Available online at www.sciencedirect.com
ScienceDirect
Journal of Microbiological Methods 72 (2008) 166–179
Journal of Microbiological Methods
www.elsevier.com/locate/jmicmeth

A simple binomial test for estimating sequencing errors in public repository 16S rRNA sequences

Young-Gun Zo [a,b], Rita R. Colwell [a,b,c,*]

[a] Center of Marine Biotechnology, University of Maryland Biotechnology Institute, 701 E. Pratt St., Baltimore, MD 21202, United States
[b] Center for Bioinformatics and Computational Biology, University of Maryland Institute of Advanced Computer Studies, University of Maryland College Park, College Park, MD 20742, United States
[c] Johns Hopkins University Bloomberg School of Public Health, 615 N. Wolfe St, Baltimore, MD 21205, United States

Received 30 May 2007; accepted 13 November 2007
Available online 23 November 2007

Abstract

Sequences in public databases may contain a number of sequencing errors. A double binomial model describing the distribution of indel-excluded similarity coefficients (*S*) among repeatedly sequenced 16S rRNA was previously developed and it produced a confidence interval of *S* useful for testing sequence identity among sequences of 400-bp length. We characterized patterns in sequencing errors found in nearly complete 16S rRNA sequences of *Vibrionaceae* as highly variable in reported sequence length and containing a small number of indels. To accommodate these characteristics, a simple binomial model for distribution of the similarity coefficient (*H*) that included indels was derived from the double binomial model for *S*. The model showed good fit to empirical data. By using either a pre-determined or bootstrapping estimated standard probability of base matching, we were able to use the exact binomial test to determine the relative level of sequencing error for a given pair of duplicated sequences. A limitation of the method is the requirement that duplicated sequences for the same template sequence be paired, but this can be overcome by using only conserved regions of 16S rRNA sequences and pairing a given sequence with its highest scoring BLAST search hit from the nr database of GenBank.

Keywords: 16S rRNA; Binomial model; Sequencing error; Sequence similarity coefficient; SSU rRNA

1. Introduction

Contemporary bacterial systematics, especially as employed in microbial ecology, depends on 16S rRNA gene nucleotide sequence polymorphism of reference strains (Clarridge, 2004). Typically, sequences for type strains are employed as reference against which other sequences are compared. Accuracy of reference nucleotide sequences of genes employed in comparative analyses is, therefore, essential for accurate identification and classification of bacteria. However, it is well known that publicly available 16S rRNA sequence data for bacteria may contain various levels of error. In a recent report, one out of twenty 16S rRNA sequence entries in public repositories was found to contain substantial sequence anomalies (Ashelford et al., 2005). Therefore, use of reference sequences in comparative analyses should be preceded by an examination of their accuracy.

Most common and problematic are chimerical sequences (Ashelford et al., 2005), typically formed during PCR reactions carried out to amplify 16S rRNA genes of a mixed population. Because an increasing number of 16S rRNA sequences in public repositories derive from metagenomic surveys of mixed populations of natural bacterial communities, recognizing and avoiding use of chimerical sequences in repositories is an important quality control for both depositing and using sequences obtained from mixed populations. Fortunately, computational tools are available to detect 16S rRNA chimera sequences (Ashelford et al., 2005).

* Corresponding author. Center for Bioinformatics and Computational Biology, University of Maryland Institute of Advanced Computer Studies, University of Maryland College Park, College Park, MD 20742, United States. Tel.: +1 301 405 9550; fax: +1 301 314 6654.
E-mail address: rcolwell@umiacs.umd.edu (R.R. Colwell).

0167-7012/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.mimet.2007.11.013

Large volumes of 16S rRNA sequence data are being generated in metagenomic studies with automatic base-calling from dye-fluorescence trace results generated by automatic sequencing instruments. Miscall of bases is a significant source of sequencing error in databases. Although quality control of base-calling can be achieved using PQ values of the Phred/Phrap Package (Ewing and Green, 1998) or other algorithms implemented in various manufacturers' base-calling software, the results of quality assessments are seldom explicitly reported in sequence records or in publications reporting sequences.

Recently, Fields et al. (2006) assessed the level of sequencing errors associated with sequences obtained from clone libraries of partial 16S rRNA fragments generated by PCR on V2-V6 domains of mixed populations. By fitting a double binomial model to the results of repeated sequencing of identical clones, they demonstrated that identical sequences generate a predictable level of sequence polymorphism due to random sequencing errors. Their empirically calibrated model showed that sequencing error could generate a measurable amount of base differences, i.e., <0.54 base difference per 100 bp length at the 95% confidence level or <0.81 base difference per 100 bp length at the 99% confidence level of sequence identity. From their results, they concluded that this level of error is small enough that the data could be used for identification of a bacterial species because a criterion of species identification, i.e., 97% identity, allows up to 3 base differences per 100 bp length. However, the estimated sequencing error is not applicable to a significant portion of the 16S rRNA sequences in public repositories because it varies with the materials and methods used (Clarridge, 2004).

A possible trend is that sequencing error may be decreasing, as new sequencing methods are developed. In other words, the earlier a sequence had been deposited, the larger the probability of error. For bacterial systematics, this trend suggests a problem because many of the 16S rRNA sequences of type strains of those bacterial species whose reference sequences are used for classification of unknown bacteria were deposited well before current methods were available. Of course, one way to avoid this problem is to re-sequence the entire, or a portion of, the 16S rRNA genes as a batch process under quality controlled conditions and, thereafter, maintaining a quality-controlled database for taxonomic analysis of a given group of bacteria (Harmsen et al., 2003). Such an approach is expensive, so it is preferable to have a mechanism for detecting sequencing error in deposited sequences to keep the level of error low enough that the data can be used for identification and classification of bacteria.

In this report, we present a method for testing the level of error in 16S rRNA sequence data deposited in public databases that may arise from various methods and/or materials used to generate the sequences. The method employs an exact binomial test, comparing the error level of a sequence to a given standard. We derived a simple binomial model from the double binomial model of Fields et al. (2006) by calibrating models to the 16S rRNA sequences for type strains of *Vibrionaceae*. We were able to apply the method to produce a set of quality-controlled reference sequences that can be used for classification of the *Vibrionaceae*, a bacterial family whose intra-family taxonomic relationships are rapidly changing because of the deluge of data arising from large scale expeditions and programs (Venter et al., 2004) and the description of new *Vibrio* species (Thompson et al., 2004).

2. Materials and methods

2.1. Sequence acquisition and analysis

2.1.1. Acquisition from public repositories

Vibrionaceae 16S rRNA sequences were downloaded from GenBank (Benson et al., 2005), RNA Database Project (RDP) (Cole et al., 2005), and European Ribosomal RNA Database (ERDB) (Wuyts et al., 2002) by selecting all sequences with the genus names *Enterovibrio*, *Grimontia*, *Photobacterium*, *Salinivibrio*, and *Vibrio* via taxonomic browsers or their equivalent. Various designations of type strains for each species of *Vibrionaceae* were compiled from Euzeby's List of Bacterial Names with Standing in Nomenclature (Euzeby, 1997) and catalogues of the American Type Culture Collection (ATCC), Collection de l'Institut Pasteur (CIP), the Deutsche Sammlung von Mikroorganismen und Zellkulturen GmbH (DSMZ), The Belgian Co-ordinated Collections of Micro-organisms/Laboratorium voor Microbiologie, Universiteit Gent (BCCM/LMG), and The National Collection of Industrial, Marine and Food Bacteria (NCIMB). By searching the compiled type strain designations in the entry name, title, organism, and strain fields of GenBank, RDP and ERRD records, all available 16S rRNA sequences for the strains were collected. From the sequence collections, entries with >1300-bp length were selected for analysis. Throughout this report, each sequence is referenced by its GenBank Accession Number (ACCN). When it was necessary to obtain sequences with high similarity to a given sequence under study, a BLASTN search (Altschul et al., 1997) was done using BLASTCL3.EXE on the GenBank nr database, with 'bacteria' serving as the filter for the organism field.

2.1.2. Sequencing

Assessment of the level of sequencing error resulting from methods used in our own laboratory was done by sequencing 16S rRNA genes of two *V. cholerae* strains, RC2 and RC782. RC2 is the laboratory number for *V. cholerae* ATCC 14035^T, and RC782 for *V. cholerae* ATCC 14547, the latter being the former type strain of *V. albensis*, now recognized as the basonym designated for bioluminescent *V. cholerae*-like strains. The RC2 culture used for sequencing has been maintained in our laboratory for more than twenty years and the RC782 culture was purchased from the American Type Culture Collection, Virginia, in 2003.

Sequencing the PCR products of genes was done using universal primers, 5'-AGA GTT TGA TCM TGG CTC AG-3' for forward direction and 5'-CGG YTA CCT TGT TAC GAC TT-3' for reverse direction. The PCR products were cloned into the pCR4TOPO sequencing vector, using the TOPO-TA cloning kit (Invitrogen, Carlsbad, CA). BigDye termination sequencing of purified plasmid vector was done using the ABI