Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Microbiological Methods 65 (2006) 144–152
Journal of Microbiological Methods
www.elsevier.com/locate/jmicmeth

Confidence intervals of similarity values determined for cloned SSU *rRNA* genes from environmental samples

M.W. Fields [a], J.C. Schryver [b], C.C. Brandt [c], T. Yan [c], J.Z. Zhou [c], A.V. Palumbo [c,*]

[a] *Department of Microbiology, Miami University, Oxford, Ohio 45056, United States*
[b] *Computational Sciences and Engineering Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, United States*
[c] *Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, United States*

Received 30 July 2003; received in revised form 1 July 2005; accepted 1 July 2005
Available online 3 August 2005

Abstract

The goal of this research was to investigate the influence of the error rate of sequence determination on the differentiation of cloned SSU rRNA gene sequences for assessment of community structure. SSU rRNA cloned sequences from groundwater samples that represent different bacterial divisions were sequenced multiple times with the same sequencing primer. From comparison of sequence alignments with unedited data, confidence intervals were obtained from both a 'double binomial' model of sequence comparison and by non-parametric methods. The results indicated that similarity values below 0.9946 are likely derived from dissimilar sequences at a confidence level of 0.95, and not sequencing errors. The results confirmed that screening by direct sequence determination could be reliably used to differentiate at the species level. However, given sequencing errors comparable to those seen in this study, sequences with similarities above 0.9946 should be treated as the same sequence if a 95% confidence is desired.

Keywords: Clone; Environmental; Error analysis; *SSU rRNA* genes; Sequencing

1. Introduction

Our current understanding of the microbial world suggests that in many environments the vast majority of species (>99.9%) are uncultivable with standard microbiological techniques (Pace, 1997). Sequencing of the small subunit ribosomal RNA (SSU *rRNA*) genes has provided new insights into the extent of microbial diversity (Woese and Fox, 1977; Woese et al., 1990). Current diversity estimates range from 350 to 500,000 operational taxonomic units (OTUs) per gram of soil (Øvreås, 2000; Dykhuizen, 1998), and most likely 100 OTUs per milliliter of seawater (Curtis et al., 2002). Sequence variation in the SSU *rRNA* genes not only provides an estimate of the

* Corresponding author. Oak Ridge National Laboratory, Environmental Sciences Division, P.O. Box 2008, MS-6038, Oak Ridge, TN 37831-6038, United States. Tel.: +1 865 576 8002; fax: +1 865 576 8646.
E-mail address: palumboav@ornl.gov (A.V. Palumbo).

0167-7012/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.mimet.2005.07.001

variety of microorganisms in a given sample (differentiation at the species level), but can also provide information about the phylogenetic composition of the sample.

Numerous published studies employ SSU *rRNA* gene clonal libraries to estimate microbial diversity. Two techniques commonly used in such studies are restriction fragment length polymorphism analysis and density gradient gel electrophoresis. In both techniques, unique clones are identified and sequenced, which can be laborious and time-consuming. However, clones can be randomly selected and screened by direct sequence determination resulting in a significant savings in time and effort. The resulting sequences are then compared pairwise and a similarity index is calculated. The convention of 97% to 98% SSU rRNA gene sequence identity for the definition of a species and 95% for the definition of a genus are generally accepted (Ludwig et al., 1998; Stackebrandt and Goebel, 1994).

Although the sequenced-based approaches have become more common, the effect of sequencing errors on diversity calculations has not been fully evaluated. In this paper we describe the calculation of confidence intervals for a pairwise similarity index using a double binomial model. Similarity scores falling below the lower confidence bound would indicate, with a confidence of $1-z$ (where z is the probability of a Type I error, e.g., 5%), that the two sequences are truly distinct. Similarity scores falling within the confidence interval are deemed to be statistically indistinguishable from scores obtained on identical sequences.

To illustrate the method, confidence intervals were calculated for a set of SSU rDNA sequence data to evaluate the possible problem of sequencing error related to the construction of environmental libraries. The SSU rDNA cloned sequences from different bacterial divisions were sequenced with multiple reactions and the same sequencing primer. By comparing sequence alignments of unedited data, confidence intervals were calculated to estimate the accuracy of sequence determination from the environmental clones. For comparison, non-parametric confidence intervals were also calculated. The results indicate that screening by direct sequence determination could be reliably used to differentiate at the species level.

2. Derivation of confidence interval

The data consist of sequences nominally of fixed length (N). Gaps in the sequence reduce the effective length to $m=N-\#$ gaps. For example, given a sequence length of 101 and a gap of 1, $m=101-1=100$. If k is the total number of nucleic acid base matches obtained by aligning paired sequences, then the similarity coefficient is $S=k/m$, where $k=\{0, 1, \ldots, m\}$. In this example, if there is one mismatch then $S=0.99$.

The same value of the similarity coefficient could be generated for multiple values of k and m if the sequence lengths were different. For example, a sequence length of 200 with two mismatches also results in a similarity of 0.99. Since N is finite, the probability density function for S is discrete and can be written as:

$$\Pr\{S=k/m\}=\sum_{(k,m)\in I_s}\Pr\{K=k \text{ and } M=m\}$$

$$=\sum_{(k,m)\in I_s}\Pr\{K=k|M=m\}\Pr\{M=m\} \quad (1)$$

where each $\Pr\{S=k/m\}$ is the sum of the probabilities for every pair of k and m yielding a given value of S.

The method we use to calculate as the 'double binomial' probability of a match depends on both the probability that the base is counted and the probability that the bases are identical. For each position in an aligned sequence there is a probability p_d that the base pair is counted (a gap does not exist at that position), and probability p_a that the bases are identical (assuming a deletion is not present). Insertions are not explicitly considered but will be treated as a deletion. We also assume the alignments are optimal in that insertions or deletions are never paired with a true base. The number of correct counts (k) resulting from each event probability (p_a, p_d) follows a binomial distribution.

For values of S, p_a, and p_d close to unity, we assume there is only one combination of k and m that generates a given value of S. This assumption results in only one term in Eq. (1) which can then be expanded into a 'double binomial':

$$\Pr\{S=k/m\}$$

$$=\binom{m}{k}p_a^k(1-p_a)^{m-k}\binom{N}{m}p_d^m(1-p_d)^{N-m}. \quad (2)$$

We now introduce random variables x and y, where $x \in \{0, 1\}$ depending on whether a given comparison is a match, and $y \in \{0, 1\}$ denotes absence or presence of a gap. These random variables are independent of each other as each x_i and y_i for the ith comparison in a sequence are independent. Let $k = \Sigma_i(x_i y_i)$ and $m = \Sigma_i(y_i)$. The summed random variables k and m each follow a binomial distribution with expected values $E(x) = p_a$ and $E(y) = p_d$.

To determine the confidence interval, we must select values of p_a and p_d that maximize the goodness of fit of Eq. (2) to empirical data. The solution chosen is the method of moments, which requires values for the mean (μ_s) and variance (σ_s^2) of S. The following result (Brunk, 1975) will prove to be useful in the calculation of σ_s^2:

$$E\left[\frac{x^n}{y^n}\right] = E\left[\frac{E(x^n|y)}{y^n}\right]. \tag{3}$$

The mean is relatively easy to obtain:

$$\mu_S = E(S) = E\left(\frac{k}{m}\right) = E\left[\frac{E(k|m)}{m}\right]$$
$$= E\left[\frac{E(\sum_{i=1}^{m} x_i)}{m}\right] = E\left[\frac{mE(x)}{m}\right] = p_a. \tag{4}$$

Next we derive one of the components needed to obtain the expectation of S^2. The derivation takes advantage of $E(x_i^2) = E(x_i)$ because x can only assume a value of 0 or 1.

$$E[k^2|m] = E\left[\left(\sum_{i=1}^{m} x_i y_i\right)^2 \middle| y_i = 1\right] + E\left[\left(\sum_{i=m+1}^{N} x_i y_i\right)^2 \middle| y_i = 0\right]$$
$$= E\left[\left(\sum_{i=1}^{m} x_i\right)^2\right] = \mathrm{Var}\left[\sum_{i=1}^{m} x_i\right] + \left[E\left(\sum_{i=1}^{m} x_i\right)\right]^2 = \sum_{i=1}^{m} \mathrm{Var}(x_i) + \left[\sum_{i=1}^{m} E(x_i)\right]^2 = \sum_{i=1}^{m} E\left(x_i^2\right) - \sum_{i=1}^{m} [E(x_i)]^2 + \left[\sum_{i=1}^{m} E(x_i)\right]^2 = \sum_{i=1}^{m} E(x_i) - \sum_{i=1}^{m} [E(x_i)]^2 + \left[\sum_{i=1}^{m} E(x_i)\right]^2 = mp_a - mp_a^2 + m^2 p_a^2$$
$$= mp_a(1 - p_a + mp_a). \tag{5}$$

Now the second expectation can be obtained:

$$E\left(S^2\right) = E\left(\frac{k^2}{m^2}\right) = E\left[\frac{E(k^2|m)}{m^2}\right]$$
$$= E\left[\frac{mp_a(1 - p_a + mp_a)}{m^2}\right]$$
$$= p_a(1 - p_a)E\left(\frac{1}{m}\right) + p_a^2. \tag{6}$$

The variance of S is:

$$\sigma_S^2 = \mathrm{Var}(S) = E\left(S^2\right) - [E(S)]^2$$
$$= p_a(1 - p_a)E\left(\frac{1}{m}\right). \tag{7}$$

The random variable m is not independent; it is a joint function of p_d and N, where $E(m) = Np_d$. We are aware of no closed form solution to $E(1/m)$, and employ the following approximation which is justified in Appendix A:

$$E\left(\frac{1}{m}\right) \cong \frac{1}{Np_d}. \tag{8}$$

We note that the approximation in Eq. (8) is equivalent to $1/E(m)$. Substituting Eq. (8) into Eq. (7) we have:

$$\sigma_S^2 \cong \frac{p_a(1 - p_a)}{Np_d}. \tag{9}$$

After substituting Eq. (4) into Eq. (9), and solving for p_d, we get:

$$p_d \cong \frac{\mu_s(1 - \mu_s)}{N\sigma_S^2}. \tag{10}$$

Eqs. (4) and (10), respectively, provide estimates of the needed parameters p_a and p_d, in terms of the mean and variance of S.

To compute the confidence interval, a cumulative distribution function (cdf) for S is needed. This function is actually a discrete integration of Eq. (1), namely $\Pr\{S \le k / m\}$. For large m, k nearly as large as m, and p_a and p_d close to unity, a computationally

tractable approximation to the cdf can be employed. The only requirement is an approximation of the cdf for values of S close to one. We select positive integers $n<N$ and δ. These parameters fix the lower bound of a selected subset of possible values for S and are related to the precision of the approximation.

First, we locate values of S that are of interest by constructing the $(N-n)\times\delta$ matrix C:

$$\mathbf{C}=\begin{bmatrix} \frac{n-\delta}{n} & \cdots & \frac{n-2}{n} & \frac{n-1}{n} & \frac{n}{n} \\ \frac{n-\delta+1}{n+1} & \cdots & \frac{n-1}{n+1} & \frac{n}{n+1} & \frac{n+1}{n+1} \\ \frac{n-\delta+2}{n+2} & \cdots & \frac{n}{n+2} & \frac{n+1}{n+2} & \frac{n+2}{n+2} \\ \vdots & & \vdots & \vdots & \vdots \\ \frac{N-\delta}{N} & \cdots & \frac{N-2}{N} & \frac{N-1}{N} & \frac{N}{N} \end{bmatrix} \tag{11}$$

Note that the rightmost column of C contains only ones. The range of C is given by the interval $[(n-\delta)/n,\ 1]$. Larger δ and smaller values of n generate a larger C matrix and a better approximation of the cdf. Second, we sort C in ascending order such that $c_1\le c_2\le\ldots\le c_p$. We proceed by backward construction to build a double binomial model of the cdf. For any c_i:

$$\Pr\{S\le c_{i-1}\}=\Pr\{S\le c_i\}-\Pr\{S=c_i\}. \tag{12}$$

Initial cdf values are obtained by proceeding from right to left in the sequence:

$$\Pr\{S\le c_p\}=1 \tag{13}$$

$$\Pr\{S\le c_{p-1}\}=1-\Pr\{S=c_p\}. \tag{14}$$

The remaining cumulative probabilities are obtained by recursive application of Eq. (12) to initial results, and can be plotted as a step function. The backward construction procedure is an approximation because it terminates at c_1, and we assume that $\Pr\{S<(n-\delta)/n\}$ is close to zero. If $\Pr\{S\le c_i\}=1-z$, then $[c_i,\ 1]$ is the zth confidence interval for the obtained value of S.

As a comparison to the above approach we used the nonparametric quantile method for estimating confidence intervals from the data. We define the quantile q_z (e.g, the 95% quantile) as $r/(y+1)$ for the rth order statistic of an ordered data series $<s_1,\ldots,s_r,\ldots s_y>$ of length y. The 'one-tailed' interval (when S is close to 1) for level of confidence $(1-z)$ is $[q_z,\ 1]$. The nonparametric estimators have the advantages of requiring few assumptions and ease of computation. The main disadvantages are: (1) the quantile-based estimates are very sensitive to noise and outliers and (2) they cannot extrapolate beyond the range of the sample data. Both issues are aggravated with smaller sample sizes.

3. Example

The DNA sequences used in this study were derived from groundwater samples collected from two wells at the Natural and Accelerated Bioremediation Research (NABIR) Program Field Research Center in Oak Ridge, Tennessee.

3.1. Materials and methods

Groundwater samples (1–2 l) were collected and transported to the laboratory in amber glass bottles. Bacteria were harvested by centrifugation (10,000 $\times g$, 4 °C for 30 min), and the pellets were stored at −80 °C until used for DNA extraction. The cell pellet was resuspended in a lysis buffer, and the cells were disrupted using a previously described grinding method (Zhou et al., 1996). DNA was extracted as previously described (Zhou et al., 1996, 1997), and the precipitated DNA was purified by gel electrophoresis plus mini-column preparation (Wizard DNA Clean-Up System, Promega, Madison, Wisconsin, USA) (Zhou et al., 1996).

The PCR reactions (20 μl) contained 2 μl of 10 × PCR reaction buffer (500 mM KCl, 100 mM Tris–HCl pH 9.0, 1% Triton X-100), 1.5 μl of 25 mM $MgCl_2$, 0.2 μl of 400 ng μl^{-1} bovine serum albumin (Boehringer Mannheim), 0.2 μl of 25 mM 4 × dNTPs (USB Chemicals), 10 pmol of each primer, 2.5 U of Taq polymerase, and 1 μl of purified DNA (5–10 ng). To minimize PCR-induced artifacts, the optimal number of cycles was determined and five PCR reactions were combined prior to cloning as described previously (Qiu et al., 2001). The combined PCR products were separated by electrophoresis in a low-melting point agarose gel (0.8%), the appropriate band excised, and the DNA extracted with a Promega

Wizard Prep Kit (Madison, Wisconsin, USA) according to the manufacturer's instructions. Recovered DNA was resuspended in 6 μl ddH_2O, 2 μl was ligated with the pCR2.1 vector from a TA-cloning kit, and competent *Escherichia coli* cells were transformed according to provided protocol (Invitrogen, San Diego, California, USA). The SSU *rRNA* genes were amplified with the FD1 and 1540R primers, and the PCR products were purified with the ArrayIt™ PCR Purification Kit (TeleChem International, Inc.) or treated with ExoSAP-IT™ (US Biochemical Corporation) according to manufacturer instructions.

DNA sequences were determined with a BigDye™ Terminator kit (Applied Biosystems) with a 3700 DNA analyzer (Perkin-Elmer) according to the manufacturer's instructions with the SSU rDNA-specific primer 529R, and compared with sequences from GenBank. Each clone was sequenced five times with the same primer (529R, *E. coli* designation). The sequences were aligned with ClustalW (Thompson et al., 1994). Phylogenetic analyses of the partial SSU 16S rRNA gene sequences were conducted using MEGA version 2.1 (Kumar et al., 2001). Neighbor-joining phylogenies were constructed from dissimilatory distances and pairwise deletion of gaps and missing data. Maximum parsimony phylogenies did not differ significantly (data not shown).

3.2. Results

Partial sequences were determined and compared to sequences in GenBank and the Ribosomal Database Project for presumptive identification. Identification was based on comparing the V2–V6 region of the SSU rRNA gene sequence (approximately first 400 to 500 nucleotides). Nine clones from the following eubacterial divisions or subdivisions were selected for sequence comparison: α-*Proteobacteria*, β-*Proteobacteria*, γ-*Proteobacteria*, low G+C Gram-positive, high G+C Gram-positive, *Bacteroidetes, Nitrospira,* and *Verrucomicrobia* (Fig. 1). The high G+C clone (300A-B08) was most closely related to an *Actinomycete* MC 13, and the low G+C clone (300G-F05) was distantly related to an uncultivated environmental clone. The *Nitrospira*-like sequence (005B-A07) was distantly related to *Nitrospira marina*, and the α-Proteobacterial sequence (300E2-A11) was most similar to *Agrobacterium tumefaciens*. The *Bacteroidetes* (formerly the *Cytophaga–Flexibacter–Bacteroides*) were represented by a *Flexibacter*-like sequence (300BHJ-G10). The β-*Proteobacteria* were represented by two sequences most closely related to *Rhodoferax fermentans* (300A-D02) and *Comamonas acidovorans* (300E2-E02), and the γ-*Proteobacteria* clone (300A-F01) was most similar to *Methylobacter* T20. The clone 300BHJ-H06 was distantly related to *Verrucomicrobia spinosum*.

The resulting sequences varied between approximately 400 and 450 nucleotides with a mean length of 425. The raw sequence data were aligned, and pairwise similarities were calculated for each clone as described above (10 similarities per clone). The resulting 90 similarity measures were pooled and treated as identically distributed observations from sequences 400 nucleotides in length ($N=400$). A larger N did not produce reasonable fits of the double binomial model to the observations, presumably due to unequal sequence lengths and increased number of gaps. The mean similarities were greater than 0.997 for each clone.

The mean and variance of the pooled similarity measures were 0.998476 and 4.1243×10^{-6}, respectively. Using Eq. (4), the estimate of p_a was 0.998476, and application of Eq. (10) resulted in 0.9222 as the estimate of p_d. A 121×8 C-matrix was constructed for which m ranged from 280–400, and $\delta=8$. The estimated cdf of S calculated by the proposed method is shown in Fig. 2. The model estimates were in good agreement with the empirical cdf. Both plots display three major step changes in the interval from 0.99 to 1.0, and the model-predicted step changes were shifted slightly to the left compared to the data.

The model results suggested that similarity values below 0.9946 were likely derived from dissimilar sequences at a confidence level of 0.95. Both the model and the quantile method generated similar 0.90 and 0.99 confidence intervals. Thus, when comparing sequences derived from operations such as cloning and sequencing, values of similarity on the order of 0.995 could be derived from the sequencing of identical clones 1 of 10 times. Similarities on the order of 0.992 can be derived from sequencing of identical clones 1 out of 100 times.

In the range close to 100% similarity, changes in the mean and variance of replicate sequencing had a

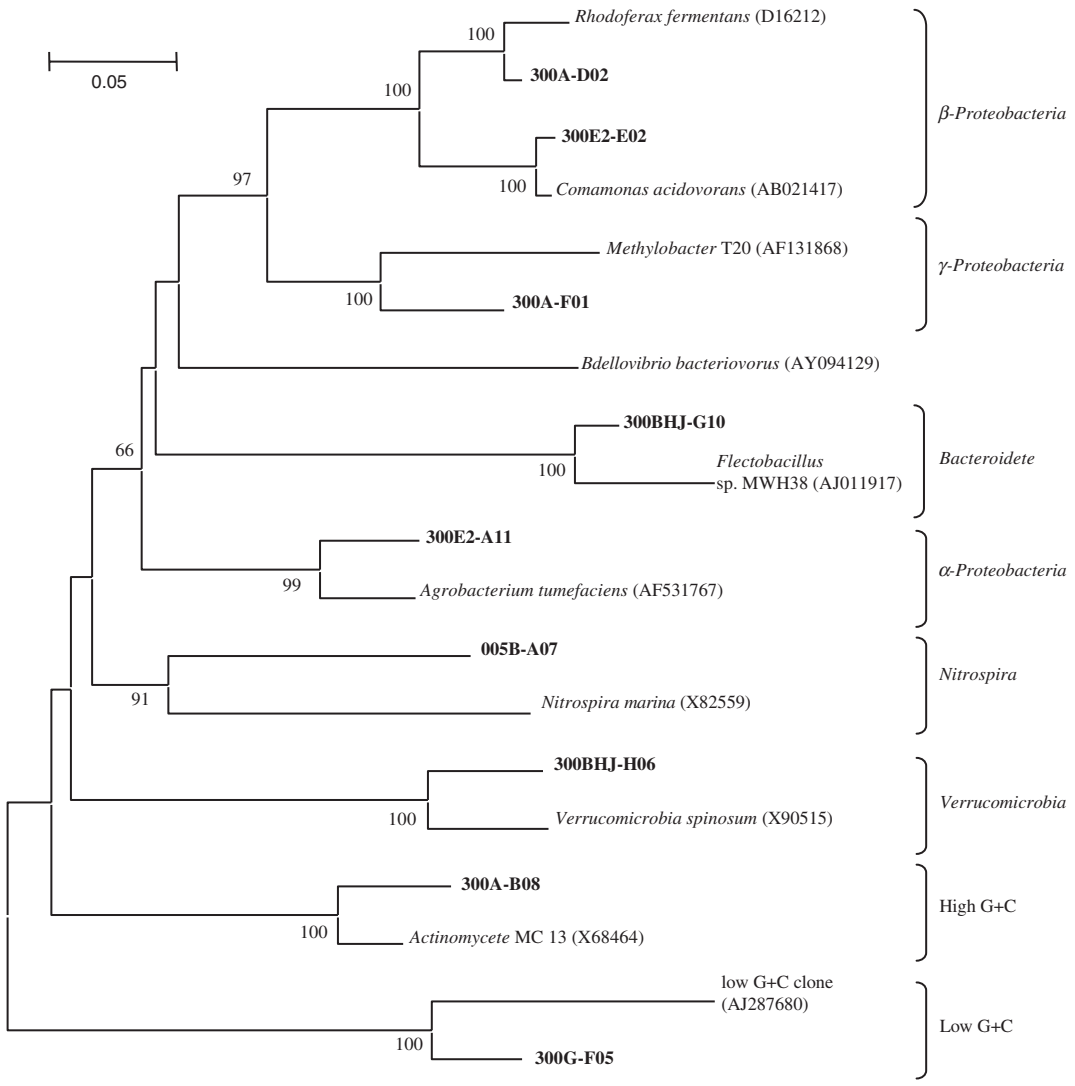

Fig. 1. Tree showing the phylogenetic placement of the nine clones (shown in bold) chosen for the sequence comparison example. The tree is based on the neighbor-joining method and a pairwise deletion. The percentage of 500 bootstrap values that supported in branch is shown, and bootstrap values below 50% are not shown. The accession numbers for the reference sequences are: *Rhodoferax fermentans* (D16212), *Comamonas acidovorans* (AB021417), *Methylobacter* T20 (AF131868), *Bdellovibrio bacteriovorus* (AY094129), *Flectobacillus* sp. MWH38 (AJ11917), *Agrobacterium tumefaciens* (AF531767), *Nitrospira marina* (X82559), *Verrucomicrobia spinosum* (X90515), *Actinomycete* MC13 (X68464), and uncultured low G+C clone (AJ287680).

small effect on the lower confidence interval. We repeated the calculation of the lower confidence intervals ($p<0.05$ and $p<0.01$) using twelve combinations of mean similarity (0.9975–0.9990) and variance (2×10^{-6} to 8.33×10^{-6}). For $p<0.05$ the resulting lower confidence interval ranged from 0.9925 to 0.9965 and for $p<0.01$ the range was 0.9896 to 0.9948. Due to the stepwise nature of the double binomial function it was not possible to reliably interpolate between values for the mean and the variance. Thus, the confidence limit can only be calculated with a high degree of precision by actually performing the calculation rather than by interpolation from a table.

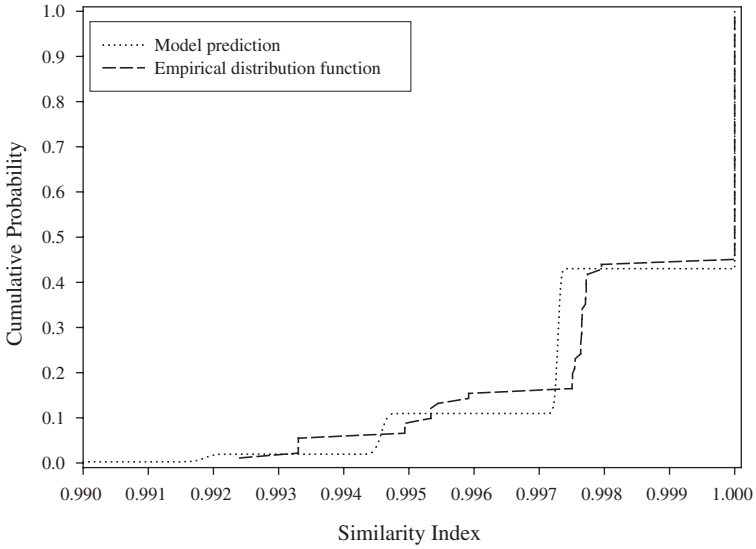

Fig. 2. Empirical cumulative distribution function and double binomial model predictions of sequence comparison.

The empirical similarity values were arranged from lowest to highest, and a quantile estimate, q_z, for a confidence level 0.90 was estimated from the highest value to the value representing 10% of $(y+1)$. The data set included a series of 90 empirical observations. The value at 10% of $90+1=91$, which is approximately the ninth order statistic in the series, was 0.9953. Consequently, we were confident at the 0.90 level that a new observation would not be below 0.9953.

For the 0.90 and 0.99 level confidence intervals, the double binomial model estimates were very similar to the quantile estimates (Table 1). Differences were less than 0.1%. However, there was a large 'step' in the 0.90 to 0.95 range for the double binomial model that seems to identify a critical breakpoint for similarity. A similarity index of 0.9946–0.9947 was associated with a much higher confidence level than similarities that were only slightly larger. It was possible to use the double binominal model to construct confidence intervals beyond the 0.99 level, whereas sample size limited the ability of quantiles to estimate confidence beyond this level.

Table 1
Lower bounds of confidence intervals for sequence similarity index based on the quantile and double binomial methods

$1-z$	Quantile	Double binomial	Difference
0.90	0.9953	0.9947	0.06
0.95	0.9933	0.9946	−0.13
0.99	0.9924	0.9919	0.05

4. Discussion

The goal of many large-scale sequencing projects is one error in 10^4 nucleotides, but a systematic evaluation of sequencing error in most genome projects has not been reported. Hill et al. (2000) estimated a sequencing error rate based on the mobile genetic element IS*10*. The results suggested that an accuracy of less than 1 sequencing error in 10^4 base-pairs was currently achieved, and that IS*10* was neither unusually difficult nor easy to sequence. The error frequency in 40 single-pass sequence reads of expressed sequence tags that contained IS*10* was approximately 3.1%, and the average sequencing error was between 0.4 and 1.3% when the SSU rRNA gene sequences from different bacterial lineages were used in our study. Large-scale sequencing projects usually incorporate overlapping sequence reads to help improve accuracy, and thus would help explain a lower error rate (1 per 10^4) compared to single reads of non-overlapping sequence (screening of a SSU rRNA gene library). The serial analysis of gene expression (SAGE) method for the estimation of transcript abun-

dance is dependent upon sequence determination, and the impact of possible error rate was recently evaluated. Colinge and Feger (2001) developed techniques to compute and model sequencing error in relation to SAGE, and Stollberg et al. (2000) developed a basic maximum likelihood estimation based on certain mathematical biases.

Our results suggested that only a small portion of the apparent diversity observed in cloning and sequencing of environmental communities was derived from sequencing error, and the model estimates that few of the similarity values were affected by sequencing errors. More significant difficulties likely arose from problems in the DNA extraction and the biases introduced by PCR amplification such as chimeric sequences and heteroduplexes (Zhou et al., 2002; Qiu et al., 2001). It should also be noted that the sequences analyzed in this study were partial sequences, and partial sequences are commonly used for relative comparisons within and between samples.

The sequence determination of clonal SSU *rRNA* genes should not be used as the sole criteria for identification (Fox et al., 1992). However, partial sequences of informative regions can be used to efficiently screen libraries and thereby reduce the number of full-length sequences needed to represent the recovered diversity from a microbial community. Further studies are needed to determine how well diversity in particular segments of SSU rRNA gene sequences can predict overall sequence relationships, but variable and conserved regions will perform differently for increasing levels of likeness.

In addition, our results indicated that there was great discrimination between strains and little chance of error at similarities greater than 99% (Fig. 2 and Table 1). At similarities of 99%, the probability of sequencing errors resulting in the misclassification of two clones from the same strain (assigning them as different OTUs) was extremely small. The results indicated that the values of 97% to 98% similarity that have been explicitly or implicitly used in previous studies could be achieved with confidence when clones are randomly selected and screened via sequence determination. The implication of our data was that sequences with similarity values significantly greater than 99% cannot conclusively be placed into phylogenetic groups based on random sequence screening, but OTU classifications of 97% to 98% can certainly be obtained with confidence. There is not currently a cohesive definition for a bacterial species, and much more work and debate are needed before comprehensive and systematic approaches are adopted for the analysis of microbial communities. However, our results indicated that a random sequencing approach of the V2–V6 region of SSU rRNA gene sequences could differentiate between environmental clones at the 97% to 98% similarity level without a significant effect from sequencing error.

Acknowledgments

This research was supported by the United States Department of Energy, Office of Science, Office of Biological and Environmental Research, Natural and Accelerated Bioremediation Research (NABIR) Program. Oak Ridge National Laboratory is managed by UT-Battelle, LLC for the U.S. Department of Energy under contract DE-AC05-00OR22725.

Appendix A

The estimation of $E(1/m)$, where m follows a binomial distribution, depends on two approximations. The binomial theorem Eq. (A1) is used in the following rationale:

$$\sum_{m=0}^{N}\binom{N}{m}x^m y^{N-m} = (x+y)^N. \tag{A1}$$

First expand the expectation and rearrange terms:

$$\begin{aligned} E\left[\frac{1}{m}\right] &= \sum_{m=0}^{N}\left(\frac{1}{m}\right)\binom{N}{m}p_d^m(1-p_d)^{N-m} \\ &= \sum_{m=0}^{N}\left(\frac{1}{m}\right)\left(\frac{N!}{m!(N-m)!}\right)p_d^m(1-p_d)^{N-m} \\ &= \sum_{m=0}^{N}\left(\frac{m+1}{m}\right)\left(\frac{1}{m+1}\right) \\ &\quad\times\left(\frac{N!}{m!(N-m)!}\right)p_d^m(1-p_d)^{N-m}. \end{aligned} \tag{A2}$$

We introduce the first approximation in the case where only large values of m close to N have

associated probabilities much greater than zero. Rearranging:

$$E\left[\frac{1}{m}\right] \cong \left(\frac{N+1}{N}\right)\sum_{m=0}^{N}\left(\frac{1}{m+1}\right)\left(\frac{N!}{m!(N-m)!}\right)p_d^m \times (1-p_d)^{N-m} \cong \left(\frac{N+1}{N}\right)\left(\frac{1}{(N+1)p_d}\right) \times \sum_{m=0}^{N}\frac{(N+1)!}{(m+1)![(N+1)-(m+1)]!}p_d^{m+1} \times (1-p_d)^{[(N+1)-(m+1)]}. \quad \text{(A3)}$$

Let $N' = N+1$ and $m' = m+1$:

$$E\left[\frac{1}{m}\right] \cong \left(\frac{1}{Np_d}\right)\sum_{m'=1}^{N'-1}\left(\frac{N'!}{m'!(N'-m')!}\right)p_d^{m'} \times (1-p_d)^{N'-m'}. \quad \text{(A4)}$$

Apply Eq. (A1) while accounting for the first and last terms of the sum:

$$E\left[\frac{1}{m}\right] \cong \frac{1}{Np_d}\left[1-(1-p_d)^{N'}-p_d^{N'}\right] \cong \frac{1}{Np_d} \times \left[1-(1-p_d)^{N+1}-p_d^{N+1}\right]. \quad \text{(A5)}$$

For large N,

$$E\left[\frac{1}{m}\right] \cong \frac{1}{Np_d}. \quad \text{(A6)}$$

References

Brunk, H.D., 1975. An Introduction to Mathematical Statistics. Xerox Publishing Company, Lexington, MA.

Colinge, J., Feger, G., 2001. Detecting the impact of sequencing errors on SAGE data. Bioinformatics 17, 840–842.

Curtis, T.P., Sloan, W.T., Scannell, J.W., 2002. Estimating prokaryotic diversity and its limits. Proc. Natl. Acad. Sci. U. S. A. 99, 10494–10499.

Dykhuizen, D.E., 1998. Santa Rosalia revisited: why are there so many species of bacteria? Antonie van Leeuwenhoek 73, 25–33.

Fox, G.E., Wisotzkey, J.D., Jurtshuk, P., 1992. How close is close – 16S ribosomal-RNA sequence identity may not be sufficient to guarantee species identity. Int. J. Syst. Bacteriol. 42, 166–170.

Hill, F., Gemund, C., Benes, V., Ansorge, W., Gibson, T.J., 2000. An estimate of large-scale sequencing accuracy. EMBO Rep. 1, 29–31.

Kumar, S., Tamura, K., Jakobsen, I.B., Nei, M., 2001. MEGA2: molecular evolutionary genetics analysis software. Bioinformatics 17, 1244–1245.

Ludwig, W., Strunk, O., Klugbauer, S., Klugbauer, N., Weizenegger, M., Neumaier, J., et al., 1998. Bacterial phylogeny based on comparative sequence analysis. Electrophoresis 19, 554–568.

Øvreås, L., 2000. Population and community level approaches for analysing microbial diversity in natural environments. Ecol. Lett. 3, 236–251.

Pace, N.R., 1997. A molecular view of microbial diversity and the biosphere. Science 276, 734–740.

Qiu, X.Y., Wu, L.Y., Huang, H.S., McDonel, P.E., Palumbo, A.V., Tiedje, J.M., et al., 2001. Evaluation of PCR-generated chimeras, mutations, and heteroduplexes with 16S rRNA gene-based cloning. Appl. Environ. Microbiol. 67, 880–887.

Stackebrandt, E., Goebel, B.M., 1994. A place for DNA–DNA reassociation and 16S ribosomal-RNA sequence analysis in the present species definition in bacteriology. Int. J. Syst. Bacteriol. 44, 846–849.

Stollberg, J., Urschitz, J., Urban, Z., Boyd, C.D., 2000. A quantitative evaluation of SAGE. Genome Res. 10, 1241–1248.

Thompson, J.D., Higgins, D.G., Gibson, T.J., 1994. CLUSTAL-W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22, 4673–4680.

Woese, C.R., Fox, G.E., 1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc. Natl. Acad. Sci. U. S. A. 74, 5088–5090.

Woese, C.R., Kandler, O., Wheelis, M.L., 1990. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc. Natl. Acad. Sci. U. S. A. 87, 4576–4579.

Zhou, J.Z., Bruns, M.A., Tiedje, J.M., 1996. DNA recovery from soils of diverse composition. Appl. Environ. Microbiol. 62, 316–322.

Zhou, J.Z., Davey, M.E., Figueras, J.B., Rivkina, E., Gilichinsky, D., Tiedje, J.M., 1997. Phylogenetic diversity of a bacterial community determined from Siberian tundra soil DNA. Microbiology 143, 3913–3919.

Zhou, J.Z., Xia, B.C., Treves, D.S., Wu, L.Y., Marsh, T.L., O'Neill, R.V., et al., 2002. Spatial and resource factors influencing high microbial diversity in soil. Appl. Environ. Microbiol. 68, 326–334.