Classification of Oligonucleotide Fingerprints: Application for Microbial Community and Gene Expression Analyses.

by

Katechan Jampachaisri, Lea Valinsky, James Borneman, and S. James Press, Departments of Statistics and Plant Pathology, University of California, Riverside, 92521 USA.

ABSTRACT

Oligonucleotide fingerprinting is an array-based approach used for analysis of microbial community composition and gene expression profiling. Oligonucleotide fingerprinting of ribosomal RNA genes (OFRG) is an approach that sorts rDNA clones into taxonomic groups through a series of hybridization experiments. For every hybridization experiment, the signal intensities are transformed into three discrete values 0, 1, and N, where 0 and 1 respectively specify negative and positive hybridization events and N designates an uncertain assignment. Various approaches have been performed to resolve the uncertainty. These include Bayesian classification with the normal distribution, Bayesian classification with the exponential distribution and gamma distribution, along with tree-based classification. All 0-1 data produced from each classification approach, including the original 0-1-N data, were afterwards clustered using the Unweighted Pair Group Method with Arithmetic Mean. The performance of each approach was compared with results in known 0-1 reference data. The comparisons indicated that the approach using Bayesian classification with normal densities followed by tree clustering out-performed all others. The results we obtained suggest a useful and improved means for identifying microorganisms using the OFRG approach. We also discuss how this Bayesian approach may be useful for analysis of gene expression data.

1. INTRODUCTION

Microorganisms are integral components of ecosystems and human civilization. They play important roles in the detoxification of polluted environments, provide essential nutrients for plants, and transform waste materials into useful commodities such as compost (Atlas and Bartha, 1998; Christensen; 1989; Tuomela et al., 2000). Microorganisms are used in fermentation processes, producing a myriad of important food and beverage products. In the biotechnology arena, they are a vital source of useful compounds and provide vehicles for the production of genetically engineered products such as pharmaceuticals (Bull et al., 2000; Chapela, 1997). Despite these important discoveries, the microbial contribution to natural ecosystems and their potential for society have yet to be fully realized, as current experimental methods do not allow thorough descriptions of the microbial communities inhabiting most environments.

One of the first steps in characterizing an ecosystem is to identify the organisms inhabiting it. Traditionally, microorganisms have been classified by characterizing their morphological and physiological traits. However, such traits do not provide a meaningful framework for evolutionary classifications. Moreover, this approach will only detect a fraction of the existing microorganisms, as the majority of them do not readily grow on laboratory media (Amann et al., 1995). In the 1970s, the development of comparative ribosomal RNA (rRNA) analysis provided an evolutionary basis for prokaryotic taxonomy (Fox et al., 1977; Sogin et al., 1972; Woese et al., 1975; Woese and Fox, 1977). The subsequent development of strategies to analyze rRNA molecules and genes (rDNA) obtained from the environment provided a culture-independent means to examine the immense diversity of microorganisms inhabiting the natural world (Giovannoni et al.; 1990, Pace, 1997).

2. THE PROBLEM

Numerous rDNA-based strategies have been developed for microbial community analysis. The most accurate approach is to analyze the entire nucleotide sequence of the rRNA molecule or gene. However, because of the high costs associated with examining such diverse communities in this manner, this approach is not practical for thorough analysis of microbial community composition. Other methods such as denaturing gradient gel electrophoresis (DGGE) (Muyzer et al., 1993), terminal restriction fragment length polymorphisms (T-RFLP) (Liu et al., 1997), ribosomal intergenic space analysis (RISA) (Borneman and Triplett, 1997), and amplified ribosomal DNA restriction analysis (ARDRA) (Vaneechoutte et al., 1992) enable relatively inexpensive and rapid analysis of many samples, but they typically generate only cursory descriptions of microbial community composition.

To overcome these experimental obstacles, a method termed oligonucleotide fingerprinting of ribosomal RNA genes (OFRG) was developed (Valinsky et al., 2002a, b). OFRG is an adaptation of a method used for gene expression profiling (Drmanac, 1999, Drmanac et al., 1991, Lennon and Lehrach, 1991). OFRG is an array-based method that enables extensive analysis of microbial community composition. OFRG works by sorting rDNA clones into taxonomic groups through a series of hybridization experiments, each using a single DNA probe. The probe sequences for the OFRG analysis are selected for their ability to differentiate known rDNA sequences in the GenBank database (NCBI) (Borneman et al., 2001). These hybridization experiments are used to produce hybridization fingerprints, which specify the presence or absence of the probe sequences in each clone. The microorganisms are identified by clustering the hybridization fingerprints of the unknown rDNA clones with those of known rDNA sequences.

A brief description of experimental process for OFRG follows. Microbial rDNA are isolated from a sample of interest by extracting DNA from the microorganisms and then PCR amplifying the rDNA. Cloned rDNA fragments are arrayed on nylon membranes and subjected to a series of hybridization experiments, each using a single DNA oligonucleotide probe. The signal intensities from these experiments are transformed into binary vectors, which we call hybridization fingerprints. Hybridization fingerprints from the unidentified rDNA clones are clustered with fingerprints from known rDNA sequences using UPGMA (Unweighted Pair

Group Method with Arithmetic Mean). The unidentified rDNA clones are identified by their association with known rDNA sequences in the UPGMA tree.

3. THE DATA

One of the crucial components of the OFRG analysis is the process by which the signal intensity data are transformed into hybridization fingerprints. The hybridization fingerprints specify whether the probes hybridize, or do not hybridize, to the rDNA clones. Probes that hybridize the rDNA clones should produce larger signal intensity values than those that do not hybridize. Following is a description of how prior OFRG studies have processed the data to produce the hybridization fingerprints.

Hybridization fingerprints have been generated by transforming the hybridization signal intensity data into three discrete values 0, 1, and N, where 0 and 1 respectively specify negative and positive hybridization events and N designates an uncertain assignment. The signal intensity data from the unidentified rDNA clones were transformed into 0, 1, and N based on the signal intensities from control clones, which are clones with defined nucleotide sequences. For most probes, the control clones expected not to hybridize with the probe (negative controls) have signal intensity values less than the control clones expected to hybridize with the probe (positive controls); conversely, the signal intensity values from the positive clones are higher than those from the negative clones. For probes that function in this manner, clones with intensity values less than or equal to x were given a 0 classification, where x is the highest intensity value generated by a negative control. Clones with intensity values greater than or equal to y were given a 1 classification, where y is the lowest value generated by a positive control. All other clones were given an N classification. For some probes, not all of the control clones perform in the predicted manner; for example, some positive control clones may have intensity values that are lower than some of the negative control values and visa versa. For probes that function in this manner, clones with intensity values less than x were given a 0 classification, where x is the lowest intensity value generated from a positive control. Clones with intensity values greater than y were given a 1 classification, where y is the highest value generated by a negative control. All other clones are given an N classification. Performing this analysis with all probes for all clones creates a hybridization fingerprint for each clone. An example of a hybridization fingerprint created by 26 probes is 000101N001000N110101111000.

This report compares and evaluates several approaches for transforming the signal intensity data into hybridization fingerprints. As with most nucleic acid hybridization experiments, the signal intensity data do not consistently fall into discrete categories. This factor along with the aforementioned method for transforming these data lead to the production of hybridization fingerprints with a considerable number of N classifications. The goal of this work was to minimize the number of N classifications, which should increase the accuracy of the OFRG analysis.

4. STATISTICAL ANALYSIS

4.1 Overview

Our data consists of signal intensities from hybridization experiments between arrayed rRNA genes and probes. We classify these hybridizations as "1", if hybridization took place, "0" if it did not, and an "N", if we were unable to determine how to classify the result. The 0-1-N classification produces hybridization fingerprints, which are used to cluster the rDNA clones into taxonomic groups. In prior studies, this grouping has been less than fully satisfactory because the number of gene/probe clone combinations in the "N" group was not small, and we had no idea how to classify the "N" combinations. (We felt confident, however, about the combinations that had been already classified as 1 or 0, according to whether hybridization had taken place in that experiment, or not). In addition, the hierarchical clustering algorithm (UPGMA) used to cluster the gene/probe combinations classifies the "N"s somewhat randomly, as it allocates them proportionally to already-established 0-1 values between pairs of genes, and then forms a distance matrix for the clustering procedure.

To illustrate the allocation procedure, we use a simple example. Suppose there are 3 genes and 10 probes, and the 0-1 sequences are given by

gene1--- (0000100010)
gene2 --- (1001110110)
gene3 --- (N011000110).

Consider each pair of genes as follows.

Gene 1-2: gene1---(0000100010)
gene2---(1001110110)

Since there is no missing value occurring in either of the sequences of these genes, the distance between gene 1 and 2 can be obtained easily by computing the proportion of mismatched characters between them. There are 4 mismatches, so the "distance" between them is: 4/10 = 0.4.

Gene 2-3: gene2---(1001110110)
gene3---(N011000110).

As we see there is a missing value occurring in gene 3 at the first position. To distribute the missing value, we consider all mismatched pairs that occurred in the other nine positions. Three out of nine pairs are mismatched with proportion 3/9=0.33. So, out of 10 probes there are 3.33 mismatches. The distance between gene 2 and 3 would be 0.333 (3.33/10).

Gene 1-3: gene1---(0000100010)
gene3---(N011000110)

The distance between gene 1 and 3 can be performed similarly by considering the proportion of mismatched pairs occurring in the nine unambiguous positions, which is 4/9=0.44. So, out of 10 probes there are 4.44 mismatches. The distance between gene 1 and 3, accordingly, would be obtained as 0.444 (4.44/10).

The (symmetric) pairwise-distance matrix for this example is given by:

	gene1	gene2	gene3
gene1	-	0.400	0.444
gene2		-	0.333
gene3			-

which the PAUP (Phylogenetic Analysis Using Parsimony) computer program, Version 4.0, Beta 9, (Swofford, 2001; Maddison *et al*., 1997) would then use to form clusters, using UPGMA.

Because the proportional allocation approach for classifying and clustering the "N"s was clearly quite arbitrary, we sought to improve upon that procedure.

The procedure we adopted to seek improvement in clustering consists of first classifying *all* of the gene/probe combinations statistically as to whether hybridization has taken place, or not, thereby eliminating the number of combinations in the unclassified, or unknown, category. Then we use UPGMA to cluster a fully-classified set of 0-1 data, with no proportional allocation clustering being necessary. We used a reference set of data for which we knew the correct clustering, to evaluate how well the statistical classification/clustering procedure would compare with the 1-0-N procedure.

The statistical procedures used were:
1) Bayesian classification; that is, the procedures involved classifying the intensities of each of the gene/probe combinations, and the procedures used were developed using Bayesian classification modeling (see, e.g., Press, 1982, 2003). We made various assumptions about the distributions of the data and we tested them, as described below in Sect. 4.2. The data were classified into one of two classes, the hybridized class, and the un-hybridized class. But to establish the structure of the populations we needed "known data", that is, data whose classifications were known with certainty. For known data we used the reference data, and we refer to it as "training data", to be consistent with terminology used in the classification literature.
2) multivariate hierarchical clustering using UPGMA.
3) "Tree modeling" for the resulting clustering procedure that minimized the "distance" between the tree corresponding to the reference population, and the statistically clustered data.

The various distributions studied and compared were: normal distributions, exponential distributions and gamma distributions. We also tried tree-based (non-parametric) classification

The complete data set consists of 27 probes and 1464 genes, after excluding genes that did not hybridize properly. – We extracted 65 genes used as training data from the 1464 genes. These 65 genes used for training data had known class membership. That is, we knew with certainty

whether these 65 had hybridized or not. Our objective was to find the best methods for both classifying and clustering the training data (65 genes), as compared with the correct cluster values, which we knew, in order to apply the resulting optimal procedure to the entire dataset (1464 genes).

The performance of each classification procedure was evaluated by calculating the *apparent error rate* (APER), defined as the fraction of misclassified observations in the training sample (Johnson and Wichern, 1992). The apparent error rate did not depend on the form of population densities and could be readily calculated by constructing the 2 x 2 "confusion matrix" (see Press, 1982), where the actual and predicted class memberships obtained from each classification approach were compared.

Prior to conducting classification and clustering analyses, all normalized intensities that were negative were truncated to zero. Then, we explored the distributions of the remaining data (in each dimension) using both histograms and normal probability plots (Venebles and Ripley, 1999). Reference data points for each probe were used to estimate the parameters of the underlying distributions. The data usually looked non-normally distributed, and very much like exponentially distributed, or gamma distributed data. We examined classifications using transformations of the non-normal data (we attempted to transform very non-normal looking data to normality). Therefore, each classification procedure would be performed on either the original or the transformed intensities, depending on the distribution assumption indicated in the procedure being applied. The derivations of the classification methods used may be found in (Press, 1982, 1989, 2003).

4.2 Bayesian Classification With Normal Distributions

The preliminary investigation of histograms and normal probability plots revealed that the normalized intensities in each probe tended to have a rather right-skewed, non-normal shape. The Box-Cox transformation (Box and Cox, 1964) technique, consequently, was utilized to transform the intensities in each probe to approximate normality. Then the Bayesian classification approach was performed on the transformed data. The un-hybridized and hybridized classes or populations are denoted by π_0 and π_1, respectively, and the distributions are denoted by: $\pi_1 \sim N(\mu_0,\sigma_0^2)$, and $\pi_1 \sim N(\mu_1,\sigma_1^2)$, respectively. All parameters, $(\mu_0,\sigma_0^2,\mu_1,\sigma_1^2)$, were assumed to be unknown.

For class π_j (j = 0, 1) in the training sample with known class memberships corresponding to each clone, the intensities, $(x_1^{(j)},x_2^{(j)},\ldots,x_{n_j}^{(j)})$, were assumed to be independent and identically distributed (i.i.d.) according to π_j. For convenience, the superscript j would be omitted when considering only one class, j. Let $\hat{\mu}_j = \bar{x}_j = \sum_{i=1}^{n_j} x_i \Big/ n_j$ and $\hat{\sigma}_j^2 = s_j^2 = \sum_{i=1}^{n_j}\left(x_i - \bar{x}_j\right)^2 \Big/ (n_j - 1)$ be sufficient, unbiased estimators of μ_j and σ_j^2, respectively. Since $(\bar{x}_j \,|\, \mu_j,\sigma_j^2,\pi_j) \sim N(\mu_j,\frac{\sigma_j^2}{n_j})$ and $(s_j^2 \,|\, \sigma_j^2,\pi_j) \sim \text{Inverted-Gamma}\left(\frac{n_j-3}{2},\frac{(n_j-1)s_j^2}{2}\right)$, the

likelihood as a function of unknown parameters could be written in term of these estimators, given by

$$p(\bar{x}_j, s_j^2 \,|\, \mu_j, \sigma_j^2, \pi_j) = p(\bar{x}_j \,|\, \mu_j, \sigma_j^2, \pi_j) \cdot p(s_j^2 \,|\, \sigma_j^2, \pi_j)$$
$$\propto \frac{1}{\left(\sigma_j^2\right)^{n_j/2}} e^{-\frac{1}{2\sigma_j^2}\left\{n_j\left(\mu_j - \bar{x}_j\right)^2 + (n_j - 1)s_j^2\right\}}, \quad (1)$$

where the symbol $\propto$ denotes proportionality. Adopting a vague prior distribution (such a distribution represents minimum prior information), $p(\mu_j, \sigma_j^2) \propto 1/\sigma_j^2$, $0 < \sigma_j^2 < \infty$. The posterior density is formed by Bayes' theorem as a product of the likelihood function and prior information, expressed as

$$p(\mu_j, \sigma_j^2 \,|\, \bar{x}_j, s_j^2, \pi_j) \propto \frac{1}{\left(\sigma_j^2\right)^{\frac{n_j}{2}+1}} e^{-\frac{1}{2\sigma_j^2}\left\{n_j\left(\mu_j - \bar{x}_j\right)^2 + (n_j - 1)s_j^2\right\}}. \quad (2)$$

The interest here was to predict class membership of an observation, *z*, which is known to belong to either π_0 or π_1. After integrating the product of the likelihood of observation *z* and the posterior distribution (2) with respect to all unknown parameters, the predictive distribution would then be obtained in the form of a Student's *t*-distribution, shown as

$$p(z \,|\, \bar{x}_j, s_j^2, \pi_j) \propto \frac{1}{\left\{1 + \frac{n_j}{n_j^2 - 1}\left(\frac{z - \bar{x}_j}{s_j}\right)^2\right\}^{n_j/2}}. \quad (3)$$

Let q_0 and q_1 be prior probabilities that *z* came from π_0 and π_1, respectively. The posterior odds ratio, afterwards, could be calculated based on the posterior classification probabilities, yielding the ratios of pairs of Student's *t*-densities as given by

$$\frac{P(z \in \pi_0 \,|\, z)}{P(z \in \pi_1 \,|\, z)} = \frac{P(z \in \pi_0) \cdot p(z \,|\, \bar{x}_0, s_0^2, \pi_0)}{P(z \in \pi_1) \cdot p(z \,|\, \bar{x}_1, s_1^2, \pi_1)}$$
$$\propto L_{01} \frac{\left\{1 + \frac{n_0}{n_0^2 - 1}\left(\frac{z - \bar{x}_0}{s_0}\right)^2\right\}^{n_0/2}}{\left\{1 + \frac{n_1}{n_1^2 - 1}\left(\frac{z - \bar{x}_1}{s_1}\right)^2\right\}^{n_1/2}}, \quad (4)$$

where $L_{01} = \left\{\frac{q_0}{q_1}\right\}\left\{\frac{(n_1 - 1)s_1^2}{(n_0 - 1)s_0^2}\right\}^{1/2}\left\{\frac{\Gamma\left(\frac{n_0}{2}\right)\Gamma\left(\frac{n_1 - 1}{2}\right)}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_0 - 1}{2}\right)}\right\}\left\{\frac{n_0(n_1 + 1)}{n_1(n_0 + 1)}\right\}^{1/2}$. As a result, the intensity *z* would be classified into π_0 if the value of odds ratio was larger than 1, and into π_0, otherwise.

4.3 Bayesian Classification With Exponential Distributions

The histograms and normal probability plots of normalized intensities in each probe appeared to have shapes slightly skewed to the right as pointed out previously. Without any transformation, the Bayesian classification, alternatively, could be conducted directly on the original intensities, approximately exhibiting similar characteristics to those in the family of gamma distributions. For convenience in computation, we first assume that the intensities in every probe follow an exponential distribution, $\pi_j \sim \text{Exp}(\beta_j)$ where $j = 0$, 1 and β_j denotes the unknown parameters in class π_j. That is, now assume $(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{n_j})$, are i.i.d. from π_j. Likewise, the likelihood function would be written in term of a sufficient estimator of β_j, given by

$$p(\hat{\beta}_j \,|\, \beta_j, \pi_j) = \frac{(n_j \beta_j)^{n_j}}{\Gamma(n_j)} \frac{1}{\hat{\beta}_j^{(n_j+1)}} e^{-\frac{n_j \beta_j}{\hat{\beta}_j}}, \tag{5}$$

where $\hat{\beta}_j = 1/\bar{\boldsymbol{x}}_j$, the maximum likelihood estimator (MLE) of β_j. We see that the density of $(\hat{\beta}_j \,|\, \beta_j, \pi_j)$ is in the form of an Inverted-gamma $(n_j, n_j \beta_j)$. With the use of vague prior, $p(\beta_j) \propto 1/\beta_j,\ 0 < \beta_j < \infty.$ After replacing the value of $\hat{\beta}_j$ by $1/\bar{\boldsymbol{x}}_j$, the posterior density of β_j could then be derived and given by

$$p(\beta_j / \hat{\beta}_j, \pi_j) = \frac{(n_j \bar{\boldsymbol{x}}_j)^{n_j}}{\Gamma(n_j)} \beta_j^{n_j - 1} e^{-n_j \bar{\boldsymbol{x}}_j \beta_j}, \tag{6}$$

which is in the form of a gamma distribution density. The predictive density of an observation z would be calculated analogously, resulting in

$$p(z \,|\, \hat{\beta}_j, \pi_j) = n_j \frac{(n_j \bar{\boldsymbol{x}}_j)^{n_j}}{(n_j \bar{\boldsymbol{x}}_j + z)^{n_j+1}}, \tag{7}$$

and the posterior odds ratio is given by the ratio of two posterior classification probabilities, shown as:

$$\begin{aligned} \frac{P(z \in \pi_0 \,|\, z)}{P(z \in \pi_1 \,|\, z)} &= \frac{P(z \in \pi_0) \cdot p(z \,|\, \hat{\beta}_0, \pi_0)}{P(z \in \pi_1) \cdot p(z \,|\, \hat{\beta}_1, \pi_1)} \\ &= \frac{q_0 n_0 \cdot \dfrac{(n_0 \bar{\boldsymbol{x}}_0)^{n_0}}{(n_0 \bar{\boldsymbol{x}}_0 + z)^{n_0+1}}}{q_1 n_1 \cdot \dfrac{(n_1 \bar{\boldsymbol{x}}_1)^{n_1}}{(n_1 \bar{\boldsymbol{x}}_1 + z)^{n_1+1}}}, \end{aligned} \tag{8}$$

where q_0 and q_1 were defined previously. Accordingly, the intensity z would be in favor of π_0 if the odds ratio was greater than 1, and vice versa.

4.4 Bayesian Classification With Gamma Distributions

As mention earlier, the intensities in each probe mostly tended to have a distribution skewed to the right, suggesting a gamma distribution. In this section, we generalize the intensities in each probe with a flexible shape parameter, instead of with a fixed shape parameter, as in the exponential case. Similarly, all intensities in class j, $(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{n_j})$, were assumed to be i.i.d.

from π_j, $\pi_j \sim \text{Gamma}(\alpha_j, \beta_j)$. With two unknown parameters and the likelihood function involving a gamma function, this causes difficulties in the derivation of the predictive density. The result is that there is no closed form for the joint posterior density, so we can't develop Bayesian estimates of the two parameters jointly To circumvent this problem, we assumed the shape parameter, α_j, was known and we carried out Bayesian estimation of the resulting scale parameter, β_j for known α_j. To estimate the fixed α_j, we estimated (α_j, β_j) jointly by maximum likelihood (MLE), and then we suppressed the estimate of β_j. We next describe the Bayesian estimation of β_j conditional on α_j.

Given all the data, and the estimated value of α_j $\left(\tilde{\alpha}_j\right)$, the likelihood as a function of the unknown parameter β_j could be written as

$$p(\hat{\beta}_j \left| \tilde{\alpha}_j, \beta_j, \pi_j \right.) = \frac{\left(n_j \tilde{\alpha}_j \beta_j\right)^{n_j \tilde{\alpha}_j}}{\Gamma\left(n_j \tilde{\alpha}_j\right)} \frac{\exp\{-(n_j \tilde{\alpha}_j \beta_j)/\hat{\beta}_j\}}{\hat{\beta}_j^{(n_j \tilde{\alpha}_j + 1)}}, \tag{9}$$

where $\hat{\beta}_j = \tilde{\alpha}_j / \bar{\boldsymbol{x}}_j$ denoted the sufficient estimator of β_j. Suppose that we adopt a vague prior, $p(\beta_j) \propto 1/\beta_j, \ 0 < \beta_j < \infty.$ The resulting posterior distribution of β_j would be obtained with the substitution of $\hat{\beta}_j = \tilde{\alpha}_j / \bar{\boldsymbol{x}}_j$, expressed as

$$p(\beta_j \left| \hat{\beta}_j, \tilde{\alpha}_j, \pi_j \right.) = \frac{\left(n_j \bar{\boldsymbol{x}}_j\right)^{n_j \tilde{\alpha}_j}}{\Gamma\left(n_j \tilde{\alpha}_j\right)} \beta_j^{n_j \tilde{\alpha}_j - 1} e^{-n_j \bar{x}_j \beta_j}, \tag{10}$$

which followed a Gamma$\left(n_j \tilde{\alpha}_j, n_j \bar{\boldsymbol{x}}_j\right)$ distribution. Likewise, the predictive distribution of an observation *z*, afterwards, would be given by

$$p(\boldsymbol{z} \left| \hat{\beta}_j, \tilde{\alpha}_j, \pi_j \right.) = \frac{\boldsymbol{z}^{\tilde{\alpha}_j - 1}}{\boldsymbol{B}(n_j \tilde{\alpha}_j, \tilde{\alpha}_j)} \frac{\left(n_j \bar{\boldsymbol{x}}_j\right)^{n_j \tilde{\alpha}_j}}{\left(n_j \bar{\boldsymbol{x}}_j + \boldsymbol{z}\right)^{\tilde{\alpha}_j (n_j + 1)}}, \tag{11}$$

where $\boldsymbol{B}(n_j \tilde{\alpha}_j, \tilde{\alpha}_j) = \frac{\Gamma\left(n_j \tilde{\alpha}_j\right)\Gamma\left(\tilde{\alpha}_j\right)}{\Gamma\left(n_j \tilde{\alpha}_j + \tilde{\alpha}_j\right)}$. Finally, the posterior odds ratio was established, resulting in

$$\begin{aligned} \frac{P(\boldsymbol{z} \in \pi_0 \left| \boldsymbol{z} \right.)}{P(\boldsymbol{z} \in \pi_1 \left| \boldsymbol{z} \right.)} &= \frac{P(\boldsymbol{z} \in \pi_0) \cdot p(\boldsymbol{z} \left| \hat{\beta}_0, \tilde{\alpha}_0, \pi_0 \right.)}{P(\boldsymbol{z} \in \pi_1) \cdot p(\boldsymbol{z} \left| \hat{\beta}_1, \tilde{\alpha}_1, \pi_1 \right.)} \\ &= \frac{q_0 \cdot \frac{\boldsymbol{z}^{\tilde{\alpha}_0 - 1}}{\boldsymbol{B}(n_0 \tilde{\alpha}_0, \tilde{\alpha}_0)} \frac{\left(n_0 \bar{\boldsymbol{x}}_0\right)^{n_0 \tilde{\alpha}_0}}{\left(n_0 \bar{\boldsymbol{x}}_0 + \boldsymbol{z}\right)^{\tilde{\alpha}_0 (n_0 + 1)}}}{q_1 \cdot \frac{\boldsymbol{z}^{\tilde{\alpha}_1 - 1}}{\boldsymbol{B}(n_1 \tilde{\alpha}_1, \tilde{\alpha}_1)} \frac{\left(n_1 \bar{\boldsymbol{x}}_1\right)^{n_1 \tilde{\alpha}_1}}{\left(n_1 \bar{\boldsymbol{x}}_1 + \boldsymbol{z}\right)^{\tilde{\alpha}_1 (n_1 + 1)}}}, \end{aligned} \tag{12}$$

where q_0 and q_1 were defined as previously. In addition, we also assumed the identical values of shape parameters in both classes, $\tilde{\alpha}_0 = \tilde{\alpha}_1 = \tilde{\alpha}$. The equation (12), then, would be rewritten as

$$\frac{P(z \in \pi_0 | z)}{P(z \in \pi_1 | z)} = \frac{q_0 \cdot \frac{1}{\boldsymbol{B}(n_0\tilde{\alpha}, \tilde{\alpha})} \frac{\left(n_0 \bar{\boldsymbol{x}}_0\right)^{n_0\tilde{\alpha}}}{\left(n_0 \bar{\boldsymbol{x}}_0 + z\right)^{\tilde{\alpha}(n_0+1)}}}{q_1 \cdot \frac{1}{\boldsymbol{B}(n_1\tilde{\alpha}, \tilde{\alpha})} \frac{\left(n_1 \bar{\boldsymbol{x}}_1\right)^{n_1\tilde{\alpha}}}{\left(n_1 \bar{\boldsymbol{x}}_1 + z\right)^{\tilde{\alpha}(n_1+1)}}}, \tag{13}$$

Consequently, the observation *z* would be assigned to π_0 if its odds ratio was larger than 1, and to π_1, otherwise.

4.5 Tree Classification

The classification tree (Venables and Ripley, 1999; Crawley, 2002) was performed based on binary recursive partitioning for which the data were split consecutively along the coordinates of the independent variables. The partition resulted in a path from the top of the tree, called the "root", and continued proceeding to one of the terminal nodes, called a "leaf", following criteria for successive splits. At each splitting point, the threshold for the response variable was chosen and the splitting continued until no further splits were allowed due to sufficient homogeneity of observations, or very small numbers of observations in each node. That is, from the root, the tree splits into 2 groups called "0", and "1" using the intensity value (threshold) as a criterion for splitting. Genes with intensity less than the threshold are placed in group 0, and those with intensity greater than the threshold are placed in group 1. Sometimes the process of splitting ended here and we got the threshold for a clear split. However, sometimes all genes in group 0 and 1 could be split further, as long as the reduction in deviance could still be achieved, or the defaults of the program are not met yet.

In this study, a threshold of intensities for a given probe at each node was selected and the deviances (*D*) of the response above and below this threshold were calculated, as defined by

$$\begin{aligned} D &= -2\sum_i \sum_k n_{ik} \log \hat{p}_{ik} \\ &= -2\sum_i \sum_k n_{ik} \log \frac{n_{ik}}{n_i}, \end{aligned} \tag{14}$$

where n_{ik} and n_i denoted, respectively, the number of clones at the *i*th split of the tree that were assigned to class *k*, where *k* = 0, 1, and the total number of clones in the *i*th split. $\hat{p}_{ik}$ is an estimate of the proportion of clones in node *i* assigned to class *k*. The whole procedure would be repeated until there is no further reduction in deviances or too few data for further subdivision. With tree-based procedures, each intensity in every probe ultimately would be classified into either classes, π_0 or π_1.

Along with the originally assigned data, all binary data obtained from the various classification approaches were then clustered using the computer program, UPGMA in PAUP. The analyses could also be performed with the presence of unknown values (N) in the original classification. With the default parameters, UPGMA assigned the unknown values proportionally to the known 0-1 values that appeared in the pairs of clones in comparison, as pointed out previously.

The pairwise-distance matrices based on the proportions of different, or mismatched, characters between pairs of clones were constructed, and the hierarchical clusters, or *trees*, would be

formed afterwards. With the same clone-probe combination, all trees drawn from different binary data corresponding to each classification approach were eventually compared to the reference tree obtained by analogously applying the UPGMA on the binary reference data.

Tree comparisons in this study were performed based on several different criteria:

1) the agreement subtrees ("Agree"), (Swofford, 2001);
2) symmetric-difference metric ("SymDiff"), (Penny and Hendy, 1985);
3) agreement-subtree metric d (AgD), (Goddard, *et al*., 1994:);
4) agreement-subtree metric d1 ("AgD1"), (Goddard, *et al*., 1994:).

With the Agree criterion, the number of conformed subtrees between pairs of trees in comparison would be counted. As a result, the higher value of Agree exhibited more agreement between both trees. According to SymDiff, the difference between two trees was measured by the number of edges (links) that produced no equivalent edges on the other trees, or in other words, no identical subtrees. The AgD1 was formed by counting the number of leaves (or clones) which had to be pruned from both trees to obtain a common substructure. However, AgD1 did not account for the location of pruned leaves and it was modified to the AgD by adding a fraction representing how far apart of pruned leaves in the common subtrees. Accordingly, the higher values of SymDiff, AgD and AgD1, the less agreement of trees in comparison.

The classification approach that outperformed all others yields small error rates of classification and indicates large agreement with the reference tree in the cluster analysis performed on the binary data produced from that classification approach. The procedure exhibiting this improvement over the original classification was finally adopted to classify the entire 1,464 clones.

5. RESULTS

With the same clone-probe combinations, all confusion matrices obtained from Bayesian classification approaches and tree-based classification, as described in the previous section, were established separately in comparison with the classification of reference data, as shown in Tables 1a-1e.

The performances of the various classification approaches conducted on the reference data were evaluated from the apparent error rate corresponding to each confusion matrix, as provided in Table 2. Most Bayesian classification methods performed on a family of Gamma distribution resulted in high percentages of misclassification. Between two classification procedures performed on the Exponential distribution and the Gamma distribution using MLE, the errors of misclassification tended to be smallest with the use of MLE to approximate the shape parameter. Using the transformed-to-normality signal intensities, the Bayesian classification exhibited a moderate rate of misclassification. The tree-based classification produced the smallest error rate of classification in this study. But as we'll see, classification error rate was only part of the story. We were ultimately concerned with how the tree results compared for the various approaches. That is, "How far from 'the reference tree' (the tree corresponding to the known classifications

and clusters) were the trees corresponding to the various methods of classification?" We show this comparison in Table 3.

In addition to the percentages of misclassification, the performance of various classification approaches could also be assessed from clustering results. All trees constructed from varied binary data with respect to each classification approach, together with that of the originally assigned data, were compared to the reference tree. As shown in Table 3, the results of this comparison based on four criteria appeared to be consistent across all classification approaches. The tree from Bayesian classification using normal densities tended to exhibit the most agreement with the reference tree, followed by those from Bayesian classifications on exponential and gamma densities, roughly producing the same level of agreement. The original approach (0-1-N classification with proportional allocation to clusters) produced a tree with the second smallest level of agreement. Among all classification methods considered here, the tree-based classification indicated the least agreement of sub-trees although it yielded the minimum rate of misclassification. One possible explanation for the poor performance of tree-based classification was possibly due to the fact that while such classification was associated with the smallest classification error rate, because it is non-parametric and therefore does not take the form of the data distribution into account when developing classification criteria, it may miss the characteristics of the data that are most relevant to tree distances. As a consequence, with a moderate rate of misclassification, and the most agreement of subtrees with the reference tree, the Bayesian classification with normal densities approach out-performed all others, including an improvement over the (0-1-N) original approach. The tree that resulted from Bayesian normal classification of the reference data, along with the reference tree, are illustrated in Figure 1 and Figure 2, respectively. Finally, this Bayesian normal classification approach was adopted to classify all the transformed signal intensities in the set of 1,464 clones. The tree that resulted from this relatively optimal approach applied to all of the 1464 clones is not displayed in this paper since it is required approximately 14 pages for presentation. Accordingly, we decided not to include it in the paper.

6. CONCLUSIONS

In this paper, we compared and evaluated several strategies for transforming hybridization signal intensity data from oligonucleotide fingerprint experiments into hybridization fingerprints composed of binary values, thereby resolving uncertainty in the originally assigned fingerprints. The performance of each classification approach was assessed in terms of both misclassification rates and the levels of agreement in the hierarchical clusters. The analysis revealed that Bayesian classification on transformed-to-normal data appeared to out-perform all others considered in the study, including the original (0-1-N) approach. However, we were still limited in that there were just a very small number of clones falling in either the hybridized or the un-hybridized group in some probes of the training data whose class memberships were established. This surely affected the performance of all statistics used in all of the Bayesian classification approaches. But most statistics have reasonably good properties as the sample size gets large. As a consequence, it is recommended that in applications, a larger training data set be used, where possible, to gain maximum advantage from the training data. Overall, utilization of the Bayesian classification scheme should increase the reliability of oligonucleotide fingerprint analyses.

This Bayesian approach may also be useful for other studies including analysis of standard gene expression data. Typical goals of gene expression analysis include identifying genes that are expressed at similar/differential levels or identifying samples that have similar/different expression patterns. These studies typically utilized cluster analyses, which group objects by their similarities. One problem with these approaches is defining what constitutes similarity, as the results of any clustering experiment can be strongly influenced by how this parameter is defined (Brazma and Vilo, 2000). One way to approach this problem is to discretize the data before it is clustered. Shmulevich and Zhang developed an approach which included discretizing gene expression data into a binary format (Shmulevich and Zhang, 2002). This approach successfully separated tumor types while reducing data noise and increasing computational efficiency. Utilization of the Bayesian approach described in this manuscript will provide an alternative approach for discretizing expression data based on prior knowledge, which could lead to new strategies for gene expression analysis.

REFERENCES

Amann, R.I., Ludwig, W. and Schleifer, K. H. 1995. Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiol. Rev*. 59, 143-169.

Atlas, R.M., and Bartha, R. 1998. *Microbial Ecology Fundamental and Applications*. Benjamin/Cummings.

Borneman, J., Chrobak, M., Della Vedova, G., Figueroa, A., and Jiang, T. 2001. Probe selection algorithms with applications in the analysis of microbial communities. *Bioinformatics*. 17**,** S39-48, supplement 1.

Borneman, J., and Triplett, E.W. 1997. Molecular microbial diversity in soils from eastern Amazonia: Evidence for unusual microorganisms and microbial population shifts associated with deforestation. *Appl. Environ. Microbiol*. 63, 2647-2653.

Box, G.E.P. and Cox, D.R. 1964. An analysis of transformations. *J. R. Statist Soc. B*, 26, 211-252.

Brazma, A., and Vilo, J. 2000. Gene expression data analysis. *FEBS Lett*. 480, 17-24.

Bull, A.T., Ward, A.C., and Goodfellow, M. 2000. Search and discovery strategies for biotechnology: The paradigm shift. *Microbiol. Mol. Biol. Rev*. 64**,** 573-606.

Chapela, I.H. 1997. Bioprospecting: Myths, Realities and Potential Impact on Sustainable Development, 238-256. In Palm, M.E. and Chapela, I.H., eds., *Mycology in sustainable development: Expanding concepts, vanishing borders*, Parkway Publishers, Boone, NC.

Christensen, M. 1989. A view of fungal ecology. *Mycologia*. 81**,** 1-19.

Crawley, M.J. 2002. *Statistical Computing: An Introduction to Data Analysis Using S-Plus*. John Wiley, West Sussex, UK.

Drmanac, R. 1999. cDNA screening by array hybridization. Methods Enzymol. **303**:165-178.

Drmanac, R., G. Lennon, S. Drmanac, I. Labat, R. Crkvenjakov, and H. Lehrach. 1991. p. 60-75. *In* C. Cantor, and H. Lim (ed.), Proceedings of the First International Conference on Electrophoresis, Supercomputing and the Human Genome. World Scientific, Singapore.

Fox, G.E., Pechman, K.R. and Woese, C.R. 1977. Comparative cataloging of 16S ribosomal ribonucleic acid: Molecular approach to prokaryotic systematics. *Int. J. Syst. Bacteriol*. 27, 44-57.

Giovannoni, S.J., Britschgi, T.B., Moyer, C.L., and Field, K.G. 1990. Genetic diversity in Sargasso Sea bacterioplankton. *Nature.* 345, 60-63.

Goddard, W., Kubicka, E., Kubicki, G. and McMorris, R. 1994. The Agreement Metric for Labeled Binary Trees. *Math. Biosci*. 123, 215-226.

Johnson, R.A. and Wichern, D.W. 1992. *Applied Multivariate Statistical Analysis*, 3rd ed., Prentice Hall, New Jersey.

Lennon, G. S., and H. Lehrach. 1991. Hybridization analyses of arrayed complementary DNA libraries. Trends Genet. **7**:314-317.

Liu, W.T., Marsh, T.L., Cheng, H. and Forney, L.J. 1997. Characterization of microbial diversity by determining terminal restriction fragment length polymorphisms of genes encoding 16S rRNA. *Appl. Environ. Microbiol*. 63, 4516-4522.

Maddison, D.R., Swofford, D.L. and Maddison, W.P. 1997. Nexus: An extensible file format for systematic information. *Syst. Biol*. 46, 590-621.

Muyzer, G., De Waal, E.D., Uitterlinden, A.G. 1993. Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction amplified genes coding for 16S rRNA. *Appl. Environ. Microbiol.* 59, 695-700.

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

Penny, D. and Hendy, M.D. 1985. The use of tree comparison metrics. *Syst. Zool.* 34(1), 75-82.

Press, S.J. 1982. *Applied Multivariate Analysis: Including Bayesian and Frequentist Methods of Inference*. Krieger Publishing Co, Malabar, Florida.

Press, S.J. 1989. *Bayesian Statistics: Principles, Models and Applications*. John Wiley, New York.

Press, S.J. 2003. *Subjective and Objective Bayesian Statistics: Principles, Models, and Applications*, 2nd ed., John Wiley, New York.

SAS Institute. 1994. *SAS/STAT User's Guide: Volume 1*, 4th ed., SAS Institute Inc, Cary, North Carolina.

Shmulevich, I., and Zhang, W. 2002. Binary analysis and optimization-based normalization of gene expression data. *Bioinformatics*. 18, 555-565.

Sogin, S.J., Sogin, M.L., and Woese, C.R. 1972. Phylogenetic measurement in prokaryotes by primary structural characterization. *J. Mol. Evol*. 1, 173-184.

Swofford, D.L. 2001. *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods) Version 4.0*. Sinauer Associates, Sunderland, Massachusetts.

Tuomela, M., Vikman, M., Hatakka, A., and Itavaara, M. 2000. Biodegradation of lignin in a compost environment: A review. *Bioresource Technol*. 72**,** 169-183.

Valinsky, L., Della Vedova, G., Jiang, T., and Borneman, J. 2002. Oligonucleotide fingerprinting of ribosomal RNA genes for analysis of fungal community composition. *Appl. Environ. Microbiol*. 68**,** 5999-6004.

Valinsky, L., Della Vedova, G., Scupham, A.J., Alvey, S., Figueroa, A., Yin, B., Hartin, J., Chrobak, M., Crowley, D.E., Jiang, T., and Borneman, J. 2002. Analysis of bacterial community composition by oligonucleotide fingerprinting of rRNA genes. *Appl. Environ. Microbiol*. 68, 3243-3250.

Vaneechoutte, M., Rossau, R., Devos, P., Gillis, M., Janssens, D., Paepe, N., Derouck, A., Fiers, T., Claeys, G., and Kersters, K. 1992. Rapid identification of bacteria of the comamonadaceae with amplified ribosomal DNA-restriction analysis (ARDRA). *FEMS Microbiology Letters*. 93, 227-234.

Venables, W.N. and Ripley, B.D. 1999. *Modern Applied Statistics with S-Plus*, 3rd ed., Springer-Verlag, New York.

Woese, C.R., and Fox, G.E. 1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc. Natl. Acad. Sci*. *USA* 74, 5088-5090.

Woese, C.R., Fox, G.E., Zablen, L., Urchida, T., Bonen, L., Pechman, K., Lewis, B. J., and Stahl, D. 1975. Conservation of primary structure in 16S ribosomal RNA. *Nature.* 254, 83-86.

Confusion Matrices of Each Classification Approach Compared with Reference Data

Normal Bayesian Classification	**Actual Class**		Total
	0	1	
0	1087	21	1108
1	44	603	647
Total	1131	624	1755

Table 1a

Exponential Bayesian Classification	**Actual Class**		Total
	0	1	
0	1066	25	1091
1	65	599	664
Total	1131	624	1755

Table 1b

Gamma Bayesian Classification	**Actual Class**		Total
	0	1	
0	1067	25	1092
1	64	599	663
Total	1131	624	1755

Table 1c

Tree Classification	**Actual Class**		Total
	0	1	
0	1098	16	1114
1	33	608	641
Total	1131	624	1755

Table 1d

Approach	**APER (%)**
Bayesian classification (Normal)	3.70
Bayesian classification (Exponential)	5.13
Bayesian classification (Gamma) with MLE	5.07
Tree-based classification	2.79

Table 2: The Apparent Error Rates From Each Classification Approach

Criteria	Original	Bayesian (Normal)	Bayesian (Exponential)	Bayesian (Gamma)	Tree
Agree	39/65	43/65	42/65	42/65	27/65
SymDiff	60	48	52	52	66
AgD1	26	22	23	23	38
AgD	26.0917	22.0664	23.0696	23.0696	38.0672

Table 3

Tree Comparison Based on Four Criteria.

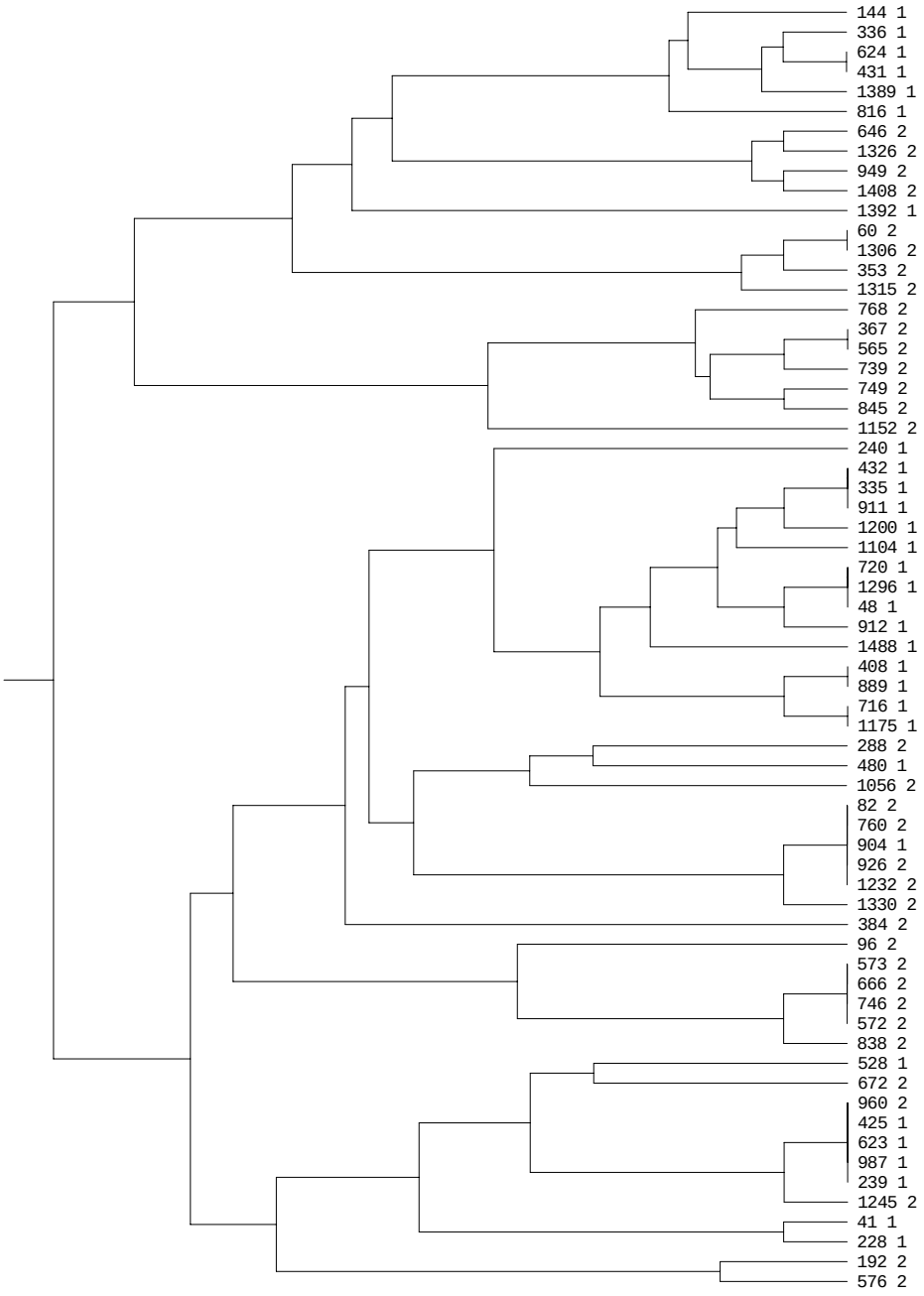

Figure 1

Hierarchical Clustering of Reference Data Obtained From Normal Bayesian Classification

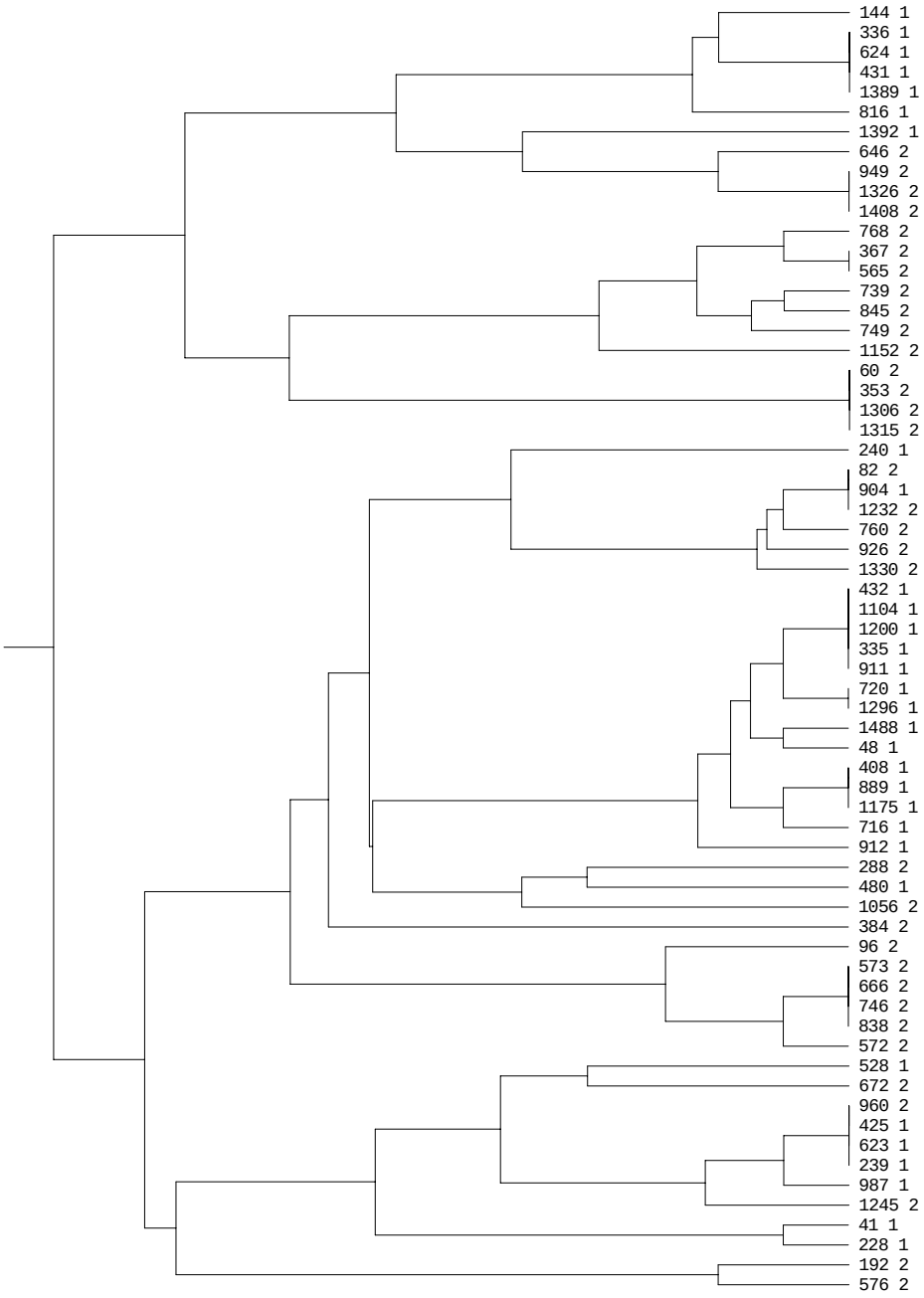

Figure 2

Reference Tree