Genome analysis

Detection of gene copy number changes in CGH microarrays using a spatially correlated mixture model

Philippe Broët1,* and Sylvia Richardson2

1Faculté de Médecine, Université Paris-XI, IFR69, 16 Avenue Paul Vaillant Couturier 94807 Villejuif Cedex, France and 2Department of Epidemiology and Public Health, Imperial College, Norfolk Place, London W2 1PG, UK

Received on December 7, 2005; revised on January 28, 2006; accepted on January 31, 2006
Advance Access publication February 2, 2006
Associate Editor: Martin Bishop

ABSTRACT

Motivation: Comparative genomic hybridization array experiments that investigate gene copy number changes present new challenges for statistical analysis and call for methods that incorporate spatial dependence between sequences along the chromosome. For this purpose, we propose a novel method called CGHmix. It is based on a spatially structured mixture model with three states corresponding to genomic sequences that are either unmodified, deleted or amplified. Inference is performed in a Bayesian framework. From the output, posterior probabilities of belonging to each of the three states are estimated for each genomic sequence and used to classify them.

Results: Using simulated data, CGHmix is validated and compared with both a conventional unstructured mixture model and with a recently proposed data mining method. We demonstrate the good performance of CGHmix for classifying copy number changes. In addition, the method provides a good estimate of the false discovery rate. We also present the analysis of a cancer related dataset.

Supplementary information: http://www.bgx.org.uk/papers.html
Contact: broet@vjf.inserm.fr

1 INTRODUCTION

Comparative genomic hybridization (CGH) microarray is an emerging tool in bioclinical research than allows to identify genomic alterations with a higher resolution than the conventional CGH (Carter et al., 2002). To study aberrations of the genome, investigators competitively hybridize fluorescein-labeled normal and pathological samples to an array containing clones designed to cover certain areas of the genome. Once hybridization has been performed, the signal intensities of the dyes are quantified. Thus, this technique provides a means to quantitatively measure DNA copy-number changes and to map them directly onto a genomic sequence. In oncology, where carcinogenesis is associated with complex chromosomal alterations, CGH arrays can be used for detailed analysis of genomic changes in copy number (in terms of gains or loss of genetic information) in the tumor sample. Indeed, amplification of an oncogene or deletion of a tumor suppressor gene are important steps in elucidating mechanisms for tumorigenesis.

In parallel to the rapid development of CGH microarray technologies, statistical research into ways of analyzing such experimental setups has become an active area. While some authors have proposed methods that do not take into account spatial dependency along the chromosome (Hodgson et al., 2001; Pollack et al., 2002; Cheng et al., 2003; Wang J. et al., 2004), others have devised methods that make use of the spatial location information for detecting gene copy number changes. Break-point detection models have been developed and implemented by Autio et al. (2003), Jong et al. (2004) and Picard et al. (2005). In a different framework, Fridlyand et al. (2004), have proposed an unsupervised hidden Markov chain models approach. More recently, Eilers et al. (2004) have implemented a smoothing algorithm based on penalized quantile regression to visualize effectively the patterns of changes. However, as pointed by Wang et al. (2005), although these methods take into account spatial dependencies along a chromosome, they do not provide at the same time a way to assess classification performance, e.g. by estimating the false discovery rate (FDR) introduced by Benjamini and Hochberg (1995). Since CGH microarray studies typically involve the simultaneous study of thousands of genomic sequences, producing incorrect conclusions must be taken into account in the same way as for transcriptome studies. In this spirit, Wang et al. (2005) recently proposed a data mining approach to select gains and losses, called CLAC, consisting in spatial clustering followed by cluster selection. As a final step, using a reference (normal/normal hybridization) array, the authors proposed to estimate the FDR by the number of genes selected in the reference array (using the same rule) over the number of genes selected in the tumor array.

In this article, we place ourselves in a mixture model based framework (McLachlan and Peel, 2000) and extend in two important directions the three-state (copy gain/copy loss/modal copy) mixture model proposed by Hodgson et al. (2001) and Wang et al. (2004) for analyzing CGH microarrays. First, we take into account the spatial dependence between genomic sequences by allowing the mixture weights to be correlated for neighbouring genomic sequences on a chromosome. Second, we embed our model in a Bayesian framework and use the allocation probabilities to classify each genomic sequence as being deleted, amplified or unmodified. Based on these posterior probabilities, an accurate estimate of the proportion of falsely allocated sequences is obtained. To extend mixture models to spatially indexed data, we link the mixture weights to Markov random fields. This extension is naturally accommodated in the flexible Bayesian hierarchical formulation that underpins our proposed method.
The paper is organized as follows. In Section 2, we present our Bayesian spatially structured mixture model and classification procedure called CGHmix and give details of the spatial modelling of the weights. In Section 3, using simulated realistic datasets, we compare the operational characteristics of our spatially structured mixture model with the conventional (unstructured) one. We also compare the performance of our model with the CLAC algorithm proposed by Wang et al. (2005). In Section 4, we present the implementation of the model on a breast cancer cell line cDNA microarray CGH dataset discussed in the paper by Pollack et al. (2002). We conclude with a discussion on interpretation and extensions of the model.

2 CGHmix: Model and Inference

2.1 Spatial mixture model

Let $Z$ be the random vector of the $N$ genomic sequence (GS) measurements for the $K$ chromosomes, with $n_k$ GS for the $k$-th chromosome ($N = \sum_{k=1}^{K} n_k$). In the following, we denote by $Z_{i,k}$ the random variable corresponding to the log-ratio measurement for the $i$-th genomic sequence ordered along the chromosome $k$. We assume in this work that the sequences are not overlapping. We assume a three-state mixture model: modal copy state (unmodified GS), loss copy state (deleted GS) and gain copy state (amplified GS). To consider a commonly used model that retains the Markov property, we chose to expand the formulation of our mixture model in the case of three latent Markov values of $c$, $c = 1, 2, 3$ with probabilities $\{\omega_c; c = 1, 2, 3\}$ respectively, such that $\sum_c \omega_c = 1$. Here, $c = 1$ corresponds to the loss copy state, $c = 2$ the modal copy state and $c = 3$ the gain copy state.

For each component $c$, we model the log ratio measurement as arising from a normal distribution with mean $\mu_c$ and variance $\sigma^2_c$. Thus, conditionally on the allocations, we have: $f(Z | L_{i,k} = c) \sim N(\mu_c, \sigma^2_c)$ and the marginal density of $Z_{i,k}$ can be written in the form:

$$f(Z_{i,k}) = \sum_c \omega_c f(Z_{i,k} | L_{i,k} = c) \times f(L_{i,k} = c).$$

The quantities $\omega_c$ are the mixing proportions or weights for the GS of chromosome $k$ with $0 \leq \omega_c \leq 1$ and $\sum_c \omega_c = 1$.

By formulating a three-state mixture, we have only translated part of the known biological knowledge (i.e. the existence of loss or gain of genomic material). In addition, we want to incorporate the fact that the probability of allocation of a GS to a state is correlated to that of its neighbours as deletions and gains occur in patches along the chromosome. There are several possible ways to create dependent allocations and it is anticipated that they would give similar results so long as the model chosen is flexible and can take into account localized structure.

We chose to expand the formulation of our mixture model in the framework of Bayesian hierarchical models, a framework that has been shown to be flexible and efficient for mixture modelling (Richardson and Green, 1997) as well as capable of accommodating diverse spatial structures arising in a variety of applications (Green et al., 2003).

### 2.1.1 Spatial structure

To introduce a spatial structure on the weights for each chromosome, in the same spirit as Fernandez and Green (2002), we relate the weights $\omega_{c,i,k}$ to three latent Markov random fields $x_{c,i,k} = \{x_{c,i,k}; 1 \leq i \leq n_k\}$ by a logistic transformation:

$$w_{c,i,k} = \frac{\exp(x_{c,i,k})}{\sum_{c} \exp(x_{c,i,k})}.$$

Here $x_{c,i,k}$ are three independent $n_k$-dimensional latent vectors, each distributed according to the intrinsic Gaussian conditional autoregressive model described below.

To be precise, we consider nearest neighbour Markov random field models where each sequence has two adjacent (preceding and following) neighbours, except for sequences at the ends of the chromosome. Gaussian conditional autoregressive models have been widely used in spatial statistics for describing interaction priors between latent variables at fixed sites, Besag and Kooperberg (1995). One key feature of these distributions is the Markovian interpretation of their conditional distributions. In this paper, we consider a commonly used model that retains the Markov property, which is the intrinsic Gaussian conditional autoregression (ICAR). For such a model, the distribution of each spatial latent variable $x_{c,i,k}$ conditional on $x_{c,i,\{i\},k}$ depends only on the neighbours:

$$x_{c,i,k} | x_{c,\{i\},k} \sim N\left(\frac{1}{m_{i,k}} \sum_{l \in \delta_{i,k}} x_{c,l,k} \frac{\tau_c^2}{m_{l,k}}, \frac{\tau_c^2}{m_{i,k}}\right),$$

where $\delta_{i,k}$ is the set of indices for the neighbours of GS $i$ for chromosome $k$, and $m_{i,k}$ is the corresponding number of neighbours (here one or two). To allow identifiability, we assume $\sum_{c} x_{c,i,k} = 0$ for all $i$. In this model, the parameter $\tau_c^2$ acts as a smoothing prior for the spatial fields and consequently controls the dependence among the weights along the chromosome. In particular, small values of $\tau_c^2$ will correspond to smoother realizations.

By modelling the weights in terms of random fields, we have introduced a large number of parameters (more than the data points). Nevertheless, these parameters are not treated as independent fixed effects. On the contrary, they are linked by the hierarchical structure adopted (here the spatial fields) which leads to borrowing information between the parameters and shrinkage (Carlin and Louis, 2000). The estimation of the weights is thus stabilized.

It is worth noting that the spatial part of the model is specified separately for each of the chromosomes. This choice implies that parameters controlling the switching structure between states are homogeneous across a same chromosome but can be different between chromosomes. On the other hand, the parameters of the mixture densities, $\mu_c$ and $\sigma^2_c$ that characterize losses and gains are common to all the chromosomes, thus borrowing strength from all the data for their estimation.

### 2.1.2 Prior settings

In the Bayesian framework, all unknown quantities are given prior distributions. For the conditional variance $\tau_c^2$, we specify a gamma prior distribution $\Gamma(\epsilon_c, \phi)$ [the gamma distribution is parametrized such as its mean is $\epsilon_c \phi$ and its variance is $(\epsilon_c / \phi^2)$]. In order to have smooth spatial fields (i.e. small $\tau_c^2$), we chose a gamma distribution $\Gamma(10^{-2}, 10^{-2})$ which tends to place most of the prior mass near zero while allowing a reasonably large coefficient of variation. With this specification, neighbouring locations are induced to have similar values of the corresponding spatial latent variables, this effect being more pronounced as the values of $\tau_c^2$ decreases.

The prior distribution for the parameters of the normal distribution of each state cannot be improper since if no genomic sequence is allocated to one of the three states, the posterior would also be improper. The mean parameter $\mu_2$ is set to 0 since it corresponds
to the modal copy state. In practice, if a dye bias is observed, we recommend to use a reference array (normal/normal) to estimate the bias. This quantity is then subtracted from the log-ratio measurements \( Z_{i,k} \). For unique labelling, we impose further that \( \mu_1 < 0 \) and \( \mu_2 > 0 \). We chose flat prior for the means in the observed range of the data \([a - b], a < 0, b > 0 \) so that \( \mu_1 \sim U(a, 0) \) and \( \mu_2 \sim U(0, b) \). For the component variances \( \sigma_i^2 \), we consider an inverse gamma distribution \( \Gamma(\alpha, \beta) \) with \( \alpha = \beta = 10^{-1} \).

The model can easily accommodate a few missing values of genomic sequences. Indeed in the Bayesian framework, missing data are treated as unknown parameters and the missing values are simulated from the joint posterior distribution. The only modification necessary in the code is to declare the missing values of \( Z_{i,k} \) as NA and to initialize them in the Monte Carlo Markov chain (MCMC) algorithm described below.

### 2.2 Inference and FDR estimation

Inference for parameters of interest is undertaken by sampling from their joint posterior distributions using MCMC samplers implemented in the WinBUGS software (Spiegelhalter et al., 2004, http://www.mrc-bsu.cam.ac.uk/bugs). All results presented correspond to 10,000 sweeps of MCMC algorithms following a burn-in period of 10,000 (period for achieving stability of the algorithm). Convergence was checked both by visual inspection of the traces plots for the parameters of the mixtures and by checking reproducibility of the results between different starting values (Gilks et al., 1996, chap. 5). Summary statistics are calculated from the full output of the MCMC algorithm, after discarding the burn-in samples. Furthermore, the samples provide information on quantities of prime interest, the posterior probabilities \( \hat{p}_{c, i, k} \) of belonging to each state for each GS \( i \) of chromosome \( k \). These posterior probabilities can be directly estimated as empirical averages of the weights from the output of the algorithm and plotted as a function of the order of the sequence along chromosome.

Using these estimates, various probabilistic clustering and classification of the data can be achieved. In the following, we compare the results obtained considering two different allocation rules. First, we propose to assign a sequence \( i \) of chromosome \( k \) to a modified state, \( c = 1 \) (copy loss) or \( c = 3 \) (copy gain) if its corresponding posterior probability \( \hat{p}_{c, i, k} \) is above a threshold value \( \hat{p}_{\text{cut}} \) (\( c = 1, 3 \)), otherwise the sequence \( i \) is assigned to the modal copy state \( c = 2 \). As seen in what follows, this allocation rule is quite flexible since it can be tailored for any target FDR. Second, and to avoid having to define a threshold value \( \hat{p}_{\text{cut}} \), we also consider the Bayes allocation rule which assigns each sequence to the component of the mixture to which it has the highest probability of belonging.

An advantageous feature of our model is that we can consider a direct posterior probability approach (Newton et al., 2004; Broët et al., 2004) to estimate the FDR (the expected proportion of false discoveries) for any subset of genomic sequences. Here, a false discovery (or false positive) corresponds to a genomic sequence which is claimed to be modified (i.e. assigned to a modified state based on the allocation rule) whereas it is not. For the entire set of genomic sequences represented on the array or for any given subset (e.g. a chromosome), the FDR can be estimated as follows. For a set \( S \) of \( m \) genomic sequences classified as modified (with subsets \( S_1 \) and \( S_2 \) for copy gain and loss states, respectively), the estimated FDR is equal to \( 1/m \sum_{i \in S_1} (1 - \hat{p}_{1, i, k}) + 1_{i \in S_2} (1 - \hat{p}_{3, i, k}) \).

### 3 PERFORMANCE EVALUATION AND COMPARISON

There are two novel aspects to our procedure: (1) the use of spatially structured instead of conventional (unstructured) mixture models and (2) the use of a fully Bayesian model-based approach rather than a data mining approach that favours contiguous clusters. To evaluate the performance of the spatial mixture model, we have designed a simulation study that compare CGHmix with both the unstructured mixture model and with CGHMiner. Without loss of generality, the simulation study mimics a CGH micro-array with a single fake chromosome having copy gains and losses.

To be precise, by conventional (unstructured) three-state mixture model we mean that the marginal density of \( Z_{i,k} \) can be written in the form: \( f(Z_{i,k}) = \sum_{k=1}^{3} w_{c,k} \times f_i(\cdot | \mu_k, \sigma_k^2) \).

Here, the class weights \( w_{c,k} \) are constant by chromosome, not indexed by the sequences and follow a Dirichlet prior distribution \( D(\delta, \beta, \delta) \) with parameter \( \delta \). The parameter of the Dirichlet distribution was set to the usual value of 1 that corresponds to uniform weights.

#### 3.1 Simulation set-up

Two configurations were generated to represent 200 fake genomic sequences with normalized log-ratio measurements \( Z_i \) \((i = 1, \ldots, 200)\). Conditionally on being modified (deleted, amplified) or unmodified, values were sampled from normal distributions \( N(\mu_a, \sigma_a^2) \) or \( N(\mu_b, \sigma_b^2) \) respectively for amplified or deleted genomic sequences. This setup is the one considered by Wang et al. (2005). In this configuration 30% of the GS are modified with pattern of modification along the fake chromosome as follows: \([1–40] \) unmodified GS, \([41–70] \) deleted GS, \([71–130] \) amplified GS. Parameters for the distributions were chosen such as: \( Z \sim N(0, 0.3^2) \) for unmodified genes and \( Z \sim N(\log(2/1), 0.3^2) \) or \( Z \sim N(\log(1/2), 0.3^2) \) respectively for amplified or deleted genomic sequences. This setup is the one considered by Wang et al. (2005). In this configuration 30% of the GS are modified with pattern of modification along the fake chromosome as follows: \([1–40] \) unmodified GS, \([41–70] \) deleted GS, \([71–130] \) amplified GS. Parameters for the distributions were chosen such as: \( Z \sim N(0, 0.3^2) \) for unmodified genes and \( Z \sim N(\log(2/1), 0.3^2) \) or \( Z \sim N(\log(1/2), 0.3^2) \) respectively for amplified or deleted genomic sequences.

The varied length of 10, 20, 30, 40 and 50 modified sequences, as well as the difference in the overall fraction of modified sequences will allow us to study the sensitivity of our procedure to the size of the modified regions. 50 replicated datasets were simulated for the two configurations. In each case, we also simulated a reference array with 200 fake genomic sequences sampled from \( N(0, 0.3^2) \). This array is used by CGHMiner for computing the FDR.

#### 3.2 Results

##### 3.2.1 Performance of the spatial versus the unstructured mixture model

Figure 1 (panels 1, 2, 4, 5) displays the operational characteristics of the spatial (CGHmix) and the unstructured mixture...
models for the first and second configuration. To be precise, we plotted the true positive fraction (TPF: the number of true positives found divided by the number of truly modified sequences) and the true negative fraction (TNF: the number of true negatives found divided by the number of truly unmodified sequences) as a function of the threshold $p_{\text{cut}}$ for the first allocation rule as defined in the previous section. Each point of the graphs is obtained by averaging the results over the 50 simulations. The panels 3 and 6 display comparative ROC curves (the true positive fraction versus the false positive fraction), where the false positive fraction is the number of false positives divided by the number of truly unmodified sequences.

As seen in Figure 1, the spatial mixture model gives strikingly better results [higher true positive fraction (TPF) and higher true negative fraction (TNF)] than the unstructured mixture model. This is well highlighted in the comparison of the ROC curves (panels 3 and 6) where we see that for both configurations the ROC curves of the spatial mixture model is well above the unstructured one and rises sharply. Moreover, the operational characteristics of the spatial model are still excellent even with more overlap between the intensity of the three states (config. #2). These results clearly show the benefit of using spatial modelling on the mixture weights in order to gain sensitivity (true positives) while controlling the false positive rate. Note that the curves in panels 1 and 4 cross at values of $p_{\text{cut}}$ respectively equal to 0.79 and 0.7. These values strike a good balance between TPF and TNF well above 90% and will be considered as good candidates for the selection rule in the analysis that follows.

Figure 2 display the estimated and realized FDR (for the two configurations) obtained from the spatially structured (panels 1 and 3) and unstructured model (panels 2 and 4). Here, the realized FDR is the quotient of the number of false discoveries to the total number of discoveries (i.e. sequences claimed as modified by the allocation rule). These comparisons show that the spatial model gives good estimate of the FDR for values of the realized FDR up to 10% and has a small downbias (maximum 10%) for higher values. Values of FDR $\leq 10\%$ correspond to those used in practice for selecting GS that are worth investigating with complementary biological methods. In contrast, the unstructured mixture model overestimates the realized FDR. Note that for the recommended values of $p_{\text{cut}}$ discussed above, we obtained reasonable realized FDR of 3.4 and 7.7% for configuration 1 and 2 respectively.

![Figure 1](http://bioinformatics.oxfordjournals.org/)

**Fig. 1.** The panels 1, 2, 4, 5 display the operational characteristics of the spatial and the unstructured mixture models as a function of $p_{\text{cut}}$ for the first (top row) and second (bottom row) configurations. Panels 3 and 6 display the ROC curves for the first and second configurations.
A graphical display of the posterior probabilities for each state obtained in one simulated dataset (config. #2) using the spatial mixture model is given in Figure 3. We note the smooth pattern of the posterior probabilities of allocation to each state obtained under the spatial mixture model which helps classifying correctly the genomic sequences.

We also calculated TPF, TNF and FDR (realized and estimated) obtained for the Bayes allocation rule (Table 1). Again, in each case, TPF and TNF are higher for the spatial model. It is worth pointing out that the Bayes rule leads to higher realized FDR values than those which would be obtained by using our first allocation rule with reasonably high $p_{cut}$ values, such as those already discussed, e.g. 0.79 for configuration 1 and 0.7 for configuration 2. On balance, these results lead us to prefer the first allocation rule based on $p_{cut}$ which allows to target small values of the FDR.

3.2.2 Performance comparison with CGHMiner

As a first point of comparison, we report in Table 2 the results obtained using our spatial mixture model (CGHmix) with the values of $p_{cut}$ discussed just above and those obtained by CGHMiner with the proposed default target FDR value (which is 1% for the CGHMiner Excel plug-in, Wang et al., 2004, http://www-stat.stanford.edu/~wp57/CGH-Miner/).

In the two configurations, CGHmix outperforms CGHMiner in terms of mean number of false positives. These comparisons also show that CGHMiner substantially underestimates the realized FDR associated with the data mining procedure whereas CGHmix gives accurate estimates of the realized FDR. This latter finding is not surprising since we previously showed that the spatial mixture has good estimation properties for low values of the FDR. An illustration of the performance of the two procedures for one simulated dataset (config. #1) is provided as Supplementary Material (Figure I).

To further compare our approach with that of CGHMiner, we calibrated our threshold value $p_{cut}$ so that the realized FDR are equal for both procedures. As the realized FDR values for CGHMiner are large, this obliges us to use low $p_{cut}$ values, 0.27 for configuration 1 and 0.55 for configuration 2 (Table 3).

Even for such inappropriate $p_{cut}$ values, our procedure gives better TPF results with some underestimation of the realized FDR, as noted before for such range of FDR values.

4 ANALYSES OF CGH-ARRAY CANCER DATA SET

As an illustration, we analysed the cDNA microarray CGH study discussed in the paper by Pollack et al. (2002), a dataset that is provided in CGHMiner. We consider the first normal reference array and the MCF7 cell line array, and investigate the 23 chromosomes with 6691 cDNA sequence (with 264 missing data). Data were already log-transformed and normalized. For each genomic sequence, we calculated the difference between MCF7 cell line and normal reference measurements (dye bias correction).

For this multiple chromosomes dataset (see Figure II, Supplementary Material), the estimated means and standard
deviation for the loss and gain states are $\mu_1 = -0.35$, $\mu_3 = 0.41$ and $\sigma_1 = 0.37, \sigma_3 = 0.54$. The variability of the modal component is estimated as $\tilde{\sigma}_2 = 0.27$. From the estimated posterior probabilities, we allocated each sequence to one of the three states based on our first allocation rule. We used a $p_{cut}$ value of 90% in order to obtain a small estimated FDR close to the FDR reported by CGHMiner.

For the whole dataset, the estimated FDR reported by CGHMiner is 1.5% and that reported by CGHmix is 2.3%. A graphical display of the GS allocations for both methods is provided as Supplementary Material, see Figure III and IV). Both methods (CGHMiner and CGHmix) found commonly observed genomic changes described for MCF7 cell lines using classical chromosomal CGH (Hiorns et al., 2004) such as gains of 1q, 8q, 12q, 17q and 20q together with losses of 1p, 9p and X. However, CGHmix found

---

### Table 1. Bayes allocation rule

<table>
<thead>
<tr>
<th>Configuration</th>
<th>Spatial (CGHmix)</th>
<th>Unstructured (CGHmix)</th>
<th>Spatial (CGHmix)</th>
<th>Unstructured (CGHmix)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TPF %</td>
<td>99.7</td>
<td>57.0</td>
<td>96.7</td>
<td>62.9</td>
</tr>
<tr>
<td>TNF %</td>
<td>93.8</td>
<td>86.1</td>
<td>77.6</td>
<td>56.1</td>
</tr>
<tr>
<td>Realized FDR (%)</td>
<td>14.2</td>
<td>18.0</td>
<td>17.6</td>
<td>35.4</td>
</tr>
<tr>
<td>Estimated FDR (%)</td>
<td>7.7</td>
<td>14.0</td>
<td>9.8</td>
<td>41.3</td>
</tr>
</tbody>
</table>

Average results on 50 simulations.

---

### Table 2. Comparative performance between CGHmix and CGHMiner

<table>
<thead>
<tr>
<th>Configuration</th>
<th>CGHmix ($p_{cut} = 0.79$)</th>
<th>CGHMiner (default setting)</th>
<th>CGHmix ($p_{cut} = 0.70$)</th>
<th>CGHMiner (default setting)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Realized false positive (mean)</td>
<td>1.9</td>
<td>16.4</td>
<td>8.1</td>
<td>16.0</td>
</tr>
<tr>
<td>Realized false positive (range)</td>
<td>0-20</td>
<td>3-39</td>
<td>0-50</td>
<td>1-48</td>
</tr>
<tr>
<td>Realized FDR (%)</td>
<td>3.4</td>
<td>23.7</td>
<td>7.7</td>
<td>14.1</td>
</tr>
<tr>
<td>Estimated FDR (%)</td>
<td>2.9</td>
<td>12.2</td>
<td>7.5</td>
<td>0.3</td>
</tr>
</tbody>
</table>

Average results on 50 simulations.

---

**Fig. 3.** The panels 1–3 present the posterior probabilities of allocation to each state obtained with the spatial mixture model for one of the dataset simulated under configuration 2. The vertical bar indicates the breakpoint between the true states.
Table 3. Comparative performance between CGHmix and CGHMiner

<table>
<thead>
<tr>
<th></th>
<th>Configuration 1</th>
<th>Configuration 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CGHmix (p_{cut} = 0.27)</td>
<td>CGHMiner (default setting)</td>
</tr>
<tr>
<td>Realized FDR (%)</td>
<td>23.7</td>
<td>14.0</td>
</tr>
<tr>
<td>Estimated FDR (%)</td>
<td>14.7</td>
<td>11.0</td>
</tr>
<tr>
<td>TPF %</td>
<td>99.9</td>
<td>96.0</td>
</tr>
<tr>
<td>TNE %</td>
<td>86.0</td>
<td>83.0</td>
</tr>
</tbody>
</table>

Average results on 50 simulations.

additional abnormalities, frequently observed in MCF7 cell lines (Hornes et al., 2004), such as two amplicons on chromosome 1q and several amplicons along chromosome 3. Considering the Bayes allocation rule leads to select larger areas and more genomic sequences but with a high estimated FDR of 12.2%.

In order to avoid the arbitrariness of the choice of a threshold value p_{cut}, the investigator can plot the posterior probabilities for any chromosomal segment of interest (e.g. see Figure V in Supplementary Material which display the posterior allocation probabilities for the spatially structured model for chromosome 1).

5 DISCUSSION

We propose in this article the procedure CGHmix based on a spatially structured three-state mixture model well-suited for identifying regions in the genome which undergo gains or losses of genomic material using CGH microarrays.

As usual for model-based analyses, one of the main issue is what model would provide a useful description of the data. In contrast to transcriptional changes for which mixtures can be flexible models for describing finely the dynamic of changes in expression levels (Broët et al., 2002), for large-scale genomic alteration experiments, the three states have a direct interpretation in terms of copy-number changes. Moreover, the proposed spatially structured mixture formulation is grounded on the physical changes of genomic segments that appear in nearby areas. Note that for transcriptome-oriented studies, the idea of incorporating biological knowledge by structuring the weights and modifying the standard mixture model formulation has been recently discussed by Pan (2005) and Xiao et al. (2005), the latter work using a spatial hidden Markov model to better estimate levels of transcription when these are linked to gene location.

In this work, we proceed from a Bayesian approach which allows to estimate posterior probabilities of state membership. CGHmix provides a practical way of clustering the genes based on these posterior probabilities, as well as an estimation of the associated FDR that will guide researchers interested in targeting chromosomal areas that are worth further investigations.

As seen in the simulation study, our approach has a very good performance for revealing gains or losses of contiguous sequences while simultaneously estimating the FDR. By comparison to a standard mixture formulation, using spatially structured weights in the mixture model increases operational characteristics. Moreover, we showed that CGHmix outperforms the hierarchical clustering algorithm procedure proposed by Wang et al. (2005). We also showed that the spatially structure mixture model gives good estimate of the FDR for values of the realized FDR up to 10% and has a small downward bias for higher values. The discrepancies between the realized and estimated FDR are related to the fact that our FDR estimate is obtained from a model-based approach which relies on the accuracy of the fitted model.

In this work, we have presented results for two different allocation rules. In practice, we recommend to use the rule based on classifying a genomic sequence into a modified state if its posterior probability of allocation is above a threshold value rather than the Bayes allocation rule since the former rule is flexible and can be tailored to any target FDR.

Since we considered a homogeneous spatial structure for each chromosome, for areas with many switches between short and long modified sequences it might be feared that the prior structure will lead to over-smooth these changes. However, as seen for the chromosome 3 from the MCF7 cell line, our method identified several short amplicons together with large deletion areas.

For a practical use, computations for CGHmix are performed using the multi-purpose freely available Bayesian software WinBugs (Spiegelhalter et al., 2004). With only few lines of code, a mixture with spatially structured weights can be specified. As with any Bayesian analysis, the choice of prior distributions and hyperparameters has to be carefully thought through and it is advisable to carry out sensitivity studies. We have run our model using different setups for the parameters of the mixture model components as well as the conditional variance of the Markov random fields and we found that posterior probabilities and allocation are robust to moderate perturbations in the prior.

A nice feature of our approach is that several extensions of the model can be envisaged that rely on the generic advantages of using Bayesian hierarchical models (Best et al., 1996). If the exact genomic locations are known, the proposed mixture model could be extended to incorporate the distance between genomic sequence location as a parameter of the Markov random fields. For overlapping BACs it might be additionally useful to distinguish two types of neighbours (overlapping and contiguous) with different weights.

In conclusion, the specificity of comparative genomic hybridization experiments strongly advocates the use of model-based approaches tailored to incorporate spatial dependencies and our novel Bayesian mixture modelling offers an efficient tool for detecting copy number changes.

ACKNOWLEDGEMENTS

Discussions of this work started while the authors were visiting the Institute for Mathematical Sciences, National University of Singapore in 2004. The visit was supported by the Institute and by a grant (No. 01/1/21/19/217) from BMRC-A*STAR of Singapore. This work was supported by Alliance: a Franco-British partnership programme (grant PN03060), Entente Cordiale Cancer Research Bursaries (grant 3), and by the BBSRC ‘Exploiting genomics’ (grant 28/EGM16093).

Conflict of Interest: none declared.

REFERENCES


