Mixture Models as a Method to Find Present and Divergent Genes in Comparative Genomic Hybridization Studies on Bacteria

Guri Feten*, Trygve Almøy, Lars Snipen, Ågot Aakra, O. Ludvig Nyquist, and Are H. Aastveit
Department of Chemistry, Biotechnology and Food Science, Norwegian University of Life Sciences, P.O. Box 5003, N-1432 Ås, Norway

Received 16 November 2005, revised 3 May 2006, accepted 3 June 2006

Summary
Comparative genomic hybridization (CGH) using microarrays is performed on bacteria in order to test for genomic diversity within various bacterial species. The microarrays used for CGH are based on the genome of a fully sequenced bacterium strain, denoted reference strain. Labelled DNA fragments from a sample strain of interest and from the reference strain are hybridized to the array. Based on the obtained ratio intensities and the total intensities of the signals, each gene is classified as either present (one copy or multiple copies) or divergent (zero copies).

In this paper mixture models with different number of components are fitted on different combinations of variables and compared with each other. The study shows that mixture models fitted on both the ratio intensities and the total intensities including the replicates for each gene improve, compared to previously published methods, the results for several of the data sets tested. Some summaries of the data sets are proposed as a guide for the choice of model and the choice of number of components.

The models are applied on data from CGH experiments with the bacteria *Staphylococcus aureus* and *Streptococcus pneumoniae*.

Key words: Analysis of variance; Comparative genomic hybridization; Microarray experiment; Mixture models; Validation criteria.

1 Introduction
Comparative genomic hybridization (CGH) using microarrays is performed on bacteria in order to test for genomic diversity within various bacterial species (Behr et al., 1999; Björkholm et al., 2001). The microarrays used for CGH are based on the genome of a fully sequenced bacterium strain, denoted reference strain. Prior to hybridization, genomic DNA from the reference strain and a sample strain is extracted and digested to smaller fragments. Fragments from the reference strain and the sample strain are then labelled with different fluorescent dyes, and hybridized to the array. The procedure of scanning and image analysis can be found in e.g. Nguyen et al. (2002). From the obtained ratio intensities and total intensities of the signals, each gene should be classified as either present (common between the strains) or divergent (specific to the reference strain).

Comparative genomic hybridization has also become a useful tool for prediction of gene amplifications and deletions in different types of cancer (Pollack et al., 1999, 2002; Veltman et al., 2003; Peng et al., 2003). The changes in the number of copies of each gene are important for the understanding of cancer and its diagnosis.

* Corresponding author: e-mail: guri.feten@umb.no, Phone: +47 64 96 58 28, Fax: +47 64 96 59 01
Studies of different methods to divide genes into groups of present (one copy or multiple copies) and divergent (zero copies) have been carried out several times. One method is based on a two-sample t-statistic where sample-reference and reference-reference CGH profiles are compared at each point along the genome to detect regions of significant differences (Moore et al., 1997). A constant cut off value could be chosen based on dividing some known (based on genome sequence) present and divergent genes into their correct group (Salama et al., 2000; Murray et al., 2001; Dziejman et al., 2002). This method demands a strain with known deletions, which is not necessarily available for every organism. A more robust method for deciding the cut off is based on the distribution of the signal ratio, where each hybridization data set gets an independent cut off (Kim et al., 2002). Repsilber et al. (2005) proposed a method that ranked the genes based on the average ratio intensities after a rotation of the data, where the rotation makes use of the total intensities of the signal.

A refinement of dividing the genes into groups of divergent and present is to divide the genes into the three groups divergent, present and present in multiple copies. A method using k-means clustering and dynamic programming have been suggested (Autio et al., 2003). Irizarry et al. (2003) used a method based on mixture models and weighted z-tests to detect differentially represented yeast mutants. Mixture models have previously been used to identify differentially expressed genes in gene expression studies (Lee et al., 2000).

A further refinement is to divide the genes into groups with equal number of copies. Different methods, such as a likelihood-based approach (Carothers, 1997), an unsupervised Hidden Markov Models approach (Fridlyand et al., 2003), regression (Feten et al., 2006), and standard t-statistic and a modification with variances smoothed along the genome (Wang and Guo, 2004), have been proposed.

In this paper a method that applies mixture models assuming a normal distribution of the ratio intensities and the total intensities is suggested in the purpose of dividing the genes into groups of present and divergent. Mixture models with different number of components are fitted to the data. For each gene the posterior probability of being present and divergent is estimated, and the genes are assigned to the group that maximizes their posterior probability. The aim is to find the minimal number of genes common to both the reference and the sample strain. The models are proposed on data from CGH experiments with the bacteria *Staphylococcus aureus* and *Streptococcus pneumoniae*. Some statistics are studied as a guide to the selection of model and number of components. Finally, analysis of variance and tests with multiple comparisons are used to test for differences both between the models, between the number of components and between the data sets.

In Section 2 the mixture models and two validation criteria are presented, in addition, an analysis of variance model for the two criteria is given. An introduction to our microarray data, both biological and statistical aspects, is found in Section 3, where the models are applied on the current microarray data. Finally, in Section 4 we discuss the results obtained.

## 2 Methods

### 2.1 Models

Let $\mathbf{x}_j$ denote an $m \times 1$ vector containing $m$ measurements of gene $j$ ($j = 1, \ldots, p$), e.g. ratio intensities, and assume that there are $c$ copies of this gene in the sample strain. Further on, assume that the probability density of $\mathbf{x}_j$ conditioned on $c$ is $f_c(\mathbf{x}_j; \theta_c)$, where $\theta_c$ is a vector of unknown parameters. Then the unconditioned probability distribution of $\mathbf{x}_j$ follows a mixture model with $C + 1$ components

$$f(\mathbf{x}_j | \Psi) = \sum_{c=0}^{C} w_c f_c(\mathbf{x}_j; \theta_c),$$

where $\Psi = [\omega_0, \ldots, \omega_{C-1}, \theta'_0, \ldots, \theta'_C]^T$, and $\omega_c$ is the probability that a gene exists in $c$ copies ($\omega_c \geq 0$ and $\sum_{c=0}^{C} \omega_c = 1$). The parameter vector $\theta_c$ depends on the assumed probability distribution, e.g. if $\mathbf{x}_j$ is assumed normally distributed, then $\theta_c$ is a vector of the expectation and covariance param-
Suppose the value of $z_{j,c}$ is known, then the log likelihood function is given by:

$$l(\Psi) = \sum_{j=1}^{p} \left( \sum_{c=0}^{C} z_{j,c} (\log f_c(x_j; \hat{\theta}_c) + \log (\omega_c)) \right)$$

(McLachlan and Peel, 2000).

The iterative EM-algorithm (Expectation Maximization) is applied for estimation of the unknown parameters in the log likelihood function (Dempster, Laird and Rubin, 1977). In step 0 of the iteration some starting values are defined for the unknown parameters, $\Psi^{(0)}$. Here, we use the method of random starting values as described in McLachlan and Peel (2000).

In the E-step of the EM-algorithm $z_{j,c}$ is found as the estimated posterior probability for $c$ copies of gene $j$, given by:

$$z_{j,c} = \Pr(c \mid x_j) = \frac{\omega_c f_c(x_j; \hat{\theta}_c)}{\sum_{c=0}^{C} \omega_c f_c(x_j; \hat{\theta}_c)}, \quad (3)$$

where $\omega_c$ and $\hat{\theta}_c$ are the current parameter estimates from the previous step in the iteration.

In the M-step of the EM-algorithm the Maximum Likelihood Estimates (MLEs) based on the estimated values of the elements in $Z$ are computed. The estimators can be found in the Appendix. The parameter $\omega_c$ is estimated by the average over genes of $z_{j,c}$.

$$\hat{\omega}_c = \frac{1}{p} \sum_{j=1}^{p} z_{j,c}.$$ 

The E-step and the M-step are proceeded until convergence in the log likelihood function, measured by $|[l(\Psi^{(1)})] - l(\Psi^{(0)})|/|l(\Psi^{(0)})| < \text{tolerance}$, obtaining the maximum likelihood values of the mixture model. Local maxima are avoided by using various starting points. In this paper three different starting points are chosen, and methods described in McLachlan and Peel (2000) are used to detect presence of spurious local maximizers.

After convergence is obtained the estimated posterior probability of being present and divergent for gene $j$

$$\Pr(\text{present} \mid x_j) = \sum_{c=1}^{C} z_{j,c},$$
$$\Pr(\text{divergent} \mid x_j) = z_{j,0},$$

are computed by inserting the estimates from (3), and each gene is assigned to the group where $\Pr(\cdot \mid x_j)$ is maximized.

In this paper we study Mixture Model (1) with the assumption that $x_j$ follows the normal distribution, hence

$$f_c(x_j \mid \mu_c, \Sigma_c) = (2\pi)^{-\frac{d}{2}} |\Sigma_c|^{-\frac{1}{2}} e^{-\frac{1}{2} (x_j - \mu_c)^T \Sigma_c^{-1} (x_j - \mu_c)},$$

where $\mu_c$ and $\Sigma_c$ are the expectation vector and the covariance matrix for genes that exist in $c$ copies.
Let $T_{ij} (i = 1, \ldots, n$ and $j = 1, \ldots, p)$ be the $i$-th intensity of the $j$-th gene in the channel with the sample strain, and $R_{ij}$ be the $i$-th intensity of the $j$-th gene in the channel with the reference strain. Further on, let $M_j$ be the log$_2$-intensity ratio $M_j = \log_2(T_{ij}/R_{ij})$, and let $A_j = \log_2(\sqrt{T_{ij}} R_{ij})$ be the average log$_2$-intensity, in this paper called ratio intensity and total intensity, respectively. In the following, let $\mathbf{M} = \frac{1}{n} \sum M_j$, $\mathbf{A} = \frac{1}{n} \sum A_j$, $\mathbf{M}_j = [M_{1j}, \ldots, M_{nj}]^T$, and $\mathbf{A}_j = [A_{1j}, \ldots, A_{nj}]^T$.

Further on, let $\mathbf{M}_j$ be the log$_2$-intensity ratio $M_{ij} = \log_2(T_{ij}/R_{ij})$, and let $\mathbf{A}_j = \log_2(\sqrt{T_{ij}} R_{ij})$ be the average log$_2$-intensity, in this paper called ratio intensity and total intensity, respectively. In the following, let $\mathbf{M}_j = \frac{1}{n} \sum M_j$, $\mathbf{A}_j = \frac{1}{n} \sum A_j$, $\mathbf{M}_j = [M_{1j}, \ldots, M_{nj}]^T$, and $\mathbf{A}_j = [A_{1j}, \ldots, A_{nj}]^T$.

The simplest way to assign the genes into two groups by applying mixture models is to fit the model on the average ratio intensity for each gene. The model where $x_j = \mathbf{M}_j$ is denoted Model 1. Since the total intensities may have additional information about which group the genes belong to, Model 2 includes the total intensities, and $x_j = [\mathbf{M}_j, \mathbf{A}_j]$. By only considering the average of the replicates, the information about the variance between replicates is discarded, hence two genes with equal average ratio intensity will be predicted to exist in $c$ copies with equal probability, independent of whether the variance within the genes is equal or not. To take advantage of the variance within each gene, Model 1 is extended to Model 3 by applying the replicates of the ratio intensities for each gene, $x_j = \mathbf{M}_j$. Since we have technical replicates, the covariance matrix of $\mathbf{M}$ is supposed to have an equi-correlation structure (McLachlan, Do, and Ambroise, 2004). The same extension is done with Model 2 by applying the replicates of both the ratio intensities and the total intensities, achieving Model 4 where $x_j = [\mathbf{M}_j, \mathbf{A}_j]^T$. Due to the technical replicates the corresponding covariance matrix has two blocks with equi-correlation structure on the diagonal, and two off diagonal blocks with constrained structure. Irizarry et al. (2003) fits a mixture model to the averages of the replicates on each array; hence Model 4 is an extended version of their model, which again is an extended version of Model 2. However, Irizarry et al. do not assign the mutants into the group with maximized posterior probability, instead they test for differentially represented mutants by the use of $z_{ij}$ as weight for a weighted $z$-test based on the standardized average log ratio.

Table 1 gives an overview of the four models and the corresponding expectation vectors and covariance matrices.

For all the models the component corresponding to the group of divergent genes is the one containing the gene with the smallest observed average ratio.

<table>
<thead>
<tr>
<th>Model</th>
<th>$x_j$</th>
<th>$\mu_j$</th>
<th>$\Sigma_j$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$\mathbf{M}_j$</td>
<td>$\mu_{M,e}$</td>
<td>$\sigma^2_{M,e}$</td>
</tr>
<tr>
<td>2</td>
<td>$[\mathbf{M}_j, \mathbf{A}_j]^T$</td>
<td>$[\mu_{M,e}, \mu_{A,e}]^T$</td>
<td>$\begin{bmatrix} \sigma^2_{M,e} &amp; \sigma_{MA,e} \ \sigma_{MA,e} &amp; \sigma^2_{A,e} \end{bmatrix}$</td>
</tr>
<tr>
<td>3</td>
<td>$\mathbf{M}_j$</td>
<td>$\mu_{M,e} = \mu_{M,e}^1$</td>
<td>$\Sigma_{M,e} = \sigma^2_{M,e} \left((1 - \rho_{M,e}) I + \rho_{M,e} \mathbf{1} \mathbf{1}^T\right)$</td>
</tr>
<tr>
<td>4</td>
<td>$[\mathbf{M}_j, \mathbf{A}_j]^T$</td>
<td>$[\mu_{M,e}, \mu_{A,e}]^T$</td>
<td>$\begin{bmatrix} \Sigma_{M,e} &amp; \Sigma_{MA,e} \ \Sigma_{MA,e} &amp; \Sigma_{A,e} \end{bmatrix}$</td>
</tr>
</tbody>
</table>

where

$\mu_{M,e} = \mu_{M,e}^1$ and $\Sigma_{M,e} = \sigma^2_{M,e} \left((1 - \rho_{M,e}) I + \rho_{M,e} \mathbf{1} \mathbf{1}^T\right)$

$\mu_{A,e} = \mu_{A,e}^1$ and $\Sigma_{A,e} = \sigma^2_{A,e} \left((1 - \rho_{A,e}) I + \rho_{A,e} \mathbf{1} \mathbf{1}^T\right)$

$\Sigma_{MA,e} = \sigma_{MA,e} \left(\left(1 - \frac{\sigma_{MA,e}}{\sigma_{M,e}}\right) I + \frac{\sigma_{MA,e}}{\sigma_{M,e}} \mathbf{1} \mathbf{1}^T\right)$
2.2 Methods of validation

2.2.1 Criteria for validation
To be able to validate the models, i.e. both the different choices of $x$ and the number of components, some kind of validation criterion, dependent on whether the true group (divergent or present) of the genes are known or not, is needed.

If it is known whether the gene is divergent or present (e.g. simulation studies or reference-reference hybridizations), the estimated probability of correctly classifying a gene can be computed. However, this is a poor measurement because the number of divergent genes is much smaller than the number of present genes. A measurement that account for this is the geometric average (GA) of the estimated probability of correct classification of a divergent gene and the estimated probability of correct classification of a present gene

$$GA = \sqrt{Pr(\bar{D} \mid D) \cdot Pr(\bar{P} \mid P)}.$$  \hspace{1cm} (4)

where $\bar{D}$ is the event that a gene is classified as divergent, $D$ is the event that a gene is divergent, $\bar{P}$ is the event that a gene is classified as present, and $P$ is the event that a gene is present (Kubat et al., 1998).

It is natural to assign a gene to the divergent group if the corresponding posterior probability is maximized, i.e. apply the limit 0.5. Instead of assigning a gene to the group where the posterior probability is maximized, the limit can be adjusted. For a good model the number of divergent genes will then increase at the cost of a decreasing number of true present genes, or the number of true present genes will increase at the cost of a decreasing number of true divergent genes. To validate the models ability to divide the genes for different limits a receiver-operating characteristic (ROC) curve can be applied (Swets, 1988). The curve is a plot with the rate of genes correctly classified as divergent on the vertical axis and the rate of genes wrongly classified as divergent on the horizontal axis. Each limit provides a point in the plot, and many different choices of limit give a curve. A good model will separate divergent from present genes, and thus have a large area under the curve (AUC).

2.2.2 The geometric average studied by an analysis of variance model

It is of interest to test for significant differences in the results obtained with different choices of $x_j$, different number of components and different data sets. To be able to test for these differences a model has to be assumed. Based on ideas from CVANOVA (Cross Validation Analysis of Variance) introduced by Indahl and Næs (1998) we have used the following analysis of variance model

$$GA_{fCh} = \alpha + \tau_f + \beta_C + \gamma_h + (\tau \beta)_{fC} + (\tau \gamma)_{fh} + (\beta \gamma)_{fCh} + \epsilon_{fCh},$$  \hspace{1cm} (5)

which can be analyzed for all data sets simultaneously. Here $\alpha$ is the overall mean, $\tau_f$ is the effect of Model $f$, $\beta_C$ is the effect of $C$ number of components, and $\gamma_h$ is the effect of data set $h$. The errors are assumed to be independent and approximately normally distributed, however, the corresponding tests are robust against deviations from the normal distribution (Lindman, 1992). If there is a significant effect of any of the factors, multiple comparisons are used to identify the differences. For an introduction to analysis of variance and multiple comparisons see e.g. Montgomery (1997).

Model (5) can also be fitted with AUC as response. In addition, removing the effect of data set and the corresponding interaction terms can reduce the model; hence each data set can be analyzed separately.

3 Implementation

3.1 Data

The models were tested on four different data sets, where the true group of the genes were known. The bacterial strains used were the Staphylococcus aureus strains COL, Mu50 and N315, and the
Streptococcus pneumoniae strains TIGR4, R6 (ATCC-BAA-255) and G54. The two first data sets, denoted Mu50 and N315, were from experiments with COL as a reference strain and Mu50 and N315, respectively, as sample strains. The third and fourth data sets, denoted G54 and R6, were from experiments with TIGR4 as a reference strain and G54 and R6, respectively, as sample strains. A gene in the reference strain is treated as present in the sample strain if a pairwise local alignment to at least one of the sample strain genes shows more than 70% sequence identity. Several BLAST searches were made for every gene, varying the score and penalty parameters around their default values. The best hit for index gene $j$ was recorded for each search, and the fraction of exact matching residue-pairs was used as the identity score. The median identity score over all searches was used as the final identity score for gene $j$. Figure 1 shows a histogram of the sequence identity (Identity index) for N315.

The microarrays used were provided by the Pathogen Functional Genomics Resource Center (PFGRC), The Institute for Genomic Research (TIGR). The *Staph. aureus* arrays consisted of 2592 ORFs (PCR products) from *Staph. aureus* COL. Also, the array contained an additional 117 ORFs from strains Mu50 (60), MW2 (51), and N315 (6) which are not present in the COL strain’s genome complement, and where the MW2 genes were filtered out in the preprocessing. In addition, the microarrays contained control spots. All PCR products were printed in triplicate on the arrays. The genome sequences of all these strains are publicly available; see e.g. TIGR’s comprehensive microbial resource (CMR), www.tigr.org. The *Strep. pneumoniae* microarrays consisted of 2131 ORFs from *Strep. pneumoniae* TIGR4. The array also contained an additional 563 ORFs from strains R6 (164) and G54 (399), and control spots. PCR products were printed in quadruplicate. The genome sequences of strains R6 and TIGR4 are available; see e.g. TIGR’s comprehensive microbial resource (CMR), http://www.tigr.org/tigr-scripts/CMR2/CMRHomePage.spl, while the G54 genome sequence can be found at http://bioinfo.cnio.es/data/Spneumo/Stpn-annotation.html. All hybridizations experiments were performed twice, as dye-swap experiments. Some spots were declared missing. To obtain a balanced experiment these values were predicted using the average of the observed values for the gene. More advanced methods for predicting missing values exist, but in Feten, Almøy and Aastveit (2005) it is showed that the average is the best method for data with weak correlation between the genes, which is the case for the data in this paper. A further discussion about prediction of missing values is beyond the scope of this paper.

![Figure 1](image)

**Figure 1** A histogram of the sequence identity for each gene in the data set N315. The marked threshold at 0.7 is used to separate present genes from divergent genes.
Hybridized arrays were scanned at wavelengths 532 nm (cyanine-3) and 635 nm (cyanine-5) at 10 μm resolution to obtain two TIFF images, with a ScanArrayExpress Microarray Scanner (Packard Bioscience). Fluorescent intensities and spot morphologies were analyzed using the QuantArray program ver. 3.0 (Packard BioScience), and spots were excluded based on slide or morphology abnormalities.

The data were normalized with a procedure similar to the self-normalization described by Yang et al. (technical report available at http://www.stat.berkeley.edu/users/terry/zarray/Html/normspie.html), but extended with a smoothing of the gene-wise dye-effects over average log-intensity.

Let $M_{jai}$ be the observed ratio intensity for gene $j$ for spot $i$ on array $a$. A gene-wise effect of dye can be estimated from the linear model

$$M_{jai} = \mu_j + \delta_j I_a + e_{jai},$$

where $\mu_j$ is the expected ratio intensity for gene $j$, $\delta_j$ is the gene-wise effect of dye, $I_a$ is 1 or -1 depending on the dye used for the reference strain on array $a$, and $e_{jai}$ is noise. An effect of dye dependent on the total intensity, $d(A)$, can thus be found by smoothing the data $(A_j, \delta_j)$ e.g. by the lowess smoother (Cleveland, 1979), where $A_j$ is the total intensity for gene $j$ and $\delta_j$ is the estimated effect of dye for gene $j$. Finally, the normalized ratio intensities for each spot $M_{jai}$ are obtained as

$$M_{jai} = M_{jai}' - d(A_{jai})I_a.$$

Figure 2 shows the MA-plots with all replicates for each gene in the four data sets, while the corresponding plots with averaged values of each gene are shown in Figure 3. The variance unconditioned on genes appears to be larger in data from experiments with strains G54 and R6 compared to

![Figure 2](image-url)
data from experiments with the strains Mu50 and N315 for both the ratio intensities and the total intensities. Notice that the total intensities were smaller for Mu50 and N315 than for the other strains.

3.2 Numerical summaries of data

As an attempt to better understand the results, we studied the amount of total variability (measured by the sum of squares) explained by the groups, i.e. the variability between the groups divided by the total variability. This amount is an indication of the possibility to divide the genes into two groups based on the information in the ratio intensities and in the total intensities. The amount of unexplained variability, i.e. the variability within the genes divided by the total variability, gives us information about the variability within the genes, and hence whether the mixture models should be fitted on the averaged values of each gene, or on all the replicates. The amount of explained and unexplained variability is given in Table 2. For Mu50 the groups hardly explain anything of the variability neither in the ratio intensities nor in the total intensities, hence poor classification must be expected. In the other strains, with R6 as the superior, the groups explain more of the variability, and better classification can be expected. Especially for Mu50 and N315 improvement with Model 3 and Model 4 compared to Model 1 and Model 2 might be achieved due to large variability within the genes.

By fitting a two-stage nested design with genes nested under groups an $F$-value for the test of group effect is obtained (Montgomery, 1997). Let $\Lambda = |W|/|T|$, where $W$ is the sum of squares and products matrix within groups and $T$ is the total sum of squares and products matrix, then $F = d_f/\sqrt{1 - \Lambda} / 0.5 d_f \sqrt{\Lambda}$, where $d_f$ and $d_{fn}$ is the degrees of freedom for the denominator and the numerator respectively (Mardia, Kent, and Bibby, 1979). The $F$-value gives more accurate information about the effect of group than just the amount of variance explained. Larger $F$-values indicates
that the groups are well separated. Table 3 show the \( F \)-values from univariate (\( M_{ij} \)) and multivariate (\( \frac{1}{2} M_{ij}; A_{ij}/C_{138} \)) analysis of variance of the data. Due to different degrees of freedom all the \( F \)-values cannot be compared directly, but the results confirm that there is a smaller effect of group in Mu50 than in the other data sets.

The information above hopefully gives valuable contribution to a better understanding of when one model is in preference to another. However, in general the true group of the genes is unknown, hence additional information is needed as a guidance in the choice of model. Table 4 shows the average of the variance within genes and the variance between the genes for each strain respectively. We notice that Mu50 has large variances within the genes, especially for the ratio intensities, while the variance between the genes is small. This indicates that poor classification must be expected, but with possible improvement for the models that include all replicates. The opposite is the case for strain G54, which have small variances within the genes and large variance between the genes. Strain N315 has large variances of the intensities within the genes, while R6 has large variance of the intensities between the genes.

### 3.3 Results

To validate the models, we applied the geometric average (GA) given in (4). The analysis of variance model (5) is used to test for significant differences between the models, between the number of components and between the data sets. If nothing else is stated, the reported differences are assumed
The interaction between model and number of components and the interaction between number of components and data set were both found insignificant, hence these terms are removed from the model before the analysis. The results for the different models with different choices of number of components are shown in Figure 5. As expected, poor classification of the genes in Mu50 was obtained. The results indicate that for all strains except G54 the best choice of model is Model 4, namely the one that includes all replicates for both the ratio intensity and the total intensity. Note that the difference between the models is only significant for N315 and R6. In N315 including the replicates is an improvement independent of the number of components in the model, while for the other strains this is only the case for most of the components.

For Mu50 best classification is obtained with three and four components in the mixture model, while for the other strains the optimal classification is obtained with two components (see Figure 5). However, the difference between the results for various number of components is not significant.

Model 1 fitted on strain G54 maximizes GA with three components, which can be explained by the fact that the ratio intensities are distributed with two heavy tails. The other strains have a heavy tail to the left, and hence a model with two components should be fitted to the data. Since N315 and R6, in opposite to Mu50 and G54, have a distinct divergent tail in the MA-plot; great improvement in the classification is achieved by including the total intensities in the model.

The geometric average (GA) for the different models are shown in Figure 6. The results are averaged over number of components and indicate that the best choice of model is the most complex one, namely the one that includes all the replicates for both the ratio and the intensity. Even if there are significant interaction effects between models and data sets, Model 4 is preferred for all data sets.

### Table 4

Variance within and between the genes for each strain, where

\[
\hat{s}^2_j = \frac{1}{p(n-1)} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2
\]

and

\[
\hat{s}^2 = \frac{1}{p-1} \sum_{j=1}^{p} (\bar{x}_j - \bar{x})^2.
\]

<table>
<thead>
<tr>
<th>Strain</th>
<th>(\hat{s}^2_M)</th>
<th>(\hat{s}^2_A)</th>
<th>(\hat{s}^2_M)</th>
<th>(\hat{s}^2_A)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mu50</td>
<td>0.089</td>
<td>0.12</td>
<td>0.08</td>
<td>0.26</td>
</tr>
<tr>
<td>N315</td>
<td>0.061</td>
<td>0.24</td>
<td>0.15</td>
<td>0.42</td>
</tr>
<tr>
<td>G54</td>
<td>0.033</td>
<td>0.09</td>
<td>0.30</td>
<td>0.54</td>
</tr>
<tr>
<td>R6</td>
<td>0.064</td>
<td>0.13</td>
<td>0.25</td>
<td>0.62</td>
</tr>
</tbody>
</table>

![Figure 4](image-url) (a) and (b)

**Figure 4** Plot of residuals versus fitted values for Model (5) with GA (a) and AUC (b) as response.
can be shown that the models with two and three components achieved better assignments to the groups than the models with five components. The genes in Mu50 are as assumed hardest to assign to the correct group.

Figure 5  Plot of GA for different models and number of components for Mu50 (a), N315 (b), G54 (c), and R6 (d). Model 1 (---), Model 2 (--•--), Model 3 (--■--), and Model 4 (--▲--).

Figure 6  Plot of GA for different models for Mu50 (---), N315 (--•--), R6 (---▲--), and G54 (--■--).
The optimal model for each strain was found with the classification limit 0.5. Validation of the models with other choices of classification limits is also interesting; in this study this is done by ROC-curves. Model (5) with AUC as response is used to test for significant differences. The reported differences are, if nothing else is stated, assumed significant at 0.05 level, where the \( p \)-values are adjusted by the Tukey method. The corresponding residual plot is shown in Figure 4(b). The interaction between number of components and data set was found insignificant, hence this term is removed from the model before the analysis. The results from every combination of model and number of components will not be presented, only the ones that are best based on GA and AUC respectively. Figure 7 shows the ROC-curves for the best choice of both model and components for the different strains. For strain Mu50, Model 4 with four components achieved the largest, but insignificantly, improvement compared to the model with highest GA, where AUC increases from 0.58 to 0.76, in addition GA is almost unchanged (decreases from 0.50 to 0.48). This is the strain with largest overlapping region between the two groups, and hence poorest possibility to assign the genes into correct group. For the other strains the improvement is not that large, and in addition GA becomes a lot smaller, especially for N315 and R6. For strain G54 AUC increases from 0.76 to 0.82 by Model 2 with four components, while GA decreases from 0.69 to 0.63, but none of the changes are significant. For strain N315 and R6 an improvement in AUC from 0.86 to 0.91 and from 0.75 to 0.83 is obtained, respectively, both by Model 2 with five components, while GA decreases from 0.78 to 0.66 and from 0.71 to 0.57. However, there is no significant difference between Model 2 and Model 4.

![Figure 7](https://www.biometrical-journal.com)

**Figure 7** ROC curves for the optimal model based on GA (lower curve, black) and for the optimal model based on AUC (upper curve, grey) for Mu50 (a), N315 (b), G54 (c), and R6 (d). The fraction of truly divergent genes is on the vertical axis, and the fraction of falsely divergent genes is on the horizontal axis.
The area under curve seems to be almost unaffected by the number of components for the models based on the ratio intensities (results not shown). For the model based on the average of the ratio intensities and the average of the total intensities (Model 2) AUC is increasing with increasing number of components, except for strain Mu50. Most of the genes in the overlapping region in Figure 3 are present. When a mixture model with three components is fitted, most of these genes will be assigned to the present group with larger probability than when a mixture model with two components is fitted. Since most of these genes are truly present, AUC will increase. Notice that the difference between AUC for different number of components is not significant.

In Table 5 the results obtained with mixture models, maximizing GA and AUC respectively, are compared with the corresponding results from the GACK approach by Kim et al. (2002). For two out of four classification with mixture models based on GA gives better results than GACK, while for three out of four classification with mixture models based on AUC gives better results than GACK. One explanation might be that GACK has its benefits when the distribution of the group of present genes is symmetric, while mixture models solve deviation from a symmetrical distribution by adding more components to the model. Notice that the difference between the method is small when GACK achieves better results, while it is larger when mixture models achieve better results.

4 Discussion

The aim of this paper was to study four different mixture models’, each with four different number of components, capability to assign genes into groups of divergent and present. Different numerical summaries of the data were given as a guide in the choice of model and number of components, while two criteria together with analysis of variance were used to validate the models.

In this paper we have used two criteria in the validation of the methods. The first criterion, denoted GA, measures the geometric average of the estimated probability of correct classification of a divergent gene and the estimated probability of correct classification of a present gene given a classification limit. The second criterion, denoted AUC, measures the models’ ability to divide the genes for different limits. Since the interest is in the classification error received for a given limit, the models were mainly validated by GA. A more thorough comparison of the validation criteria are important, but beyond the scope of this paper.

The mixture models in this paper assume a normal distribution of the log₂-intensity ratios (ratio intensities) and the average log₂-intensities (total intensities), and hence of the log₂ intensities themselves. Different tests for normality rejected the hypothesis that the log₂-intensities of the data in our study follow a normal distribution. However, the method based on mixture models seems to be robust against small deviations from this assumption.
It is common to assume that the genome of the reference strain consists of regions each of which is dominantly present or dominantly divergent in the sample strain (Wang and Guo, 2004). This spatial information was used to smooth the estimated posterior probabilities by moving average. However, the genes’ position on the genome did not improve the classification, hence the results were not included in the paper.

The structure of the covariance matrices does not consider that the experiment is a dye swap experiment, i.e. the replicates are from two arrays. Since the complexity of the models increase when including array effect and the effect of array is supposed to be removed in the normalization, we found this refinement not necessary.

If the interest is to find the minimal genome common to both the reference and the sample strain, it may be an advantage to increase the classification limit above 0.5. Then the fraction of truly present genes is increasing, and the predicted minimal genome is correct with larger probability.

In the data sets analyzed in our study mixture models improved the classification of present and divergent genes compared to the previously published method GACK in some cases, but not in all. Mixture models seems to be in preference for data sets where the main peak is non-symmetric around its maximum and for data sets with two apparent regions in the MA-plot.

The results indicated that large variance between the genes give good classification. This is not necessarily universal, since the groups still might be overlapping. However, data sets with two relative clear regions in the MA-plot seemed to give good classification independent of model and component choice. Each of the two regions consists of divergent and present genes, while the overlapping region consists of both, but most present genes since they are in majority on the genome.

If the mixture models are fitted with the replicates of the genes and not the average within each gene, the increase in number of parameters is small compared to the total number of observations, and the classification is in general improved, even for data with small variances within the genes. When the amount of variance explained by the group is extremely small, the improvement in classification is negligible.

This study showed that best classification is obtained by fitting mixture models with both ratio intensities and total intensities for all replicates of the genes. Based on the geometric average of the estimated probability of correct assignment of divergent genes and present genes (GA) the models should be fitted with two components, but by increasing the number of components the area under the ROC-curve is increased without too much loss in GA.

Acknowledgements We thank the Pathogen Functional Genomics Resource Center (PFGRC), The Institute for Genomic Research (TIGR), Rockville, MD for providing microarrays. The *Staph. aureus* Mu50 were kindly provided by Lisa Herron, University of Minnesota, St. Paul, MN, while prof. Hans-Georg Sahl, University of Bonn, Germany provided us with the *Staph. aureus* COL and N315 strains. The *Strep. pneumoniae* G54 strain was provided to us by Dr. Francesco Iannelli, University of Siena, Italy, while Dr. Ingeborg S. Aaberge, The Norwegian Institute of Public Health, offered the *Strep. pneumoniae* TIGR4 strain to us.

We would like to thank two anonymous referees for constructive criticism. The Norwegian University of Life Sciences has financed this work.

**Appendix-Maximum likelihood estimators**

For cases were there is no constraints neither on the expectation vectors nor on the covariance matrices, the maximum likelihood estimators (MLEs) of the unknown parameters in the \((q+1)\)-th iteration are given by

\[
\hat{\mu}_c^{(q+1)} = \frac{\sum_{j=1}^{p} \hat{z}_{c,j}^{(q)} \mathbf{x}_j}{\sum_{j=1}^{p} \hat{z}_{c,j}^{(q)}},
\]

\[\text{(6)}\]
and

\[
\Sigma_{(q+1)} = \frac{\sum_{j=1}^{n} \xi^{(q)}_{(j)}(x_j - \hat{\mu}^{(q+1)}_{(j)})(x_j - \hat{\mu}^{(q+1)}_{(j)})^t}{\sum_{j=1}^{n} \xi^{(q)}_{(j)}}, \tag{7}
\]

respectively (McLachlan and Peel, 2000). For simplicity the iteration number is omitted in the following.

For the univariate model (Model 1) and the bivariate model (Model 2) the MLEs in the log likelihood function in (2) are given in Equation (6) and Equation (7).

For the models with constraints either in the expectation vectors or in the covariance matrices, or in both, the MLEs can be found in most books in advanced multivariate analysis, e.g. Mardia et al. (1979). In Model 3 the MLE of \(\mu_{M,e}\) is given by

\[
\hat{\mu}_{M,e} = \frac{1}{n} \sum_{i=1}^{n} \hat{\mu}_{(i),e},
\]

where \(\hat{\mu}_{(i),e}\) is element \(i\) in \(\hat{\mu}_e\) from (6), while the MLE of the variance of \(M_{ij}\) and the correlation between \(M_{ij}\) and \(M_{i'j'}\) is given by

\[
\hat{\sigma}^2_{M,e} = \frac{1}{n} \sum_{i=1}^{n} \Sigma_{(i),e} / n,
\]

and

\[
\hat{\rho}_{M,e} = 2 \sum_{i<j} \Sigma_{(i',j'),e} / [n(n-1) \hat{\sigma}^2_{M,e}],
\]

respectively, where \(\Sigma_{(i',j'),e}\) is element \((i',j')\) in \(\Sigma_e\) as given in (7).

With the constraints in Model 4 the MLE of the expectations are given by

\[
\hat{\mu}_{M,e} = \frac{1}{n} \sum_{i=1}^{n} \hat{\mu}_{(i),e},
\]

and

\[
\hat{\mu}_{A,e} = \frac{1}{n} \sum_{i=1}^{n} \hat{\mu}_{(i),e}.
\]

The MLE of the variances are given by

\[
\hat{\sigma}^2_{M,e} = \frac{1}{n} \sum_{i=1}^{n} \Sigma_{(i),e} / n,
\]

and

\[
\hat{\sigma}^2_{A,e} = \frac{1}{n} \sum_{i=1}^{n} \Sigma_{(i),e} / n,
\]

while the corresponding correlations are estimated by

\[
\hat{\rho}_{M,e} = 2 \sum_{i<j} \Sigma_{(i',j'),e} / [n(n-1) \hat{\sigma}^2_{M,e}],
\]

and

\[
\hat{\rho}_{A,e} = 2 \sum_{n+1<j<i} \Sigma_{(i',j'),e} / [n(n-1) \hat{\sigma}^2_{A,e}],
\]
The MLE of the covariance between $M_i$ and $A_i$ for $i = i'$ is given by

$$\hat{\Sigma}_{(MA)_{i,i'}} = \frac{1}{n} \sum_{i=1}^{n} \hat{\Sigma}_{(n+i),i,i'} / n,$$

while for $i \neq i'$

$$\hat{\Sigma}_{(MA)_{i,i'}} = 2 \sum_{i < j \leq n} \hat{\Sigma}_{(n+i'),i,j}/[n(n-1)].$$

References


