence alignment (PSA) have been applied in TI methods. MSA is attractive for two reasons; first,

each sequence needs to be compared with a reference alignment only once, avoiding the need to compare all pairs of sequences, and second, properties of the reference alignment such as secondary structure can be used to improve the multiple alignment. However, several recent studies show that PSA leads to a much more accurate estimate than MSA (Huse *et al.*, 2010; Sun *et al.*, 2011). A frequent criticism of using the classic Needleman–Wunsch (NW) alignment method in PSA is that NW ignores the secondary structure information. However, the assumption that incorporating secondary structure information indeed improves TI methods and subsequent OTU assignment has never been rigorously tested. Thus, we applied three TI methods, two of them including secondary structure information, to compare their performances on a test data set based on partial 16S rRNA sequences. The sequences had an average length of 231 nucleotides (range 200–317), covered the V2 region and were previously generated to study the association between obesity and the composition of human gut microbiota (Turnbaugh *et al.*, 2009). We used the NW algorithm, one of the most commonly used PSA methods that does not consider secondary structure information, and two secondary structure-based alignment algorithms (i) PARTS (Harmanci *et al.*, 2008), an accurate but computationally expensive PSA algorithm, and (ii) INFERNAL (Nawrocki *et al.*, 2009), a MSA algorithm widely applied in many studies and incorporated into many workflows.

As noted previously (Sun *et al.*, 2011), one major obstacle to comparing OTU assignment methods is the lack of ground truth information for performance evaluation. To overcome this difficulty, we constructed a reference database from the RDP-II database (Cole *et al.*, 2005) using TaxCollector (Giongo *et al.*, 2010), so that each reference sequence was fully annotated. We then ran a MegaBlast search of the gut data against the reference database, and used a stringent criterion (>97% identity over an aligned region >97% of the total length of the sequences) to retain the annotated sequences, which resulted in a total of about 750 000 reads that can be confidently classified into 671 species. Due to large computational costs associated with PARTS, we generated 10 test subsets, each containing 500 sequences selected from the order Clostridiales in the TaxCollector annotated data set. Selecting sequences from within an order increases the difficulty of OTU picking, which is required for testing the performance of the algorithms. We applied NW, PARTS and INFERNAL (using their default settings) to align the sequences in each data set and to produce distance matrices. The commonly used normalized mutual information (NMI) metric (Strehl and Ghosh, 2002; Sun *et al.*, 2011) was applied to evaluate how the clustering outcomes of the three competing approaches agreed with the ground truth.

The pairwise sequence distances computed by NW, PARTS and INFERNAL are shown in Figure 1. We removed all pairs of sequences with NW distances greater than 0.15, resulting in a total of 580 295 sequence pairs to compute pairwise distances. As expected, all distances generated by PARTS and INFERNAL were greater than those from NW, because by definition, NW is the optimal pairwise sequence alignment method and yields the minimum genetic distance for a given sequence pair. From the scatter plot, it can be seen that the variance of PARTS distances is proportional to the NW distances, whereas the variance of INFERNAL distances appears independent of NW distances. The r^2 between NW distances and PARTS distances is 0.95, indicating a high linear association between distances from NW and PARTS, whereas r^2 is 0.79 between NW and INFERNAL distances.

The pairwise distances were then fed into ESPRIT using the average linkage mode to group the sequences into OTUs at various distance levels (0.01–0.15 incremented by 0.01), and corresponding clustering performance was evaluated by NMI

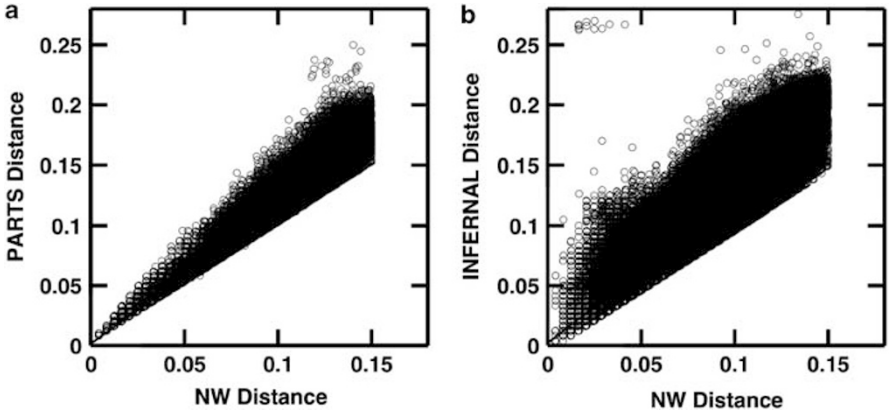

Figure 1 Scatter plot of pairwise distances. (**a**) pairwise distances generated by NW and PARTS ($r^2 = 0.9497$). The data is fitted by a line ($y = 1.0995x + 0.0018$) using linear least squares method. (**b**) Scatter plot of pairwise distances generated by NW and INFERNAL ($r^2 = 0.7943$). The data is fitted by a line ($y = 1.2066x + 0.0021$) using linear least squares method.

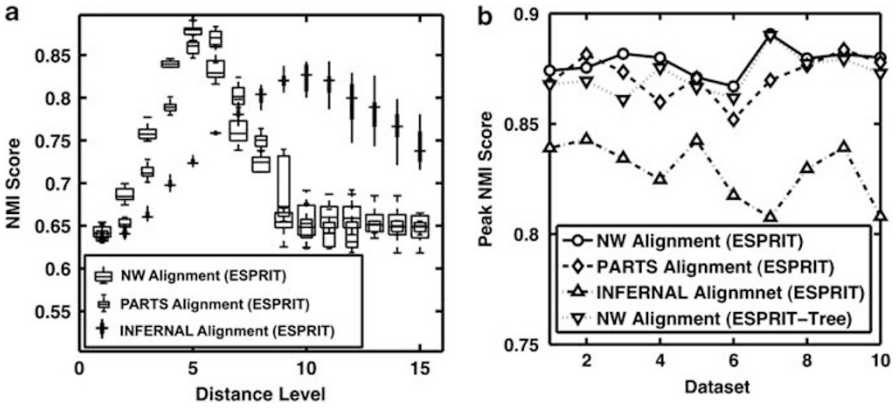

Figure 2 NMI scores of three approaches. (**a**) Boxplots of NMI scores. The NW-based results are shown in wide boxes, the PARTS-based results are shown in narrow boxes, and the INFERNAL-based results are shown in filled boxes. (**b**) Peak NMI scores generated by ESPRIT, PARTS and ESPRIT-Tree for 10 test data sets.

scores. Figure 2a presents the NMI scores of NW-based, PARTS-based and INFERNAL-based approaches at 15 distance levels, over the 10 data sets. The three approaches performed differently at various distance levels. Overall, the profiles of NMI scores from NW-based and PARTS-based approaches are similar. In contrast, the peak score of INFERNAL-based approach shifts to the right of the other two, which implies that as a MSA, INFERNAL tends to overestimate pairwise distances and that its best clustering performance occurs at a larger distance level.

We compared the peak NMI scores of the three approaches, which by definition correspond to the best clustering results that these methods can achieve for each test data set (Figure 2b). The NW-based ESPRIT-Tree clustering results were also reported for reference. For each test data set, NMI scores were computed by each method at various distance levels (0.01–0.15), and the peak score was recorded to evaluate the method's performance on the data sets. The mean value of peak NMI scores over the 10 test data sets was computed using each method and the results were 0.88 for NW-based ESPRIT, 0.87 for NW-based ESPRIT-Tree, 0.87 for PARTS, and 0.83 for INFERNAL, which indicates on average NW-based ESPRIT outperforms others and INFERNAL generates the worst results. To further illustrate how significant those results are different from each other, two-tailed *t*-tests were performed comparing the mean value of peak scores computed by NW-based ESPIRT with those by other methods. We observe that at significance level 0.05, NW-based ESPRIT performed similarly to NW-based ESPRIT-Tree ($P=0.10$) and PARTS-based ESPRIT ($P=0.08$), whereas it performed significantly better than INFERNAL ($P<10^{-8}$). All of the results suggest that the secondary structure information incorporated in PARTS and INFERNAL does not significantly contribute to the clustering accuracy and may even decrease it.

Although the results of the PARTS alignment are similar to those of NW alignments, their computational costs differ greatly. NW required only 0.16 CPU hours on a desktop computer (2.8 GHz) to align the sequences of the 10 test data sets, but PARTS required a total of 58 900 CPU hours (on a cluster of nodes each with six-core Opteron 4184 processors (AMD, Pasadena, CA, USA) at 2.8 GHz).

Our argument, that incorporating secondary structure information into 16S rRNA sequence alignments does not improve the performance of TI methods, is supported by the simulation study. An additional simulation study performed on a 16S rRNA data set against the V9 region from a soil sample confirmed this observation (data not shown). The use of multiple sequence alignment techniques that include secondary structure actually appears to diminish performance and increase computational costs. Secondary structure is crucial for the biological function of ribosomal RNA and helpful for constructing valid phylogenetic trees of distantly related organisms. However, for assigning OTUs in large 16S rRNA data sets, adding secondary structure information currently appears to be of little utility. Our findings should reassure microbial ecologists that binning algorithms already exist that, at least for partial 16S rRNA sequence reads, perform better than existing secondary structure-based algorithms, which, in any case, are too inefficient for scaling to the analysis of data sets containing many millions of sequences. Microbial ecologists need to be aware that another issue exists with current binning algorithms for 16S rRNA sequences in large data sets, particularly the high frequency of multiple OTU's that contain sequences of high similarity. Rather than focusing on secondary structure, addressing this other issue will allow microbial ecologists more accurate insight into the true diversity of complex communities.

Acknowledgements

YS and VM are supported in part by a grant from NSF (#DBI-1062362). RK is supported in part by the Howard Hughes Medical Institute and the National Institute of Health (HG4872, HG4866, DK78669).

X Wang is at Department of Microbiology and Cell Science, University of Florida, Gainesville, FL, USA; X Wang is also at Emerging Pathogens Institute, University of Florida, Gainesville, FL, USA; Y Cai is at Reseach Center for Biomedical Information Technology, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China; Y Sun is at Interdisciplinary Center for Biotechnology Research, University of Florida, Gainesville, FL, USA; Y Sun is also at Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL, USA; R Knight is at Department of Chemistry and Biochemistry, Univesity of Colorado, Boulder, CO, USA; R Knight is also at Howard Hughes Medical Institute, Boulder, CO, USA; V Mai is at Department of Microbiology and Cell Science, University of Florida, Gainesville, FL, USA and V Mai is also at Emerging Pathogens Institute, University of Florida, Gainesville, FL, USA
E-mail: vmai@ufl.edu

References

Cai Y, Sun Y. (2011). ESPRIT-Tree: hierarchical clustering analysis of millions of 16S rRNA pyrosequences in quasilinear computational time. *Nucleic Acids Res*, 2011 **39**: e95.

Cole JR, Chai B, Farris BJ, Wang Q, Kulam SA, McGarrell DM *et al.* (2005). The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis. *Nucleic Acids Res* **33**: 294–296.

Giongo A, Richardson AGD, Crabb DB, Triplett EW. (2010). TaxCollector: modifying current 16S rRNA databases for the rapid classification at six taxonomic levels. *Diversity* **2**: 1015–1025.

Harmanci AO, Sharma G, Mathews DH. (2008). PARTS: Probabilistic Alignment for RNA joinT Secondary structure prediction. *Nucleic Acids Res* **36**: 2406–2417.

Huse SM, Welch DM, Morrison HG, Sogin ML. (2010). Ironing out the wrinkles in the rare biosphere through improved OTU clustering. *Environ Microbiol* **12**: 1889–1898.

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

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.

Strehl A, Ghosh J. (2002). Cluster ensembles - a knowledge reuse framework for combining multiple partitions. *J Mach Learn Res* **3**: 583–617.

Sun Y, Cai Y, Huse SM, Knight R, Farmerie WG, Wang X *et al.* (2011). A large-scale benchmark study of existing algorithms for taxonomy-independent microbial community analysis. *Brief Bioinform*; e-pub ahead of print 27 April 2011; doi: 10.1093/bib/bbr009.

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.