Genetics and population analysis

Classification of oligonucleotide fingerprints: application for microbial community and gene expression analyses

Katechan Jampachaisri\textsuperscript{1}, Lea Valinsky\textsuperscript{3}, James Borneman\textsuperscript{2} and S. James Press\textsuperscript{1,*}

\textsuperscript{1}Department of Statistics and \textsuperscript{2}Department of Plant Pathology, University of California, Riverside, CA 92521, USA and \textsuperscript{3}Central Laboratories, Israeli Ministry of Health, Yaakov Eliav 9, 94467 Jerusalem, Israel

Received on February 3, 2005; revised and accepted on April 13, 2005
Advance Access publication April 21, 2005

\textbf{ABSTRACT}

\textbf{Motivation:} Oligonucleotide fingerprinting of ribosomal RNA genes (OFRG) is a procedure that sorts rRNA gene (rDNA) clones into taxonomic groups through a series of hybridization experiments. The hybridization signals are classified 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. This study examined various approaches for classifying the values including Bayesian classification with normally distributed signal data, Bayesian classification with the exponentially distributed data, and with gamma distributed data, along with tree-based classification. All classification data were clustered using the unweighted pair group method with arithmetic mean.

\textbf{Results:} The performance of each classification/clustering procedure was compared with results from known reference data. Comparisons indicated that the approach using the Bayesian classification with normal densities followed by tree clustering out-performed all others. The paper includes a discussion of how this Bayesian approach may be useful for the analysis of gene expression data.

\textbf{Contact:} james.press@ucr.edu

\section{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 great number of important food and beverage products. In the biotechnology arena, they are a vital source of useful compounds and provide the means 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

\textsuperscript{*}To whom correspondence should be addressed.

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).

\section{2 THE PROBLEM}

Numerous rDNA-based strategies have been developed for microbial community analysis. The most accurate approach is to analyze the nucleotide sequence of rRNA molecules or genes (Giovannoni et al., 1990; Olsen et al., 1986; Pace et al., 1986; Ward et al., 1990). However, because of the high costs associated with examining such diverse communities in this manner, this approach is usually impractical for thorough analysis of microbial community composition. Other methods such as denaturing gradient gel electrophoresis (Muyzer et al., 1993), terminal restriction fragment length polymorphisms (Liu et al., 1997), ribosomal intergenic space analysis (Borneman and Triplett, 1997) and amplified ribosomal DNA restriction analysis (Vanechoute et al., 1992) enable relatively inexpensive and rapid analysis of many samples, but they typically generate only partial 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 the experimental process for OFRG follows. Microbial rDNAs 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 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. To date, we have developed OFRG probe sets for bacteria and fungi. The bacterial analysis is capable of classifying rRNA gene sequences to the species level (Valinsky et al., 2002b). For fungal analysis, OFRG was able to resolve clones with average sequence identities of 99.2% (Valinsky et al., 2002a).

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 to the rDNA clones should produce larger signal intensity values than those that do not hybridize. The 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 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 controls are higher than those from the negative controls. For probes that function in this manner, clones with intensity values \( y \) were given a 1 classification, where \( y \) is the lowest intensity value generated by a negative control. Clones with intensity values \( x \leq y \) were given a 0 classification, where \( x \) is the highest intensity value generated by a positive control. All other clones were given an N classification.

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

<table>
<thead>
<tr>
<th>Gene 1</th>
<th>Gene 2</th>
<th>Gene 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

The distance between these genes is obtained easily by computing the proportion of mismatched characters between them. There are four mismatches, so the ‘distance’ between them is: 4/10 = 0.4.

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

4 STATISTICAL ANALYSIS

4.1 Overview

Our data consist of signal intensities from hybridization experiments between arrayed rDNA 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 small, and we had no idea how to classify the ‘N’ combinations. (However, we felt confident 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 classified 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

<table>
<thead>
<tr>
<th>Gene 1-2</th>
<th>Gene 1-3</th>
<th>Gene 2-3</th>
</tr>
</thead>
<tbody>
<tr>
<td>0000100010</td>
<td>0000100010</td>
<td>0000100010</td>
</tr>
<tr>
<td>1001110110</td>
<td>1001110110</td>
<td>1001110110</td>
</tr>
<tr>
<td>0000100010</td>
<td>0000100010</td>
<td>0000100010</td>
</tr>
<tr>
<td>1001110110</td>
<td>1001110110</td>
<td>1001110110</td>
</tr>
<tr>
<td>0000100010</td>
<td>0000100010</td>
<td>0000100010</td>
</tr>
<tr>
<td>1001110110</td>
<td>1001110110</td>
<td>1001110110</td>
</tr>
<tr>
<td>0000100010</td>
<td>0000100010</td>
<td>0000100010</td>
</tr>
<tr>
<td>1001110110</td>
<td>1001110110</td>
<td>1001110110</td>
</tr>
<tr>
<td>0000100010</td>
<td>0000100010</td>
<td>0000100010</td>
</tr>
<tr>
<td>1001110110</td>
<td>1001110110</td>
<td>1001110110</td>
</tr>
</tbody>
</table>

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 gene 3 would be 0.333 (3.33/10).

The distance between gene 1 and 3 can be performed similarly by considering the proportion of mismatched pairs occurring in the nine
Table 1. Symmetric pairwise-distance matrix for the test dataset

<table>
<thead>
<tr>
<th></th>
<th>Gene 1</th>
<th>Gene 2</th>
<th>Gene 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene 1</td>
<td>—</td>
<td>0.400</td>
<td>0.444</td>
</tr>
<tr>
<td>Gene 2</td>
<td>0.333</td>
<td>—</td>
<td>0.333</td>
</tr>
<tr>
<td>Gene 3</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

unambiguous positions, which is $4/9 = 0.44$. So, out of 10 probes there are 4.44 mismatches. The distance between gene 1 and gene 3, accordingly, would be obtained as $0.444 (4.44/10)$.

The symmetric pairwise-distance matrix for this example is given in Table 1, 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 to confirm 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 nucleotide sequences of the rDNA clones to evaluate how well the statistical classification/clustering procedure would compare 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 APER did not depend on the form of population densities and could be readily calculated by constructing the $2 \times 2$ ‘confusion matrix’ (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 (Venables 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 (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. 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 (Venables 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 (Press, 1982, 1989, 2003).

The symmetric pairwise-distance matrix for this example is given in Table 1, 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 to confirm 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 nucleotide sequences of the rDNA clones to evaluate how well the statistical classification/clustering procedure would compare 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 APER did not depend on the form of population densities and could be readily calculated by constructing the $2 \times 2$ ‘confusion matrix’ (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 (Venables 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 (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. 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 (Venables 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 (Press, 1982, 1989, 2003).

The symmetric pairwise-distance matrix for this example is given in Table 1, 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 to confirm 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 nucleotide sequences of the rDNA clones to evaluate how well the statistical classification/clustering procedure would compare 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 APER did not depend on the form of population densities and could be readily calculated by constructing the $2 \times 2$ ‘confusion matrix’ (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 (Venables 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 (Press, 1982, 1989, 2003).

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. 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 (Venables 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 (Press, 1982, 1989, 2003).
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

$$p(\mu_j, \sigma_j^2|\bar{x}_j, s_j^2, \pi_j) \propto \frac{1}{(\sigma_j^2)^{n_j/2}\Gamma(n_j)} e^{-1/2\sigma_j^2[s_j(\mu_j-\bar{x}_j)^2+(\sigma_j^{-1}+n_j)^2]}.$$  

(2)

The interest here was to predict class membership of an observation $z$, which is known to belong to either $\pi_0$ or $\pi_1$. 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

$$p(z|\bar{x}_j, s_j^2, \pi_j) \propto \frac{1}{\left[1 + \frac{1}{n_j^0} \left(\frac{z - \bar{x}_j}{s_j}\right)^2\right]^{n_j/2}}.$$  

(3)

Let $q_0$ and $q_1$ be prior probabilities that $z$ came from $\pi_0$ and $\pi_1$, 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

$$\frac{P(z \in \pi_1|z)}{P(z \in \pi_0|z)} = \frac{P(z \in \pi_1) \cdot p(z|\bar{x}_j, s_j^2, \pi_1)}{P(z \in \pi_0) \cdot p(z|\bar{x}_j, s_j^2, \pi_0)}$$

$$L_{01} = \left\{ \begin{array}{c} \frac{q_0}{q_1} \left(\frac{(n_1 - 1)s_j^2}{(n_0 - 1)\bar{x}_j^2}\right)^{1/2} \left\{ \Gamma\left(\frac{n_1}{2}\right) \frac{1}{\Gamma\left(\frac{n_1 - 1}{2}\right)} \right\} \\
\frac{n_0(n_1 + 1)}{n_1(n_0 + 1)} \end{array} \right\}^{1/2}.$$  

(4)

where

$$L_{01} = \left\{ \begin{array}{c} \frac{q_0}{q_1} \left(\frac{(n_1 - 1)s_j^2}{(n_0 - 1)\bar{x}_j^2}\right)^{1/2} \left\{ \Gamma\left(\frac{n_1}{2}\right) \frac{1}{\Gamma\left(\frac{n_1 - 1}{2}\right)} \right\} \\
\frac{n_0(n_1 + 1)}{n_1(n_0 + 1)} \end{array} \right\}^{1/2}.$$  

As a result, the intensity $z$ would be classified into $\pi_1$ if the value of odds ratio was larger than 1, and into $\pi_0$, 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, $\pi_j \sim \exp(\beta_j)$, where $j = 0, 1$ and $\beta_j$ denotes the unknown parameters in class $\pi_j$. That is, now assume $(x_1, x_2, ..., x_{n_j})$ are i.i.d. from $\pi_j$. Likewise, the likelihood function would be written in terms of a sufficient estimator of $\beta_j$, given by

$$p(\hat{\beta}_j|\beta_j, \pi_j) = \frac{(n_j\beta_j)^{n_j/2}}{\Gamma(n_j)} \frac{1}{\beta_j^{n_j+1}} e^{-n_j\beta_j/\beta_j},$$  

(5)

where $\hat{\beta}_j = 1/\bar{x}_j$, the maximum-likelihood estimator (MLE) of $\beta_j$. 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 $1/\bar{x}_j$, the posterior density of $\beta_j$ could then be derived and given by

$$p(\beta_j|\hat{\beta}_j, \pi_j) = \frac{(n_j\hat{\beta}_j)^{n_j/2}}{\Gamma(n_j)} \frac{1}{\beta_j^{n_j+1}} e^{-n_j\hat{\beta}_j/\beta_j},$$  

(6)

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

$$p(z|\hat{\beta}_j, \pi_j) = n_j \cdot (1 + z/\hat{\beta}_j)^{n_j-1},$$  

(7)

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

$$\frac{P(z \in \pi_0|z)}{P(z \in \pi_1|z)} = \frac{P(z \in \pi_0) \cdot p(z|\hat{\beta}_0, \pi_0)}{P(z \in \pi_1) \cdot p(z|\hat{\beta}_1, \pi_1)}$$

$$= \frac{q_0 q_1 \cdot \Gamma\left(\frac{n_0}{2}\right)}{q_1 q_1 \cdot \Gamma\left(\frac{n_1}{2}\right)} \left\{ \frac{\Gamma\left(\frac{n_0}{2}\right)}{\Gamma\left(\frac{n_1}{2}\right)} \right\} \left\{ \frac{\Gamma\left(\frac{n_0 - 1}{2}\right)}{\Gamma\left(\frac{n_1 - 1}{2}\right)} \right\}^{1/2},$$  

(8)

where $q_0$ and $q_1$ were defined previously. Accordingly, the intensity $z$ would be in favor of $\pi_0$ if the odds ratio was $> 1$, and vice versa.

### 4.4 Bayesian classification with gamma distributions

As mentioned 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 $j$, $(x_1, x_2, ..., x_{n_j})$, were assumed to be i.i.d. from $\pi_j, \pi_j \sim \Gamma(a_j, \beta_j)$. 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 cannot develop Bayesian estimates of the two parameters jointly. To circumvent this problem, we assumed the shape parameter, $a_j$, was known, and we carried out Bayesian estimation of the resulting scale parameter, $\beta_j$, for known $a_j$. To estimate the fixed $a_j$, we estimated $(a_j, \beta_j)$ jointly by MLE, and then we suppressed the estimate of $\beta_j$. We next describe the Bayesian estimation of $\beta_j$ conditional on $a_j$.

Given all the data, and the estimated value of $a_j$, the likelihood function as a function of the unknown parameter $\beta_j$ could be written as

$$p(\hat{\beta}_j|\hat{\beta}_j, a_j, \pi_j) = \frac{(n_j\hat{\beta}_j)^{n_j/2}}{\Gamma(n_j)} \frac{1}{\beta_j^{n_j+1}} e^{-n_j\hat{\beta}_j/\beta_j},$$  

(9)

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

$$p(\beta_j | \tilde{\alpha}_j, \bar{x}_j, \pi_j) = \frac{\bar{x}_j^\beta_j}{B(n_j, \tilde{\alpha}_j)} \left( \frac{\tilde{\alpha}_j + 1}{\tilde{\alpha}_j} \right)^\beta_j, \quad (10)$$

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

$$p(z | \beta_j, \tilde{\alpha}_j, \pi_j) = \frac{z^{\tilde{\alpha}_j-1} \left( n_j \bar{x}_j \right)^{\bar{x}_j} \beta_j}{B(n_j, \tilde{\alpha}_j)} \left( \frac{\bar{x}_j + z}{\tilde{\alpha}_j} \right)^{\bar{x}_j}, \quad (11)$$

where

$$B(n_j, \tilde{\alpha}_j) = \frac{\Gamma(n_j) \Gamma(\tilde{\alpha}_j)}{\Gamma(n_j + \tilde{\alpha}_j)}.$$ 

Finally, the posterior odds ratio was established, resulting in

$$P(z \in \pi_0 | z) = \frac{P(z \in \pi_0) \cdot p(z | \beta_0, \tilde{\alpha}_0, \pi_0)}{P(z \in \pi_1) \cdot p(z | \beta_1, \tilde{\alpha}_1, \pi_1)}, \quad (12)$$

$$= \frac{q_0 \cdot \frac{1}{\beta_0} \left( n_0 \tilde{\alpha}_0 \right)^{\beta_0} \tilde{\alpha}_0^{\beta_0-1}}{q_1 \cdot \frac{1}{\beta_1} \left( n_1 \tilde{\alpha}_1 \right)^{\beta_1} \tilde{\alpha}_1^{\beta_1-1}}, \quad (13)$$

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

$$P(z \in \pi_0 | z) = \frac{q_0 \cdot \frac{1}{\beta_0} \left( n_0 \tilde{\alpha}_0 \right)^{\beta_0} \tilde{\alpha}_0^{\beta_0-1}}{q_1 \cdot \frac{1}{\beta_1} \left( n_1 \tilde{\alpha}_1 \right)^{\beta_1} \tilde{\alpha}_1^{\beta_1-1}}.$$ 

Consequently, the observation $z$ would be assigned to $\pi_0$ if its odds ratio was larger than 1, and to $\pi_1$, otherwise.

### 4.5 Tree classification

The classification tree (Venebles 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 owing to sufficient homogeneity of observations, or very small numbers of observations in each node. That is, from the root, the tree splits into two 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 groups 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 yet met.

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

$$D = -2 \sum_i \sum_k n_{ik} \log \hat{p}_{ik}, \quad (14)$$

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 class $\pi_0$ or class $\pi_1$.

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 with 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 $d_1$ (‘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) that 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 the difference of pruned leaves in the common subtrees. Accordingly, the higher the values of SymDiff, AgD and AgD1, the lesser agreement of trees in comparison.

The classification approach that out-performed 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 all the 1464 clones.
The performances of the various classification approaches conducted on the reference data were evaluated from the APER corresponding to each confusion matrix, as provided in Table 3. Most Bayesian classification methods performed on a family of gamma distributions 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 an 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 see, classification error rate was only part of the story. We were ultimately concerned with how the tree results compared for the various approaches, i.e. ‘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 4.

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 with the reference tree. As shown in Table 4, the results of this comparison based on four criteria appeared to be consistent across all classification approaches. The tree from the Bayesian classification using normal densities tended to exhibit the most agreement with the reference tree, followed by those from the 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 subtrees 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 a 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 outperformed all others, including an improvement over the (0-1-N) original approach. The tree that resulted from the Bayesian normal classification of the reference data along with the reference tree are illustrated in Figures 1 and 2, respectively. Finally, this Bayesian normal classification approach was adopted to classify all the transformed signal intensities in the set of 1464 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 required ~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 the 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 unhybridized group in some
Fig. 1. Hierarchical clustering of reference data obtained from normal Bayesian classification.

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 utilize 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,
Classification of oligonucleotide fingerprints

Fig. 2. Reference tree.

2000). One way to approach this problem is to discretize the data before it is clustered. Shmulevich and Zhang (2002) developed an approach which included discretizing gene expression data into a binary format. This approach successfully separated tumor types while reducing data noise and increasing computational efficiency. Utilization of the Bayesian approach described in this paper will provide an alternative approach for discretizing expression data based on prior knowledge, which could lead to new strategies for gene expression analysis.

ACKNOWLEDGEMENTS

We wish to thank Professor Tao Jiang and Andres Figueroa (Department of Computer Science and Engineering), Professor Mark S. Springer (Department of Biology, University of California, Riverside, CA) for their useful advice about tree comparisons. This work was funded in part by grants from the NSF BDI Program (J.B. and S.J.P.) and the UC Biotechnology Research & Training programme (K.J., J.B. and S.J.P.). L.V. was supported by Vaddia-BARD postdoctoral award FI-306-00 from BARD,
The United States-Israel Binational Agricultural Research and Development Fund. K.J. was supported in part by a graduate student fellowship from the Government of Thailand.

REFERENCES


Benjamin/Cummings.


3130