Regression as a Method to Predict Copy Numbers in Comparative Genomic Hybridization Studies on Bacteria

Guri Feten*, Trygve Almøy, Lars Snipen, Ågot Aakra, 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 6 April 2005, revised 15 October 2005, accepted 28 October 2005

Summary

Comparative genomic hybridizations (CGH) using microarrays are performed with bacteria in order to determine the level of genomic similarity between various strains. The microarrays applied in CGH experiments are constructed on the basis of the genome sequence of one strain, which is used as a control, or reference, in each experiment. A strain being compared with the known strain is called the unknown strain. The ratios of fluorescent intensities obtained from the spots on the microarrays can be used to determine which genes are divergent in the unknown strain, as well as to predict the copy number of actual genes in the unknown strain. In this paper, we focus on the prediction of gene copy number based on data from CGH experiments.

We assumed a linear connection between the $\log_2$ of the copy number and the observed $\log_2$ -ratios, then predictors based on the factor analysis model and the linear random model were proposed in an attempt to identify the copy numbers. These predictors were compared to using the ratio of the intensities directly.

Simulations indicated that the proposed predictors improved the prediction of the copy number in most situations. The predictors were applied on CGH data obtained from experiments with *Enterococcus faecalis* strains in order to determine copy number of relevant genes in five different strains.

Key words: Microarray experiment; Comparative genomic hybridization; Factor analysis model; Linear random model.

1 Introduction

Comparative genomic hybridization (CGH) studies are performed to examine genomic similarity between organisms (Behr et al., 1999; Björkholm et al., 2001). An introduction to CGH-technology can be found in e.g. Sassetti and Rubin (2002), and Mantripragada et al. (2004). In such comparative studies, the genome sequence of one of the strains is known, and a microarray has been constructed on this basis, where every gene has been (repeatedly) spotted. In the CGH experiment the gene content of an unknown strain is compared to the known strain by hybridization of fluorescently labelled genomic DNA from the known and the unknown strains to the microarray. Prior to labelling and hybridization, the genomic DNA from the two strains is usually fragmented, either physically or enzymatically. The fragmented DNA is labelled with dyes (e.g. Cy3 and Cy5), one colour for the known strain and one colour for the unknown strain. Upon scanning the dyes fluoresce, and the fluorescent intensities from each spot (representing a gene) are estimated. Scanning and image analysis are described by Nguyen et al. (2002).

From the obtained intensities we try to draw conclusions about which genes are divergent in the unknown strain. For genes that are present in the unknown strain, we try to draw conclusions about the copy number of the genes.

* Corresponding author: e-mail: guri.feten@umb.no, Phone: +47 64 96 58 28, Fax: +47 64 96 59 01

Kim et al. (2002) introduced a method for dividing the genes into two groups, divergent and present. The method selects a $\log_2$ -ratio value, based on the variation in strain composition and the quality of hybridization, used as a cutoff between the two groups. In addition the method provides an estimate of the probability that a given gene is present.

Autio et al. (2003) used k -means clustering and dynamic programming to divide the genes into three groups, divergent, equally present, and amplified (genes with higher copy number in the unknown strain compared to the known strain).

A refinement of dividing the genes into three groups is to divide the genes into sets with equal copy number. Instead of finding the copy numbers, methods for identifying copy number changes have been worked out by aim of an unsupervised hidden markov models approach (Fridlyand et al., 2004), and standard t -statistic and a modification with variances smoothed along the genome (Wang and Guo, 2004).

The observed ratios for each gene are unknown functions of the copy number. There are, of course, many disturbances before the ratios are observed: In CGH experiments, fragments of labelled genomic DNA are hybridized to the microarray. In general, the strategy chosen for DNA fragmentation will not give fragments with exact match to the DNA fragments (here: PCR products) on the arrays. We may therefore be faced with an unspecific hybridization, which could be viewed as noise in the end.

Traditionally, genes in bacterial genomes have been regarded as non-redundant, i.e. the majority of bacterial genes were believed to be present in one copy per genome. Natural selection has been assumed to favour the elimination of gene redundancy in bacteria. Today, this view is being somewhat revised. The complete genome sequences of bacteria that are available today show that nearly all bacteria contain multiple copies of certain genes, so called “paralog” gene families (Pushker, Mira, and Rodríguez-Valera, 2004).

In this paper we will describe various predictors to predict the gene copy number, along with two different criteria to validate these predictors (Section 2). The predictors are well known in statistics, but, as far as we know, previously not used in prediction of copy numbers. In Section 3 the predictors are illustrated by an example. Section 4 contains the design and the results of a simulation study in order to compare the different predictors. Finally, in Section 5 we discuss the results obtained from both the study of real data and the simulation study, and raise some questions for further studies.

2 Methods

2.1 General model

Let y be the $\log_2$ -ratio of the intensity of a gene with copy number c , with $c = [0, 1, \dots, K]$. Assume that $\mathbf{w} = [y, x]^T$ follows the Probability Density Function (PDF) $f(\mathbf{w}, \boldsymbol{\theta})$, where $x = \log_2(c + \varepsilon)$, and ε is a small positive number introduced since the logarithm of zero does not exist. The Best Predictor (BP) for x (under expected quadratic loss) is

$$\hat{x}_{BP} = E(x|y),$$

where $E(x - \hat{x})^2 = \text{Var}(x|y)$ (Bickel and Doksum, 1977). Since the PDF of $\mathbf{w}$ is unknown, $\hat{x}_{BP}$ cannot be found, and we need to take some further assumptions into account. Let index j always refers to genes. We assume that for each of p genes there is a linear relationship between x and the $\log_2$ -ratio y . Hence for gene j we have

$$y_j = \beta_0 + \beta_1(x_j - \bar{x}) + e_j, \quad (1)$$

where $\bar{x}$ is the average of x_j over all genes, and where common assumptions about e_j are made.

If calibration data with known copy numbers were available, and one assumes x in (1) to be random, the natural predictor of x is (Brown, 1993)

$$\hat{x} = \bar{x} + \frac{s_{xy}}{s_{yy}}(y - \bar{y}),$$

where $\bar{x}$ and $\bar{y}$ are the sample means from the calibration data and $s_{xy} = \sum_{j=1}^p (x_j - \bar{x})(y_j - \bar{y})$, and $s_{yy} = \sum_{j=1}^p (y_j - \bar{y})^2$.

Let index i always refers to spots. An extension of Model (1) arises if we have n_j measures on gene j ,

$$\mathbf{y}_j = \beta_0 \mathbf{1} + \beta_1 (x_j - \bar{x}) \mathbf{1} + \mathbf{e}_j, \quad (2)$$

where $\mathbf{y}_j$ is a vector of the n_j observations for gene j , and where common assumptions about the errors $\mathbf{e}_j$ are made. If nothing else is mentioned, a balanced experiment is assumed ($n_j = n$ for all genes).

If we assume that x_j is a random variable with variance ϕ^2 , the unconditional variance of $\mathbf{y}_j$ and the covariance between $\mathbf{y}_j$ and x_j are

$$\text{Var}(\mathbf{y}_j) = \phi^2 \beta_1^2 \mathbf{1}\mathbf{1}' + \sigma^2 \mathbf{I}, \quad \text{and} \quad \text{Cov}(\mathbf{y}_j, x_j) = \phi^2 \beta_1 \mathbf{1}. \quad (3)$$

Since a calibration set is not available, other techniques have to be taken into account in order to estimate the unknown parameters in (3). Three predictors, all based on a linear model, are proposed here. In addition, a predictor without any model assumptions is proposed. The predictors are summarized in Table 1.

2.2 The predictors

2.2.1 Ratio

The simplest predictor, in this paper called Ratio (R), predicts the copy number by considering the average of the observed ratios for a gene, i.e. the predictor does not consider the $\log_2$ -ratio (Björkholm et al., 2001). If we let $z_{ij} = 2^{y_{ij}}$ be the observed ratio for gene j on spot i , then the predicted copy number for gene j is based on

$$\bar{z}_j = \frac{1}{n} \sum_{i=1}^n z_{ij}. \quad (4)$$

2.2.2 The factor analysis model

Model (2) with x regarded as random can be viewed as a factor analysis model with one common factor given by

$$\mathbf{y}_j = \boldsymbol{\mu} + \boldsymbol{\gamma} f_j + \mathbf{e}_j, \quad (5)$$

where $\boldsymbol{\mu} = \mu \mathbf{1}$ ($n \times 1$) is a vector of constants, $\boldsymbol{\gamma} = \gamma \mathbf{1}$ ($n \times 1$) is a vector of factor loadings, f_j is the factor score, and $\mathbf{e}_j$ ($n \times 1$) is a vector of specific factors (Lawley and Maxwell, 1973). Assume f_j and $\mathbf{e}_j$ to be independent, f_j distributed with expectation 0 and variance 1, and $\mathbf{e}_j$ independently distributed $\mathbf{e}_j \sim N(\mathbf{0}, \psi \mathbf{I})$. (In factor analysis it is common to assume that f_j is normally distributed.) The unconditional variance of $\mathbf{y}_j$ and the covariance between $\mathbf{y}_j$ and f_j are given by (3) with $x_j = f_j$, $\phi^2 = 1$, $\beta_1 = \gamma$, and $\sigma^2 = \psi$. The factor analysis model in (5) corresponds to the model in (2) if $\boldsymbol{\mu} = \mu \mathbf{1}$ and $\boldsymbol{\gamma} = \gamma \mathbf{1}$. Known methods to predict f_j can now be applied to predict $(x_j - \bar{x})$.

The Best Linear Predictor (BLP) for f_j based on Model (5) becomes

$$\hat{f}_{T1,j} = \frac{1}{\frac{\psi}{n\gamma} + \gamma} (\bar{y}_j - \mu), \quad (6)$$

where $\bar{y}_j$ is the average of the $\log_2$ -ratios for gene j . (If f_j is normally distributed this is an unbiased predictor). This predictor is also known as Thompson's factor score (Mardia, Kent, and Bibby, 1979).

The unknown parameters (γ , ψ , and μ) are estimated by

$$\begin{aligned}\hat{\gamma}^2 &= \frac{2}{n(n-1)} \sum_{l < l'} s_{(y)ll'} = \bar{s}_{(y)ll'}, \\ \hat{\psi} &= \frac{1}{n} \sum_{l=1}^n s_{(y)ll} - \frac{2}{n(n-1)} \sum_{l < l'} s_{(y)ll'} = \bar{s}_{(y)ll} - \bar{s}_{(y)ll'}, \\ \hat{\mu} &= \frac{1}{np} \sum_{i=1}^n \sum_{j=1}^p y_{ij} = \bar{y},\end{aligned}\quad (7)$$

where $s_{(y)ll'}$ is the element l, l' in the covariance matrix $S_{yy} = \frac{1}{p} \sum_{j=1}^p (y_j - \bar{y})(y_j - \bar{y})^t$, $\bar{y} = \frac{1}{p} \sum_{j=1}^p y_j$, and y_j is the vector containing the n observations for gene j . (If f_j is normally distributed, these estimators are maximum likelihood estimators, Mardia et al., 1979.) The estimators in (7) inserted in (6) lead to the predictor we call T1 in this paper.

The previous models do not take into account any effect of array, neither in the expectation, nor in the variance. Let index k refers to arrays. Each of the q arrays consists of m_{kj} replicates of each gene $\left(\sum_{k=1}^q m_{kj} = n_j \right)$. A balanced experiment is assumed ($m_{kj} = m$ for all arrays and genes). An extension of the model given by Eq. (5) can be given as

$$y_{kj} = \mu_k \mathbf{1} + \gamma f_j + e_{kj},$$

where μ_k is the array effect and e_{kj} is assumed independently distributed, $e_{kj} \sim N(0, \psi_k \mathbf{I})$. Since $\bar{y}_{kj} = \frac{1}{m} \mathbf{1}' y_{kj}$, a factor model which simultaneously models all arrays is

$$\bar{y}_j = \mu + \gamma f_j + \bar{e}_j, \quad (8)$$

with $\gamma = \gamma \mathbf{1}$, and the $\bar{e}_j$'s independently distributed $\bar{e}_j \sim N(0, \frac{1}{m} \Psi)$, where Ψ is a diagonal matrix with ψ_k as element k, k .

Thompson's factor score based on the model given by Equation (8) becomes

$$\hat{f}_{T2,j} = m\gamma(1 + m\gamma^2 \mathbf{1}' \Psi^{-1} \mathbf{1})^{-1} \mathbf{1}' \Psi^{-1} (\bar{y}_j - \mu). \quad (9)$$

The estimators for γ , Ψ , and μ inserted in (9) give a predictor we call T2.

The parameters are estimated by

$$\begin{aligned}\hat{\gamma}^2 &= \frac{2}{q(q-1)} \sum_{l < l'} s_{(\bar{y})ll'} = \bar{s}_{(\bar{y})ll'}, \\ \hat{\Psi} &= \text{diag} \left(m \left(s_{(\bar{y})kk} - \frac{2}{q(q-1)} \sum_{l < l'} s_{(\bar{y})ll'} \right) \right) = \text{diag} (m(s_{(\bar{y})kk} - \bar{s}_{(\bar{y})ll'})), \\ \hat{\mu} &= \frac{1}{p} \sum_{j=1}^p \bar{y}_j,\end{aligned}$$

where $s_{(\bar{y})ll'}$ is the off diagonal element l, l' in the covariance matrix $S_{\bar{y}\bar{y}} = \frac{1}{p} \sum_{j=1}^p (\bar{y}_j - \tilde{\bar{y}})(\bar{y}_j - \tilde{\bar{y}})^t$, $\tilde{\bar{y}} = \frac{1}{p} \sum_{j=1}^p \bar{y}_j$. (If f_j is normally distributed, these estimators are also maximum likelihood estimators, Mardia et al., 1979.)

2.2.3 The linear random model

For unbalanced experiments iterative procedures have to be used to achieve estimates of the parameters in Eq. (6). If Model (2) is treated as a random model, the unknown parameters can be directly estimated from the sum of squares within genes, the sum of squares between genes, and the total sum of squares.

If the genes are assumed to be randomly chosen from the population of genes, the effect of gene j , here denoted a_j , is assumed to be random. The effect of gene j is a function of the corresponding copy number. The linear random model is given by

$$y_{ij} = \mu + a_j + e_{ij}, \quad (10)$$

where a_j and e_{ij} are independent random variables. Assume a_j is distributed with expectation 0 and variance σ_a^2 , hence $\frac{a_j}{\sigma_a}$ is distributed with variance 1, and $e_{ij} \sim N(0, \sigma^2)$. (It is common in analysis of variance to assume that a_j is normally distributed.)

A closer inspection of the models in (5) and (10) shows they are similar if $\psi = \sigma^2$, $\gamma^2 = \sigma_a^2$, and

$$f_j = \frac{a_j}{\sigma_a}. \quad (11)$$

The estimators of σ_a^2 and σ^2 can be found in e.g. Montgomery (1997).

The Best Linear Predictor (BLP) of a_j is

$$\hat{a}_j = \frac{\sigma_a^2}{\frac{\sigma^2}{n_j} + \sigma_a^2} \left(\bar{y}_j - \frac{\sum_{j=1}^p \frac{n_j}{n_j \sigma_a^2 + \sigma^2} \bar{y}_j}{\sum_{j=1}^p \frac{n_j}{n_j \sigma_a^2 + \sigma^2}} \right) \quad (12)$$

(Searle, 1987). By inserting the estimated values of σ_a^2 and σ^2 into (12) a prediction of a_j is achieved. If these predicted values, together with the square root of the estimate of σ_a^2 are substituted into (11), we receive $\hat{f}_{L,j}$, a predictor we call L.

For equal sample sizes we get a predictor equal to Thompson's factor score in (6), but due to different methods of estimation, the maximum likelihood procedure does not take into account the degrees of freedom lost for estimation of the fixed parameters, their predictions are only approximately equal.

Table 1 Overview of the predictors with the acronyms used in the text. The numbers in the brackets refer to the number of the corresponding equation in the text. The unknown parameters in the equations are estimated (see the text). The predicted copy number for the regression predictors is given by $\hat{c}_j = \text{round} \left(\frac{1}{2t+1} \sum_{r=j-t}^{j+t} 2\hat{f}_{r+\bar{y}} \right)$, while the predicted copy number for Ratio is given by $\hat{c}_j = \text{round} \left(\frac{1}{2t+1} \sum_{r=j-t}^{j+t} \bar{z}_j \right)$.

Name of predictor		Equation		Model
Ratio	R	$\bar{z}_j = \frac{1}{n} \sum_{i=1}^n z_{ij}$	(4)	No model
Thompson's factor score	T1	$\hat{f}_{T1,j} = \frac{1}{\frac{\gamma}{n_j} + \gamma} (\bar{y}_j - \mu)$	(6)	$y_j = \mu + \gamma f_j + e_j$ (5)
Thompson's factor score	T2	$\hat{f}_{T2,j} = \frac{m\gamma \mathbf{1}' \Psi^{-1}}{1+m\gamma^2 \mathbf{1}' \Psi^{-1} \mathbf{1}} (\bar{y}_j - \mu)$	(9)	$\bar{y}_j = \mu + \gamma f_j + \bar{e}_j$ (8)
Linear random model	L	$\hat{f}_{L,j} = \frac{a_j}{\sigma_a}$	(11)	$y_{ij} = \mu + a_j + e_{ij}$ (10)

2.3 Prediction of copy numbers

If we assume that $\hat{f}_j = (\hat{x}_j - \bar{x}) = \hat{x}_j - \hat{\bar{x}}$, the predicted values of f_j could have been used to predict x_j , i.e. $\hat{x}_j = \hat{f}_j + \hat{\bar{x}}$, but $\hat{\bar{x}}$ is unknown, hence $\hat{x}_j$ cannot be found. To avoid this problem we assume that the average of the x 's and the average of the y 's ($\log_2$ -ratios) are approximately equal, then $\hat{x}_j \approx \hat{f}_j + \bar{y}$. Our final interest is the predicted value of c_j , where $c_j = 2^{x_j} - \varepsilon$ with ε ignorable.

It is common to assume that a genome consists of regions of genes with equal copy number (Wang and Guo, 2004). To take advantage of this spatial information, the predicted copy numbers could be smoothed by moving average with $2t$ neighbours. The predicted copy number of gene j is then

$$\tilde{c}_j = \frac{1}{2t+1} \sum_{r=j-t}^{j+t} 2^{\hat{f}_r + \bar{y}}$$

for the regression methods, and

$$\tilde{c}_j = \frac{1}{2t+1} \sum_{r=j-t}^{j+t} \bar{z}_r$$

for Ratio. In this paper we have used two neighbours ($t = 1$). Since $\tilde{c}_j$ is continuous, and the actual copy number is discrete, we convert the continuous numbers into discrete by rounding the predicted values to the nearest integer, or

$$\hat{c}_j = \text{round}(\tilde{c}_j).$$

2.4 Validation of predictors

If the true copy numbers were known, there are different criteria, depending on the aspect of interest, to use in the validation of the predictors. If the focus is on predicting the copy number as well as possible, the common criterion is the Root Mean Square Error of Prediction (RMSEP),

$$\text{RMSEP} = \sqrt{\frac{1}{p} \sum_{j=1}^p (c_j - \hat{c}_j)^2},$$

where p is the number of genes, c_j is the copy number of gene j , and $\hat{c}_j$ is the predicted copy number of the corresponding gene. The RMSEP is an estimator of the expected square error of prediction, $\sqrt{E(c - \hat{c})^2}$, where large errors are heavier weighted than small errors.

When the main focus is on grouping the genes into three classes, absent ($c = 0$), equally present ($c = 1$), and present in multiple copies ($c \geq 2$), a natural choice for validation is the Error of Classification (EC), where

$$\text{EC} = \frac{1}{p} \left(\sum_{j=1}^p I(c_j = 0 \text{ and } \hat{c}_j \neq 0) + \sum_{j=1}^p I(c_j = 1 \text{ and } \hat{c}_j \neq 1) + \sum_{j=1}^p I(c_j \geq 2 \text{ and } \hat{c}_j < 2) \right),$$

where $I(\cdot)$ equals 1 if the expression is true and 0 otherwise. The EC measure is an estimator of the probability of predicting wrong group.

The practical difficulty is that the true copy number is seldom known, the exceptions are experiments where the genome sequences are available (known-known hybridization). Since we have no such data available, we simulate data for validation of the predictors. The predictors are also compared, but not validated, based on results from real data where the genome sequences are unknown.

3 Example

3.1 Data

As an example we applied the predictors on data where the known strain is the *Enterococcus faecalis* V583 bacterium (Paulsen et al., 2003). The genomic DNA of V583 was compared to five other *E. faecalis* strains, denoted NCDO 581, NCDO 642, EF BRIDGE, JH2SS and V24.

There were 3245 potential open reading frames (ORFs, genes) from V583 represented on the microarrays, as PCR products of approximately 500 bp. Each PCR product was spotted, independently, five times each. The experiments were performed in duplicate, as dye-swap experiments; hence we obtained up to ten observations per gene. In addition to the PCR products representing V583 ORFs, the arrays contained negative controls: three genes from *Arabidopsis thaliana*, empty spots (i.e. no liquid spotted) and spots containing buffer ($3\times$ SSC, 0.1% SDS, only; no DNA). As a positive control, restriction digested genomic DNA from V583 was used. Negative and positive controls were also spotted five times each.

Hybridized arrays were scanned at wavelengths 532 nm (cyanine-3) and 635 nm (cyanine-5) at 10 μ m 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.

Due to the asymmetric distribution of $\log_2$ -ratios, a standard intensity-dependent normalization (e.g. lowess-normalization) could not be used. All our data are from dye-swap experiments, and we have used a normalization 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 genewise dye-effects over average log-intensity.

Some spots were declared missing, hence the number of observations was lower than ten for some genes. For Ratio and the predictor based on the factor analysis model (T1), these values were estimated using the average of the observed values for the gene. For the predictor based on the extended factor analysis model (T2), the values were estimated using the average of the observed values for the gene within each array. For bacterial data with replicates these estimators of missing values are to prefer compared to more advanced methods (Feten et al., 2005).

Figure 1 Histogram of the observed ratios.

Table 2 The average for each array ($\bar{y}_k$) and the variances between the genes for each array ($\text{Var}(\bar{y}_{kj}) = s_k^2$).

Strain	$\bar{y}_1$	$\bar{y}_2$	s_1^2	s_2^2	$\max(s_k^2)/\min(s_k^2)$
NCDO 581	-0.62	-0.62	1.06	0.52	2.04
NCDO 642	-0.70	-0.71	1.60	1.14	1.40
EF BRIDGE	0.55	0.55	1.79	1.93	1.08
JH2SS	0.38	0.37	0.75	0.93	1.24
V24	0.51	0.50	1.47	1.38	1.07

Restriction enzyme digestion of genomic DNA was performed using the restriction endonuclease *RsaI*. The recognition sequence of *RsaI* is $\text{GT}^{\downarrow}\text{AC/CA}^{\downarrow}\text{TG}$. Digestion of a completely random DNA sequence with *RsaI* would generate fragments of $4^4 \text{ bp} = 256 \text{ bp}$. The actual average fragment length obtained by *RsaI* digestion of V583 genomic DNA was 402 bp (http://www.tigr.org/tigr-scripts/CMR2/restrict_display.pl). It was assumed that a similar average fragment size would be obtained by digestion of genomic DNA from other strains within the *E. faecalis* species.

Figure 1 shows histograms of the observed ratios. Notice that strain EF BRIDGE and strain V24 have some relatively large observed ratios.

In Table 2 the total averages for each array, the variances between the genes for each array, and the ratios between the corresponding largest and smallest variance are given. The strains with some large observed ratios (EF BRIDGE and V24) have large average for each array. The variances for the arrays are approximately equal for strain EF BRIDGE and V24, while for strain NCDO 581 the variance for one of the arrays is approximately twice as large as for the other array.

3.2 Results

Table 3 shows the number of genes predicted as divergent for the different strains using the different predictors. For strains EF BRIDGE and V24 Ratio predicted more genes as divergent than L and T1, while for strains NCDO 581 and JH2SS the results were opposite. For strain NCDO 642 the number of genes predicted as divergent were approximately equal for these three predictors. Notice that T1 and L predicted approximately the same number of genes as divergent. For all the strains, and especially for NCDO 581, but except V24, there was a large difference between the number of genes predicted as divergent by T2 and T1. The strains NCDO 581 and V24 were the strains with the largest and smallest effect of array. The last column in the table shows the total number of genes predicted as divergent by all predictors.

Table 4 shows the number of genes predicted as equally present ($\hat{c} = 1$) for each strain. The difference between Ratio and the regression predictors was relatively large for all strains, especially for strain EF BRIDGE. For four of the strains, strain NCDO 581, EF BRIDGE, JH2SS, and V24, there

Table 3 Number of genes of *Enterococcus faecalis* strains predicted as divergent ($\hat{c} = 0$).

Strain	R	L	T1	T2	# consensus genes
NCDO 581	524	585	588	795	420
NCDO 642	709	700	700	773	621
EF BRIDGE	463	313	317	281	271
JH2SS	281	323	323	259	234
V24	395	350	351	362	338

Table 4 Number of genes of *Enterococcus faecalis* strains predicted as present with equal number of copies ($\hat{c} = 1$).

Strain	R	L	T1	T2	# consensus genes
NCDO 581	2707	2636	2633	2295	2147
NCDO 642	2477	2529	2529	2394	2261
EF BRIDGE	635	808	806	913	610
JH2SS	1156	1103	1107	1313	996
V24	658	749	747	719	596

Table 5 Number of genes of *Enterococcus faecalis* strains predicted to be present in higher copy number ($\hat{c} \geq 2$).

Strain	R	L	T1	T2	# consensus genes
NCDO 581	14	24	24	155	7
NCDO 642	59	16	16	78	6
EF BRIDGE	2147	2124	2122	2051	2026
JH2SS	1808	1819	1815	1673	1585
V24	2192	2146	2147	2164	2086

was a small deviation between T1 and L. The difference between T1 and T2 was large for all strains except V24. The strain V24 had almost no effect of array. The last column is the total number of genes predicted as equally present by all predictors.

Table 5 shows for each strain the number of genes the different predictors predicted as present in higher copy number. The difference between Ratio and L/T1 was not that large here. In strains NCDO 642 all the predictors predicted very few genes as present in higher copy number. Especially for strains with some large ratios, the predictors predicted many genes as present in higher copy number. The last column is the total number of genes predicted as present in higher copy numbers by all predictors.

The difference in the results of L and T1 is ignorable. Notice that T1 and L gave equal numbers for strain NCDO 642, both in divergent, equally present, and present in higher copy number.

4 Simulation Studies

4.1 Simulation design

We performed a simulation study to examine the prediction ability of the different predictors presented above. The data were simulated according to what we believe is close to real data.

Copy number zero to five were drawn from different probabilities denoted p_c , $c = 0, \dots, 5$. Four different values of p_0 , namely 0.05, 0.10, 0.15, and 0.20 were applied, and for each value of p_0 the five different distributions shown in Table 6 were applied. The distributions are ordered after increasing expectation of the copy number. The probability that a gene has copy number c is given by $p_c = (1 - p_0)p_{c|c>0}$.

The $\log_2$ -ratios for gene j ($j = 1, \dots, 2200$) on spot i ($i = 1, \dots, 5$) and array k ($k = 1, 2$), y_{ikj} , were drawn from the normal distribution

$$y_{ikj} \sim N(\log_2(\mu_{c_j}), a_k \sigma^2),$$

Table 6 The conditional probability of copy number c given that the gene is present, $P(C = c|C > 0)$, for distribution D1, ..., D5.

	$P_{c c>0}$				
	$c = 1$	$c = 2$	$c = 3$	$c = 4$	$c = 5$
D1	0.990	0.010	0.000	0.000	0.000
D2	0.830	0.100	0.030	0.025	0.015
D3	0.615	0.260	0.100	0.020	0.005
D4	0.370	0.345	0.185	0.070	0.030
D5	0.230	0.450	0.230	0.050	0.040

where a_1 was equal to one, and the values of a_2 were 1 and $1/6$, the values of σ were 0.25, 0.50, 0.75, and 1.00, and where μ_{c_j} was drawn from the gamma distribution

$$\mu_{c_j} \sim Ga\left(\frac{c_j^2}{\tau^2}, \frac{\tau^2}{c_j}\right).$$

Hence $E(\mu_{c_j}) = c_j$ and $\text{Var}(\mu_{c_j}) = \tau^2$, where the values of τ are 0.25, 0.50, 0.75, and 1.00. For divergent genes $c_j = 0.5$ was used. A smaller value could have been set, as long as it is truly greater than zero. The variance of the replicates in one gene is reflected by the value of $a_k\sigma^2$, where the effect of array is given by the value a_k . The variance between the genes with equal copy number is reflected by the value of τ^2 .

Figure 2 Histogram of data, with and without grouping with respect to copy number, for two different data sets.

The choice of μ_{c_j} as a random parameter is due to the fact that there may exist some discrepancy between the observations and the underlying copy number. Imagine all replicates for one gene have ratios from 1.08 to 1.12, and another gene has replicates from 0.88 to 0.92. In both cases we will probably say that the underlying copy number is one, with a clear discrepancy between the copy number and the observations.

Figure 2 shows histograms of ratios, with and without grouping with respect to copy number, for two different data sets. Both data sets are simulated with $p_0 = 0.05$ and $p_{c|c>0} = [0.230, 0.450, 0.230, 0.050, 0.040]$. Data in Figure 2a are generated from a distribution with $a_k = 1$, $\sigma = 0.25$, and $\tau = 0.25$, while data in Figure 2b are generated from a distribution with $a_k = 1$, $\sigma = 1.00$, and $\tau = 1.00$. Notice that in the latter situation it is hard to separate the different groups from each other, and prediction of copy numbers is more difficult.

Studying the histograms of the simulated data (Figure 2) together with the histograms of the observed data (Figure 1) we see that the data are well comparable.

4.2 Results

In Table 7 RMSEP and EC are presented for different combinations of the variances and no effect of array, while Table 8 presents the results when there was array effect. The results are averaged over all distributions (D1, ..., D5), and all values of p_0 . Since L is approximately equal to T1 for balanced experiments, L was omitted from the simulation study. Differences between the predictors were tested using ANOVA with 5% level of significance (Indahl and Næs, 1998). When the variances were smallest, Ratio gave significantly better results validated by RMSEP, while the regression predictors were better if validated by the criterion EC. When the variances were largest, the regression predictors gave significantly better results validated by both RMSEP and EC. This was also the case when only τ , the standard deviation between genes with equal copy number, was small. In the latter situation the regression predictors did not produce as similar results as they did in the other situations. As expected, T1

Table 7 RMSEP and EC for the different predictors when $a_2 = 1$ (no effect of array).

	RMSEP			EC		
	T1	T2	R	T1	T2	R
$\sigma = 0.25 \wedge \tau = 0.25$	0.48	0.49	0.35	0.13	0.13	0.44
$\sigma = 0.25 \wedge \tau = 1.00$	1.08	1.08	1.08	0.55	0.55	0.72
$\sigma = 1.00 \wedge \tau = 0.25$	0.59	0.71	0.87	0.18	0.20	0.56
$\sigma = 1.00 \wedge \tau = 1.00$	1.08	1.08	1.54	0.55	0.55	0.73

Table 8 RMSEP and EC for the different predictors when $a_2 = 1/6$ (effect of array).

	RMSEP			EC		
	T1	T2	R	T1	T2	R
$\sigma = 0.25 \wedge \tau = 0.25$	0.48	0.47	0.33	0.13	0.13	0.43
$\sigma = 0.25 \wedge \tau = 1.00$	1.07	1.07	1.06	0.55	0.55	0.71
$\sigma = 1.00 \wedge \tau = 0.25$	0.55	0.49	0.63	0.15	0.13	0.50
$\sigma = 1.00 \wedge \tau = 1.00$	1.07	1.08	1.30	0.55	0.55	0.72

Figure 3 RMSEP (a) and EC (b) for the different predictors for different values of σ . RMSEP (c) and EC (d) for the different predictors for different values of τ . RMSEP (e) and EC (f) for the different predictors for different values of p_0 . RMSEP (g) and EC (h) for the different predictors for different distributions. Data were simulated with no effect of array. Ratio (— • —), T1 (— ■ —), and T2 (— ♦ —).

Figure 4 RMSEP (a) and EC (b) for the different predictors for different values of σ . RMSEP (c) and EC (d) for the different predictors for different values of τ . RMSEP (e) and EC (f) for the different predictors for different values of p_0 . RMSEP (g) and EC (h) for the different predictors for different distributions. Data were simulated with effect of array. See Figure 3 for explanation of the symbols.

seemed to be better than T2 when there was no effect of array, while T2 was better than T1 (significantly better when validated by RMSEP) when there was effect of array. In these situations the variance of the $\log_2$ -ratios was large for genes with equal copy number, but the conditional distribution was symmetrical. When only σ , the standard deviation for each gene, was small, the regression predictors gave significantly better results validated by EC, but not by RMSEP. The regression predictors showed a great improvement compared to random classification, even when the data were noisy. (For random classification EC equals 0.67.) On the contrary, Ratio did not improve compared to random classification in situations where the variance was large (and skewed) between genes with equal copy number.

How RMSEP depended on σ and τ is shown in Figures 3a and 3c without array effect and in Figures 4a and 4c with array effect. The results are averaged over all distributions (D1, ..., D5), and all values of p_0 . Ratio was a significantly better predictor than the regression predictors for small values of σ . When there was no effect of array and σ increased (≥ 0.50), T2, and especially T1, gave significantly better results than Ratio. For larger values of σ (≥ 0.75) and with array effect included, all the predictors were significantly different, with T2 as the superior. For small values of τ all the predictors were significantly different, with both T1 and T2 better than Ratio. As expected, T1 was the superior when there was no array effect, while T2 was the superior when there was effect of array. The difference between the regression predictors increased for increasing values of σ , while it decreased for increasing values of τ .

Figures 3b and 3d show how EC altered with different values of σ and τ when there was no effect of array. Figures 4b and 4d show the corresponding figures with effect of array. The regression predictors were significantly better than Ratio for all values of both σ and τ .

The dependence of RMSEP and EC on p_0 , without array effect, is shown in Figures 3e and 3f, and with effect of array in Figures 4e and 4f. The results are averaged over all distributions (D1, ..., D5), and all values of σ and τ . In the case of no array effect the predictors were ranked equally for all values of p_0 , but with the regression predictors significantly better than Ratio. When the predictors were validated by RMSEP, T1 was significantly better than T2 for p_0 equal to 0.05 and 0.10. When there was an array effect, the predictors were equally ranked by RMSEP and EC, with equal ranking for all values of p_0 . The regression predictors were significantly better than Ratio when evaluated by EC. Evaluated by RMSEP, T2 was significantly better than T1 for p_0 equal to 0.10 and 0.15, while for p_0 equal to 0.20, only T2 was significantly better than Ratio.

Figures 3g and 3h show RMSEP and EC for the different distributions introduced in Table 6 when there was no effect of array, while Figures 4g and 4h are with effect of array. The results are averaged over all values of p_0 , σ , and τ . For all distributions and no effect of array the regression predictors gave significantly better results than Ratio with respect to both RMSEP and EC. With array effect, the regression predictors gave significantly better results than Ratio with respect to both RMSEP and EC for all distributions except D4 and D5. For distribution D4, Ratio was significantly better than T2, while T2 was significantly better than T1. For distribution D5 Ratio and T2 were significantly better than T1. The regression predictors gave significantly better results than Ratio for the two latter distributions when validated by EC.

5 Discussion and Conclusion

5.1 Discussion

Predictors based on factor analysis models and linear random models can be used to test for genomic similarities between two strains in comparative genomic hybridization analyses. Making use of these models (among others), we have studied the ability of four predictors to predict the copy number of genes.

The predictors based on linear models achieved better prediction than Ratio in almost all situations. When there was differences between the arrays, either in the mean level of the ratios, or in the var-

iance between genes, the results differ between the predictor that takes this information into account and the other predictors.

The predictors, and especially Ratio, do not have the quality to adjust for systematic deviations from symmetrical distributed intensities. When the variance of the mean of genes with equal copy number increases, the distributions of the intensities get skewed, and the predictors fail to divide the genes into groups of correct copy number.

Since the ratios cannot be negative, all observed ratios for divergent genes have to be greater than or equal to zero, and the assumption about a linear relationship between the copy number and the observed ratio is not true for small copy numbers. Due to this, the intercept is overestimated and the slope is underestimated, and the regression predictors predict larger copy numbers too small. On the contrary, this involves that Ratio predicts the copy numbers too high.

As an attempt to improve the prediction ability by the linear models, we could divide the genes into two groups, divergent and present. Then the divergent genes are removed, and a linear model may be fitted to the group of present genes. Hopefully this will improve the estimation of the intercept and the slope. For higher copy numbers the linear models are probably an advantage, few genes with high copy number make other methods (e.g. methods based on mixture models) less suited.

Another attempt to improve the prediction ability is the use of mixture models, either on the original $\log_2$ -ratio or on the predicted copy numbers on the continuous scale. This might improve the transformation from continuous to discrete copy numbers, but the mixture models also handles the fact that the mode of the graphs not necessarily corresponds to an integer. Another feature of mixture models is the assignment of a posterior probability for each gene belonging to each class (copy number). However, the predicted values of the copy numbers on the continuous scale also contain information about the uncertainty of the predicted copy numbers. As an example a gene with value 2.1 is more likely to have copy number two than a gene with the value 2.4.

In this paper all the model fits are made on the $\log_2$ -ratios instead of the ratios, for the purpose of an approximation to the normal distributions. The models could be fitted using ratio to keep the integer of the copy number, but this will obscure the normality of the data. Methods for model adequacy checking and outlier detection known from analysis of variance and regression analysis could be used to identify replicates that should be considered omitted from the analysis (Montgomery, 1997).

We assumed all genes in the known strain to be in one copy, hence we assumed the true ratio (unknown/known) to be an integer. If this ratio is not an integer, the copy number of the known strain could be modelled as known, and the copy number of the unknown strain could be modelled directly.

Apart from studying both the root mean square error of prediction and the error of classification, we have not discussed the cost of predicting incorrectly. In some situations it may be more serious to predict e.g. a gene as present when it is truly divergent, or vice versa, and the methods should take this into consideration.

5.2 Conclusion

The simulation study showed that predictors based on factor analysis models and linear random models could be used to predict the copy number of a gene. These predictors achieved better results than using the ratio when the data were noisy. If there are effect of array in the data, either in the expected level of the ratios, or in the variance between genes, the predictor based on the factor model that models different expectation and different variance should be applied. If there are no effect of array the predictor based on the linear random model, or the predictor based on the factor analysis model with common mean and variance are to prefer, dependent on whether the experiment is balanced or not.

Acknowledgements We would like to thank two anonymous referees for constructive criticism. This work has been financed by the Norwegian University of Life Sciences.

References

- Autio, R., Hautaniemi, S., Kauraniemi, P., Yli-Harja, O., Astola, J., Wolf, M., and Kallioniemi, A. (2003). CGH-Plotter: MATLAB toolbox for CGH-data analysis. *Bioinformatics* **19**, 1714–1715.
- Behr, M. A., Wilson, M. A., Gill, W. P., Salamon, H., Schoolnik, G. K., Rane, S., and Small, P. M. (1999). Comparative genomics of BCG vaccines by whole-genome DNA microarray. *Science* **284**, 1520–1523.
- Bickel, P. J. and Doksum, K. A. (1977). *Mathematical Statistics: Basic Ideas and Selected Topics*. Prentice-Hall, New Jersey.
- Björkholm, B., Lundin, A., Sillén, A., Guillemin, K., Salama, N., Rubio, C., Gordon, J. I., Falk, P., and Engstrand, L. (2001). Comparison of genetic divergence and fitness between two subclones of *Helicobacter pylori*. *Infection and Immunity* **69**, 7832–7838.
- Brown, P. J. (1993). *Measurement, Regression, and Calibration*. Oxford University Press.
- Feten, G., Almøy, T., and Aastveit, A. H. (2005). Prediction of missing values in microarray and use of mixed models to evaluate the predictors. *Statistical Applications in Genetics and Molecular Biology* **4** No. 1, Article 10. <http://www.bepress.com/sagmb/vol4/iss1/art10>
- Fridlyand, J., Snijders, A. M., Pinkel, D., Albertson, D. G., and Jain, A. N. (2004). Hidden Markov models approach to the analysis of array CGH data. *Journal of Multivariate Analysis* **90**, 132–153.
- Indahl, U. G., and Næs, T. (1998). Evaluation of alternative spectral feature extraction methods of textural images for multivariate modeling. *Journal of Chemometrics* **12**, 261–278.
- Kim, C. C., Joyce, E. A., Chan, K., and Falkow, S. (2002). Improved analytical methods for microarray-based genome-composition analysis. *Genome Biology* **3**, research0065.1–0065.17.
- Lawley, D. N. and Maxwell, A. E. (1973). Regression and factor-analysis. *Biometrika* **60**, 331–338.
- Mantripragada, K. K., Buckley, P. G., de Ståhl, T. D., and Dumanski, J. P. (2004). Genomic microarrays in the spotlight. *Trends in Genetics* **20**, 87–94.
- Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979). *Multivariate Analysis*. Academic Press, London.
- Montgomery, D. C. (1997). *Design and Analysis of Experiments*. John Wiley Sons.
- Nguyen, D. V., Arpat, A. B., Wang, N. Y., and Carroll, R. J. (2002). DNA Microarray Experiments: Biological and Technological Aspects. *Biometrics* **58**, 701–717.
- Paulsen, I. T., Banerjee, L., Myers, G. S. A., Nelson, K. E., Seshadri, R., Read, T. D., Fouts, D. E., Eisen, J. A., Gill, S. R., Heidelberg, J. F., Tettelin, H., Dodson, R. J., Umayam, L., Brinkac, L., Beanan, M., Daugherty, S., DeBoy, R. T., Durkin, S., Kolonay, J., Madupu, R., Nelson, W., Vamathevan, J., Tran, B., Upton, J., Hansen, T., Shetty, J., Khouri, H., Utterback, T., Radune, D., Ketchum, K. A., Dougherty, B. A., and Fraser, C. M. (2003). Role of mobile DNA in the evolution of vancomycin-resistant *Enterococcus faecalis*. *Science* **299**, 2071–2074.
- Pushker, R., Mira, A., and Rodríguez-Valera, F. (2004). Comparative genomics of gene-family size in closely related bacteria. *Genome Biology* **5**, R27.
- Sassetti, C. and Rubin, E. J. (2002). Genomic analyses of microbial virulence. *Current Opinion in Microbiology* **5**, 27–32.
- Searle, S. R. (1987). *Linear Models for Unbalanced Data*. John Wiley Sons.
- Wang, Y. D. and Guo, S. W. (2004). Statistical methods for detecting genomic alterations through array-based comparative genomic hybridization (CGH). *Frontiers in Bioscience* **9**, 540–549.