BIOINFORMATICS

Vol. 20 no. 18 2004, pages 3413–3422
doi:10.1093/bioinformatics/bth418

Analysis of array CGH data: from signal ratio to gain and loss of DNA regions

Philippe Hupé[1,2,*], Nicolas Stransky[2], Jean-Paul Thiery[2], François Radvanyi[2] and Emmanuel Barillot[1]

[1]Service Bioinformatique and [2]UMR 144 CNRS/Institut Curie, 26, rue d'Ulm, Paris, 75248 cedex 05, France

Received on March 19, 2004; revised on June 18, 2004; accepted on July 12, 2004
Advance Access publication September 20, 2004

ABSTRACT

Motivation: Genomic DNA regions are frequently lost or gained during tumor progression. Array Comparative Genomic Hybridization (array CGH) technology makes it possible to assess these changes in DNA in cancers, by comparison with a normal reference. The identification of systematically deleted or amplified genomic regions in a set of tumors enables biologists to identify genes involved in cancer progression because tumor suppressor genes are thought to be located in lost genomic regions and oncogenes, in gained regions. Array CGH profiles should also improve the classification of tumors. The achievement of these goals requires a methodology for detecting the breakpoints delimiting altered regions in genomic patterns and assigning a status (normal, gained or lost) to each chromosomal region.

Results: We have developed a methodology for the automatic detection of breakpoints from array CGH profile, and the assignment of a status to each chromosomal region. The breakpoint detection step is based on the Adaptive Weights Smoothing (AWS) procedure and provides highly convincing results: our algorithm detects 97, 100 and 94% of breakpoints in simulated data, karyotyping results and manually analyzed profiles, respectively. The percentage of correctly assigned statuses ranges from 98.9 to 99.8% for simulated data and is 100% for karyotyping results. Our algorithm also outperforms other solutions on a public reference dataset.

Availability: The R package GLAD (Gain and Loss Analysis of DNA) is available upon request

Contact: glad@curie.fr

INTRODUCTION

Array Comparative Genome Hybridization (array CGH) is a recently developed technology based on DNA microarrays (Pinkel *et al*., 1998; Snijders *et al*., 2001; Solinas-Toldo *et al*., 1997; Ishkanian *et al*., 2004) and dedicated to the investigation and mapping of changes in DNA copy number. The array generally consists of spotted genomic sequences inserted into bacterial artificial chromosomes (BACs), e.g. (for ease of notation, we will refer to genomic sequences as BACs): each sample DNA is labeled with a fluorescent dye and the reference DNA is labeled with another fluorescent dye. This mixture is then hybridized to the array CGH. Typical applications of arrays CGH are cancer studies since chromosome aberrations frequently occur during tumor progression (Albertson *et al*., 2003) and human genetic disease research (Albertson and Pinkel, 2003; Shaw-Smith *et al*., 2004). In cancer studies, tumor DNA samples are compared with a normal reference DNA sample. The normal sample should have two copies of each genomic region (with the exception of the non-pseudo-autosomal regions of sexual chromosomes, for which a single copy is expected in males), whereas tumor DNA may present a loss or gain of DNA regions. In the simplest case, for a diploid tumor, the loss of a region will result in there being 0 or 1 copy whereas the gain of a region will result in there being three or more copies (the reality is more complex because a tumor is often not diploid). Measurement of the signal intensities of the reference and tumor samples for each BAC should make it possible to determine which regions have been gained or lost in the tumor sample.

Once a microarray has been constructed and hybridization carried out, several steps must be completed to determine which regions have been gained or lost: image acquisition, image analysis (including gridding, spot addressing, spot segmentation, spot quantification and outlier detection), signal normalization (e.g. to correct for systematic spatial or intensity biases) and duplicate treatment (each BAC is generally spotted in several copies to make possible statistical assessment of confidence). Once these steps have been completed, a synthetic value for the signal ratio is obtained, corresponding to the amount of DNA in the BAC concerned in the tumor with respect to that in the reference sample. The regions gained and lost can then be inferred from the ratio profile. Finally, correlation of the loss and gain profiles for a sufficiently high number of tumor samples should provide insight into the regions involved in tumorigenesis or tumor

*To whom correspondence should be addressed.

progression: oncogenes are likely to be present in the regions gained and tumor suppressor genes in the regions lost.

In this study, we assume that signal ratios for each BAC, such as those provided by SPOT 2.0 (Jain *et al*., 2002) or GenePix (Axon Instruments, 2003) software, are available and we focus on the problem of identifying the regions gained and lost from the ratio profile. Let us define the *status* of a homogeneous genomic region as the number of copies of the DNA of this region (here homogeneous means that all points in the region have the same DNA copy number) and Maximum Spanning Homogeneous Region (MSHR) as a region of homogeneous DNA status bordered either by a chromosome end or by another region of different status. Plots of BAC ratios (in fact, we use the ratio base 2 logarithm $\log_2$-ratio) versus BAC position (or rank) along the genome typically generate patterns in which MSHRs should be composed of spots distributed around a mean value that characterizes the status (cf. Fig. 1). Two adjacent MSHRs are separated by a breakpoint. Our approach can be broken down into two main steps: the detection of breakpoints and the assignment of a status to each MSHR. In some cases, a point deletion or amplification may affect the DNA, appearing on the ratio profile as an outlier among BACs with the same DNA status. This special case needs a particular treatment, called outlier detection.

In the absence of experimental biases, ratios should be 0 for double loss, $\frac{1}{2}$ for a single loss, 1 for the normal situation, $\frac{3}{2}$ for a single gain and more generally $\frac{n}{2}$ for a sample with n copies of DNA. In practice, microarray experiments are subject to various sources of variation, including differences in incorporation efficiency between the two fluorescent dyes, an intensity-dependent effect and a print-tip effect, as reported by Yang *et al*. (2001) for expression data. These variations create noise and bias the theoretical values. In addition, tumor biopsy samples generally contain a mixture of normal and tumor cells, and tumor cells may even present heterogeneity in terms of genome losses and gains, corresponding to different stages of tumor progression; these heterogeneities result in smaller signal gaps between regions.

To our knowledge, only two articles have dealt with the problem of breakpoint detection and none have considered the question of region assignment. Jong *et al*. (2003) used a genetic algorithm and local optimization to detect breakpoints. The algorithm developed by Olshen and Vankatraman (2002) is based on circular binary segmentation, as described by Sen and Srivastava (1975). This paper is organized as follows: we will begin by describing our breakpoint detection algorithm; we will then present the region assignment method, followed by a validation of our approach based on simulations, karyotyping results, loss of heterozygosity (LOH) (Vogelstein *et al*., 1989) and manually analyzed data. Finally, we discuss the result obtained and perspectives.

BREAKPOINT DETECTION

The problem of chromosomal breakpoint detection can be approached by estimating a piecewise constant function defining each MSHR of the chromosome. A solution to this problem of estimation has been proposed by Polzehl and Spokoiny (2002), with application in two dimensions to image segmentation. We present here the main principles of their algorithm—adaptive weights smoothing (AWS)—and describe how this algorithm should be applied to chromosomal breakpoint detection with array CGH data. The AWS procedure is an iterative, data-adaptive smoothing technique that was designed for smoothing in regression problems involving discontinuous regression function. It is assumed that the regression function can be approximated, e.g. by a simple local constant model. The regression function is estimated as a weighted maximum-likelihood estimate (MLE), with weights chosen in a completely data-adaptive way. The algorithm finds, around each point, the maximal neighborhood in which the local constant assumption holds. In our case, the maximal neighborhood of every BAC should allow us to delineate in a straightforward manner the MSHRs and the parametric estimation should provide its copy number. The procedure has a number of features of potential value for our problem: it has been shown to preserve contrasts and edges between regions (and should therefore detect breakpoints accurately), it requires very little prior information about the data to model and has a numerical complexity of nM with n the number of points (BACs) and M the size of the maximum neighborhood. The AWS is more general than simple piecewise constant function estimation, but it is straightforward to restrict it to our case.

Statistical model

Let us consider a series of N independent observations $S = \{(X_1, Y_1), \ldots, (X_N, Y_N)\}$ in which each X_i is valued in a metric space $\mathcal{X}$ and determines the location (the BAC rank on the chromosome) and each Y_i is valued in another metric space $\mathcal{Y}$ and is the observation at X_i (the measured $\log_2$-ratio); the locations X_i are ordered such that $X_1 < \cdots < X_i < \cdots < X_N$. We also assume that the observation Y_i depends on the location X_i via a parameter $\theta \in \Theta$, where Θ is a subset of a finite-dimensional space $\mathbb{R}^d$. Conditionally on $X_i = x$, the random variable Y_i is distributed with the density probability function $p[y, \theta(x)]$ for some unknown $\theta(x)$ on $\mathcal{X}$ valued in Θ. Here, we consider the local constant gaussian regression model $Y_i = \theta(X_i) + \epsilon_i$, where the ϵ_i are i.i.d. $\mathcal{N}(0, \sigma^2)$. We wish to infer the function θ such that θ is of the form $\theta(x) = \sum_{m=1}^{M} a_m \mathbf{1}(x \in \mathcal{X}_m)$ with disjoint regions $\mathcal{X}_1, \ldots, \mathcal{X}_M$ and $\mathcal{X} = \bigcup_{m=1}^{M} \mathcal{X}_m$. The regions $\mathcal{X}_m$, the values a_m and even the total number of regions M are unknown. Estimation of the parameter θ is a local estimation problem in that this parameter depends on the location.

The AWS procedure

The approach used for the local estimation of θ is based on local-likelihood modeling (Polzehl and Spokoiny, 2002) and extends the AWS procedure proposed by Polzehl and Spokoiny (2000). An iterative algorithm finds, around every location X_i, the maximal possible neighborhood in which the parameter θ is constant: a weight $w_{ij}(0 \leq w_{ij} \leq 1)$ is assigned to every observation Y_j at X_j, which depends on the previous step of the algorithm. The weighted MLE $\hat{\theta}(X_i) = \hat{\theta}_i$ is of the form:

$$\hat{\theta}_i = \operatorname*{argsup}_{\theta \in \Theta} L(W_i, \theta, \theta')$$

with

$$L(W_i, \theta, \theta') = \sum_{j=1}^{N} w_{ij} \log \frac{p(Y_j, \theta)}{p(Y_j, \theta')},$$

where θ' is an arbitrary point in Θ and $W_i = \operatorname{diag}\{w_{i1}, \ldots, w_{iN}\}$.

At each iteration k, the geometric increase in $h^{(k-1)}$ by a growth rate $a > 1$ defines a new larger neighborhood around each X_i, which is used to calculate the MLE of θ_i. New weights are calculated by means of a location penalty kernel function K_l, which takes into account the proximity of the X_j's in the neighborhood, and a statistical penalty kernel function K_s, which takes into account the comparison of two local models. The kernels K_s and K_l are non-increasing functions and must fulfill $K_s(0) = K_l(0) = 1$. Moreover, a parameter λ controls the statistical penalty and a memory parameter $\eta(0 \leq \eta \leq 1)$ is used to stabilize the procedure. The detail of the procedure is given below (see Polzehl and Spokoiny, 2002):

(1) *Initialization*: Calculate the global MLE $\hat{\theta}^{(0)}$ of θ:

$$\hat{\theta}^{(0)} = \operatorname*{argsup}_{\theta \in \Theta} \sum_{i=1}^{N} \log p(Y_i, \theta) = \frac{1}{N} \sum_{j=1}^{N} Y_j.$$

For every $i = 1, \ldots, N$, set $\hat{\theta}_i^{(0)} = \hat{\theta}^{(0)}$ and define $W_i^{(0)}$ as the unit matrix. Set $k = 1$.

(2) *Iteration*: for every $i = 1, \ldots, N$

(a) *Calculate the adaptive weights*: For every point X_j, calculate the penalties

$$\begin{aligned} \boldsymbol{l}_{ij}^{(k)} &= |\rho(X_i, X_j)/h^{(k)}|^2, \\ \boldsymbol{s}_{ij}^{(k)} &= \lambda^{-1} \Big[L\left(W_i^{(k-1)}, \hat{\theta}_i^{(k-1)}, \hat{\theta}_j^{(k-1)}\right) \\ &\quad + L\left(W_j^{(k-1)}, \hat{\theta}_j^{(k-1)}, \hat{\theta}_I^{(k-1)}\right) \Big] \Big/ 2, \end{aligned}$$

where $\rho(x, x')$ is a metric in $\mathcal{X}$ and $h^{(k)}$ controls the size of the neighborhood of each X_i.

calculate

$$\tilde{w}_{ij}^{(k)} = K_l\left(\boldsymbol{l}_{ij}^{(k)}\right) K_s\left(\boldsymbol{s}_{ij}^{(k)}\right)$$

and define the weight $w_{ij}^{(k)}$ as

$$w_{ij}^{(k)} = \eta w_{ij}^{(k-1)} + (1 - \eta)\tilde{w}_{ij}^{(k)}.$$

Denote by W_i^k the diagonal matrix $W_i^k = \operatorname{diag}\left\{w_{i1}^{(k)}, \ldots, w_{iN}^{(k)}\right\}$.

(b) *Estimation:* Calculate the new local MLE $\hat{\theta}_i^{(k)}$ of θ_i

$$\hat{\theta}_i^{(k)} = \operatorname*{argsup}_{\theta \in \Theta} L\left(W_i^{(k)}, \theta, \theta'\right).$$

(3) *Stopping*: Stop if $ah^{(k)} > h^*$, otherwise increase k by 1, set $h^{(k)} = ah^{(k-1)}$ and continue with step 2.

According to the assumption of our local constant gaussian model we have:

$$\hat{\theta}_i = \min_{\theta \in \Theta} \frac{1}{2\sigma^2} \sum_{j=1}^{N} w_{ij}(Y_j - \theta)^2.$$

$$L(W_i, \hat{\theta}_i, \theta') = \frac{\sum_{j=1}^{N} w_{ij}}{2\sigma^2}(\hat{\theta}_i - \theta')^2.$$

For the local constant gaussian regression model, the AWS procedure requires the parameter σ to be known. An estimate of σ is given by:

$$\frac{IQR(Z_1, \ldots, Z_{N-1})}{IQR(\mathcal{N}(0,1)) \times \sqrt{2}}, \tag{1}$$

where $Z_i = Y_{i+1} - Y_i$ and IQR defines the interquartile range.

The results of the AWS procedure provide one estimate of $\hat{\theta}_i$ for every $i = 1, \ldots, N$. Based on these estimates, we define a breakpoint as a location X_i such that $\hat{\theta}_i \notin [\hat{\theta}_{i+1} - \epsilon; \hat{\theta}_{i+1} + \epsilon]$ (in our case, $\epsilon = 10^{-2}$). Thus, a breakpoint corresponds to the last position of an MSHR. The chromosome can be split into $N' + 1$ MSHRs for a total number N' of breakpoints: $(X_1, \ldots, X_{B_1}), (X_{B_1+1}, \ldots, X_{B_2}), \ldots, (X_{B_{N'}+1}, \ldots, X_N)$. Note that we apply a particular process for singularity or outlier detection (detailed below). The procedure is run for each chromosome separately.

AWS parameters

The procedure requires the tuning of various parameters. We apply the exponential kernel $K_l(u) = K_s(u) = \exp(-u)$. For the neighborhood, we have chosen $h^{(0)} = 1$, $a = 1.2$ and $h^* = 10X_N$. The parameter λ has been set to the 0.999-quantile of the $\chi^2(1)$ distribution, to prevent there being too many breakpoints. The value of η has been set to 0.5 and corresponds to the memory parameter of the algorithm. Polzehl and Spokoiny (2002) suggested using the symmetric statistical

penalty $s_{ij}^{(k)}$ to detect fine structures, as might occur in cancer data. Nevertheless, very fine structures, such as single amplicons or deletions, may be missed and a special procedure is proposed in the next paragraph.

Outlier detection

The AWS procedure is based on the assumption that the maximal neighborhood on which parametric estimation can be carried out is large compared with the distance between two neighboring points. This procedure may therefore fail to detect very fine structures such as a BAC located in a MSHR for which the signal Y_i differs significantly from the expected values of this MSHR. Such a BAC is called an outlier (we point out that our definition of an outlier is purely statistical, and therefore an outlier corresponds either to a biological effect—local amplicon or deletion—or to an experimental artefact). To overcome this limitation in the detection of outliers, we have designed a special procedure based on median-absolute-deviation (MAD) for detecting the remaining outliers. It should be noted that when an outlier presents a large deviation, it is detected at the breakpoint detection step. This first type is called AWS outlier and is characterized by a location X_i such that $\hat{\theta}_{i-1} \in [\hat{\theta}_{i+1} - \epsilon; \hat{\theta}_{i+1} + \epsilon]$ and $\hat{\theta}_i \notin [\hat{\theta}_{i-1} - \epsilon; \hat{\theta}_{i-1} + \epsilon]$ (N.B. a special treatment is applied for starting location and ending location: if $\hat{\theta}_1 \notin [\hat{\theta}_2 - \epsilon; \hat{\theta}_2 + \epsilon]$ (respectively $\hat{\theta}_{N-1} \notin [\hat{\theta}_N - \epsilon; \hat{\theta}_N + \epsilon]$) then X_1 (respectively X_N) is considered as well as an outlier). The second type of outlier is called MAD outlier and such outliers are identified as follows: for each MSHR, we remove all the AWS outliers; based on the assumption that the observations Y_i in an MSHR are drawn from the normal distribution $\mathcal{N}\left(\mu_k, \sigma_k^2\right)$, a location X_i for which the observation Y_i lies in the $\alpha/2$-quantile upper or lower tail of the normal distribution $\mathcal{N}\left(\hat{\mu}_k, \hat{\sigma}_k^2\right)$ is considered to be a MAD outlier (α has been set to 0.001). As we are looking for outliers, $\hat{\mu}_k$ is estimated by the median and $\hat{\sigma}_k^2$ by the square of the median-absolute-deviation for robustness considerations.

Optimization of the number of breakpoints

Our data show that despite the use of a strong statistical penalty λ, the AWS procedure may in some cases identify breakpoints which correspond to small shifts and define regions of $\sim$10–20 BACs. This is probably due to specific local effects on the chromosome, unrelated to the biological variation we want to investigate but nevertheless real. Thus, a filtering step was added to remove these undesirable breakpoints. Before this step, all the outliers are excluded from the analysis. The likelihood of our data can be written as:

$$L = \prod_{i=1}^{B_1} \frac{1}{\sigma_1\sqrt{2\pi}} \; \mathrm{e}^{-\frac{1}{2}\left(\frac{Y_i-\mu_1}{\sigma_1}\right)^2} \cdots \prod_{i=B_{N'+1}}^{N} \frac{1}{\sigma_{N'+1}\sqrt{2\pi}} \; \mathrm{e}^{-\frac{1}{2}\left(\frac{Y_i-\mu_{N'+1}}{\sigma_{N'+1}}\right)^2}.$$

We calculate the following function:

$$f = \sum_{k=1}^{N'+1} (B_k - B_{k-1}) \log\left(\hat{\sigma}_k^2\right) + \lambda' \sum_{k=1}^{N'} K(\hat{\sigma}^{-1}|\hat{\mu}_k - \hat{\mu}_{k+1}|) \log(N)$$

with $B_0 = 0$, $B_{N'+1} = N$, $\hat{\sigma}_k^2$ and $\hat{\mu}_k$ are the usual MLE of σ_k^2 and μ_k, and $\hat{\sigma}$ is calculated from Equation (1). The function f corresponds, up to an additive constant, to a penalized form of $-\log L$. The function $K(x)$ is the tricubic kernel function and takes the value $[1-(x/6)^3]^3$ for $x \in [0;6]$ and zero elsewhere. A kernel function in the penalty term is chosen mainly to prevent the removal of true breakpoints defining a MSHR of very small cardinality. The algorithm is then very similar to the JOIN procedure of the GLSo algorithm proposed by Jong *et al*. (2003): the breakpoint for which removal leads to the largest decrease in the function f is eliminated and the procedure is iterated until the function f ceases to decrease. When a breakpoint is removed, a new larger MSHR appears and its MAD outliers are re-evaluated.

REGION ASSIGNMENT

The purpose of the region assignment is to assign a gain, loss or normal status to each MSHR. Our algorithm involves three steps:

- First, for each chromosome, MSHR are grouped in classes, each class containing MSHRs of the same expected (but unknown) DNA copy number.
- Second, the resulting classes for all chromosomes are clustered to produce superclasses, of same expected DNA copy number; these superclasses are called homogeneous chromosomal status regions (HCSR).
- Finally, each HCSR is given a label: gain, normal or loss. An evaluation of the ratios is computed and corresponds to different levels of gain or loss.

This two-step clustering (chromosome, then genome level) ensures that label assignments are consistent for all clusters within a chromosome. This refinement is necessary since the signal measured on the array may be chromosome-dependent.

MSHR clustering by chromosome

The aim of this step is to cluster the MSHR identified on a chromosome such that each cluster corresponds to a set of MSHR with identical statuses. In practice, we do not know a priori the number of clusters for a given chromosome, and we therefore propose criteria for determining the most appropriate number of clusters. We do this as follows: first, we eliminate all the outliers detected previously; then, we calculate the mean and cardinality of each MSHR; finally, we perform hierarchical clustering of the means of MSHRs with centroid criteria,

taking into account the cardinality of each MSHR. From the dendrogram produced, we then try to find the optimal number of clusters for chromosomes with more than one breakpoint (if there is only one breakpoint then the chromosome has two clusters). We successively cut the dendrogram to obtain sets S_i of clusters $C_1^i, \ldots, C_i^i$ with $i = 2, 3, \ldots, N_{\max}^*$ ($N_{\max}^*$ is less than or equal to the number of MSHR). We now use all the points belonging to each cluster (except outliers) to calculate the likelihood as follows:

$$L_i = \prod_{j \in C_1^i} \frac{1}{\sigma_1 \sqrt{2\pi}} \; \mathrm{e}^{-\frac{1}{2}\left(\frac{Y_j - \mu_1}{\sigma_1}\right)^2} \ldots$$

$$\prod_{j \in C_i^i} \frac{1}{\sigma_i \sqrt{2\pi}} \; \mathrm{e}^{-\frac{1}{2}\left(\frac{Y_j - \mu_i}{\sigma_i}\right)^2}.$$

We calculate the following function:

$$f_i^* = \sum_{k=1}^{i} \#C_k^i \log\left(\hat{\sigma}_k^2\right) + \lambda^* \sum_{k=1}^{i-1} K\left(\hat{\sigma}^{-1} |\hat{\mu}_k - \hat{\mu}_{k+1}|\right) \log(N),$$

where $\hat{\sigma}_k^2$ and $\hat{\mu}_k$ are the usual MLE of σ_k^2 and μ_k, $\hat{\sigma}$ is calculated from Equation (1) and $\#C_k^i$ is the cardinality of the cluster (N.B. the clusters are sorted in increasing order of means). This function corresponds, up to an additive constant, to a penalized form of $-\log L_i$. The optimal number of clusters is $i^* = \mathrm{arginf}_i f_i^*$. The clusters identified correspond to HCSRs. The value of λ^* has been set to 8.

HCSR clustering throughout the genome

The preceding step provides us with a set of HCSRs for each chromosome. At this stage of the analysis, we now consider globally the HCSRs of the whole genome. Based on the same principle, we cluster HCSRs according to their means, using the centroid agglomeration method and taking into account the cardinality of each HCSR. We retain the number of clusters for which the new function f_i^* is minimal. In this case, the minimal value of i is 1 (and for this value of i, the f_i^* function is calculated without the penalty term). The estimate $\hat{\sigma}$ is calculated from the data for the whole genome. For our analysis, λ^* has been set to 40.

Label assignment

We now have to decide which regions are normal, and which have been lost or gained. In array CGH experiments, as in standard microarray experiments, a bias results from differences in dye incorporation efficiency such that, even for normal/normal hybridization, the expected $\log_2$-ratios are not centered around zero. Thus, $\log_2$-ratiosare median-centered before identification of the normal DNA regions. Once clustering has been achieved for the whole genome, the cluster with the median closest to zero is considered to be normal DNA. Clusters with higher medians are considered to reflect gains and those with lower values are considered to reflect losses.

VALIDATION

Validation on simulated data

We simulated 210 genomic profiles of three types: normal profiles, profiles displaying moderate rearrangement and profiles displaying high levels of rearrangement. For each profile, we generated a series of 2457 points drawn from a normal distribution with a mean of zero and an SD of 0.079, evaluated from 12 normal/normal hybridization arrays. For moderate- and high-rearrangement profiles, a status (loss, normal or gain) was defined according to a three-state first-order Markov process with a probability transition matrix:

$$\begin{pmatrix} 0.99 & 0.008 & 0.002 \\ 0.0005 & 0.999 & 0.0005 \\ 0.002 & 0.008 & 0.99 \end{pmatrix}$$

and

$$\begin{pmatrix} 0.995 & 0.004 & 0.001 \\ 0.0025 & 0.995 & 0.0025 \\ 0.001 & 0.004 & 0.995 \end{pmatrix}$$

respectively. We added realistic values of 0.3 for gain status and -0.3 for loss status to the profile generated. We also used a Poisson process to add outliers such that the expected number of outliers in the series was 20. A value of 0.3 was either added to the value or subtracted, with a probability of 0.5. The global performance of our methodology was assessed according to the following criteria: (#correctly labeled BACs + #true positive outliers)/total number of BACs.

For both values of λ', this criterion ranges from 98.94 to 99.84%. For a total of 1195 breakpoints, 81.9% were correctly located and 15.1% were incorrectly located, with a maximum localization error of 3 BACs (cf. Table 1) for $\lambda' = 8$. For $\lambda' = 10$, no improvement was observed because the decrease in false positive rate did not counteract the increase in false negative rate. A total of 278 and 283 breakpoints were removed for λ' values of 8 and 10, respectively. We found that 66.2% of the outliers were correctly identified (cf. Table 1). The large number of false negatives may be accounted for by these points being picked up in a distribution with only a small shift (± 0.3) with respect to their neighborhood.

We have estimated the resolution of our method by simulating a chromosomal profile of 200 BACs. In the middle position, an alteration of length 1, 2, 4 or 8 has been added with a signal mean amplitude of 0.15, 0.20, 0.25 or 0.30 and a gaussian distribution. The SD is 0.079, as measured on our real data. For each combination of length and signal, 1000 simulations have been done (*nb* the HCSRclustering step has been ignored since we are working on only one chromosome). The resolution is estimated both by the percentage of correctly assigned BACs in the altered region and the number of times

Table 1. The results for the detection of breakpoints and outliers on 210 simulated genomic profiles for two values of λ'

	$\lambda' = 8$	$\lambda' = 10$
Total number of breakpoints	1195	1195
Number of breakpoints correctly identified	979	978
Number of breakpoints mislocated	181	178
Number of missed breakpoints	35	39
Number of additional breakpoints	26	25
Difference in position for mislocated breakpoints		
1	167	164
2	13	13
3	1	1
Outlier detection		
True positives	2679	2678
False negatives	1364	1365
False positives	1243	1249

We obtained 98.9–99.8% correct assignations, see text for details.

Table 2. Resolution of the method estimated on a chromosomal profile of 200 BACs depending on the length of the altered region and the signal amplitude

Signal	Length of altered region			
	1	2	4	8
Percentage of correctly labelled BACs				
0.15	$10 \pm .95$	$9 \pm .66$	$9 \pm .47$	$14 \pm .85$
0.20	23 ± 1.33	$21 \pm .92$	$26 \pm .92$	56 ± 1.33
0.25	48 ± 1.58	45 ± 1.14	56 ± 1.11	$90 \pm .66$
0.30	67 ± 1.49	69 ± 1.04	$81 \pm .85$	$97 \pm .19$
Percentage of altered regions				
0.15	$10 \pm .95$	17 ± 1.17	28 ± 1.42	43 ± 1.55
0.20	23 ± 1.33	38 ± 1.52	60 ± 1.55	81 ± 1.23
0.25	48 ± 1.58	70 ± 1.45	88 ± 1.04	$98 \pm .41$
0.30	67 ± 1.49	$90 \pm .95$	$98 \pm .47$	100 ± 0

The performance (mean $\pm$ SD) are estimated by the percentage of correctly assigned BACs in the altered region and the number of times that at least an alteration has been found in this region. SD on signal ratios was estimated on real data and set to 0.079.

Table 3. The results for breakpoint detection and label assignment on 15 human cell strains (Snijders' dataset)

Cell strain/chromosome	$\lambda' = 8$	$\lambda' = 10$
GM00143/False	8	0
GM01524/6	Yes	Yes
GM01524/False	0	0
GM01535/5	Yes	Yes
GM01535/12	Yes	Yes
GM01535/False	0	0
GM01750/9	Yes	Yes
GM01750/14	Yes	Yes
GM01750/False	0	0
GM02948/False	1	0
GM03134/8	Yes	Yes
GM03134/False	4	4
GM03563/3	Yes	Yes
GM03563/9	Yes	Yes
GM03563/False	8	4
GM03576/False	0	0
GM04435/False	2	2
GM05296/10	Yes	Yes
GM05296/11	Yes	Yes
GM05296/False	8	6
GM07081/7	Yes	Yes
GM07081/15	No	No
GM07081/False	6	6
GM07408/False	2	2
GM10315/False	3	0
GM13031/17	Yes	Yes
GM13031/False	4	4
GM13330/1	Yes	Yes
GM13330/4	Yes	Yes
GM13330/False	0	0

Following the / after the cell strain name is the number of the chromosome on which a breakpoint is present or 'False', indicating the number of false-positive breakpoints identified by the procedure in each cell strain. Yes means that breakpoints have been correctly located for the chromosome under consideration. All breakpoints were detected and all label assignments are correct except for GM07081/15 (not detected by the array CGH technology) and BAC RP11-237j07 of GM05296. In this last case, the breakpoint was located on the neighboring BAC.

that at least an alteration has been found in this region. The results are presented in Table 2 and show that a signal less or equal to 0.2 give low performance unless the length of the region is greater than 8 BACs and a signal greater or equal to 0.25 give good performance. Note that the results of our simulations depends only on the signal-to-noise ratio of the data, that should be kept higher than approximately 2.5 to avoid deterioration of performances.

Validation on the dataset from Snijders *et al*. (2001)

We present here the results obtained with our methodology applied to a public dataset (Snijders *et al*., 2001). The data correspond to 15 human cell strains with known karyotypes (12 fibroblast cell strains, 2 chorionic villus cell strains and 1 lymploblast cell strain) from the NIGMS Human Genetics Cell Repository (http://locus.umdnj.edu/nigms). Each cell strain has been hybridized with an array CGH of 2276 BACs, spotted in triplicate. The variable used for the analysis is the test over reference $\log_2$-ratio, as described by the authors. This dataset had already been analyzed with another algorithm; the results obtained are presented in Olshen and Vankatraman (2002).

Our results for breakpoint detection and label assignment are shown in Table 3 for two values of λ'. Our algorithm gave perfect detection of breakpoints: none was missed in the nine cell strains that had breakpoints. For strain GM05296, the first breakpoint of chromosome 10 was detected on BAC RP11-14i14 instead of RP11-237j07, which immediately follows it: visual checking showed that the conclusion in favor of BAC RP11-237j07 was far from clear. The number of false-positive breakpoints decreases dramatically if the

value of λ' is increased from 8 to 10. However, for some cell strains, false-positive breakpoints remain (especially for GM00143 and GM03563): such false-positive breakpoints may result from local trends on the chromosome (a BAC effect or a drift along the genome can be observed, even for normal/normal hybridizations). Similar false-positive breakpoints were reported by Olshen and Vankatraman (2002) for the cell strain GM03563, on chromosome 11. All label assignments were correct, except for the monosomic region on chromosome 15 of GM07081, which was not detected by array CGH technology (Snijders *et al*., 2001). For cell strains GM04435, GM07081 and GM07408, our algorithm identified a small monosomic region (although karyotyping did not show this region to be monosomic) of two BACs on chromosome 8 (RP11-122N11 and RP11-287P18), corresponding to the region identified in strain GM03134. If we compare our results with those obtained by Olshen and Vankatraman (2002), our algorithm gave fewer false-positive breakpoints. For cell strain GM03134, our algorithm identified the small monosomic region on chromosome 8 whereas Olshen and Vankatraman (2002) did not identify this region. For cell strain GM01535, Olshen and Vankatraman (2002) did not find the monosomic region consisting of a single BAC located at the end of chromosome 12, whereas this BAC was detected as an AWS Outlier by our algorithm.

Validation on bladder cancer data

We have applied our algorithm to bladder cancer data from tumors collected at Henri Mondor Hospital (Créteil, France) (Billerey *et al*., 2001) and hybridized on arrays CGH composed of 2464 BACs (F. Radvanyi, D. Pinkel *et al*., unpublished data). The data consist of 13 arrays CGH experiments (using DNA from 13 different bladder tumors with the following stages-grades: 1 T1G2, 1 T1aG3, 1 T2G2, 2 T3G3 and 8 T4G3) hybridized according to Pinkel's protocol (Pinkel *et al*., 1998) (Table 4). Images were analyzed with SPOT 2.0 software (Jain *et al*., 2002). A pre-processing step was used to remove poor-quality spots. Spots with a reference signal intensity (and DAPI signal intensity) below 125% of the background reference signal (DAPI signal) were discarded. Triplicates with an SD of $\log_2$-ratio >0.1 were removed from the analysis and spots located in areas of spatial bias (unpublished data) were also eliminated. The value used is the mean for each BAC of the Log2Rat variable calculated by SPOT 2.0, which corresponds to the test over reference $\log_2$-ratio (as each BAC was spotted three times on the array CGH). For our data, the karyotype is unknown. Thus, we mainly focused on breakpoint detection validation on the basis of visual expertize. Nevertheless, supporting evidence for the location of breakpoints was provided by LOH analysis.

Based on visual expertize, AWS smoothing gave an excellent fit to the CGH profile (cf. Figs 1 and 2) and this algorithm seems highly appropriate for array CGH analysis. Despite the small number of errors observed, the optimization procedure for incorrect breakpoint removal is necessary to remove false positives. A total of 108 and 116 breakpoints were removed (for $\lambda' = 8$ and 10, respectively), even though some were of biological interest. For four tumors, label assignment was highly problematic, even from visual expertize. These tumors corresponded to high-stage and high-grade tumors (1 T2G2 and 3 T4G3) with many genome rearrangements. Indeed, signal variation at breakpoint may be blurred by several biological limitations of the technology: tumor biopsy samples generally contain a mixture of normal and tumor cells, and cells within a tumor may display differences in genomic losses and gains, a phenomenon known as tumor heterogeneity. Moreover, aneuploidy may affect several chromosomes differently. These limitations make breakpoint detection and label assignment difficult. For the other nine tumors, label assignment was consistent with visual expertize.

Table 4. The results for the detection of breakpoints and outliers on 13 bladder tumor genomic profiles for two values of λ'

	$\lambda' = 8$	$\lambda' = 10$
Total number of breakpoints	267	267
Number of breakpoints correctly identified	251	245
Number of breakpoints mislocated	7	7
Number of missed breakpoints	9	15
Number of additional breakpoints	9	8
Difference in position for mislocated breakpoints:		
1	6	6
2	1	1

Performances are similar to those of a human expert.

These 13 bladder tumors had been assessed for LOH on chromosome 10, using polymorphic markers (Cappellen *et al*., 1997). Although CGH and LOH studies do not provide the same information (Albertson *et al*., 2003), the results of the two studies were consistent: the regions of gains and losses detected by array CGH correspond to regions of allelic imbalance detected with polymorphic markers. For example, Cappellen *et al*. (1997) found an allelic imbalance for polymorphic markers between D10S185 and D10S168 on chromosome 10 of tumor 1533e: these markers are located between BACs RP11-9M11 (Position 1402) and RP11-32I9 (Position 1431), which delineate the lost region detected by array CGH for the same tumor (cf. Fig. 2).

A region of amplification including the *CCND1* (cyclin D1) gene was detected on the long arm of chromosome 11 for tumor 1533e (cf. Fig. 2). Interestingly, the breakpoints defining this previously identified region on chromosome 11 of tumor 1533e were also detected in the peritumoral urothelium of the patient concerned, although the mean $\log_2$-ratio of this region was only 0.25 (data not shown), demonstrating the sensitivity of our method.

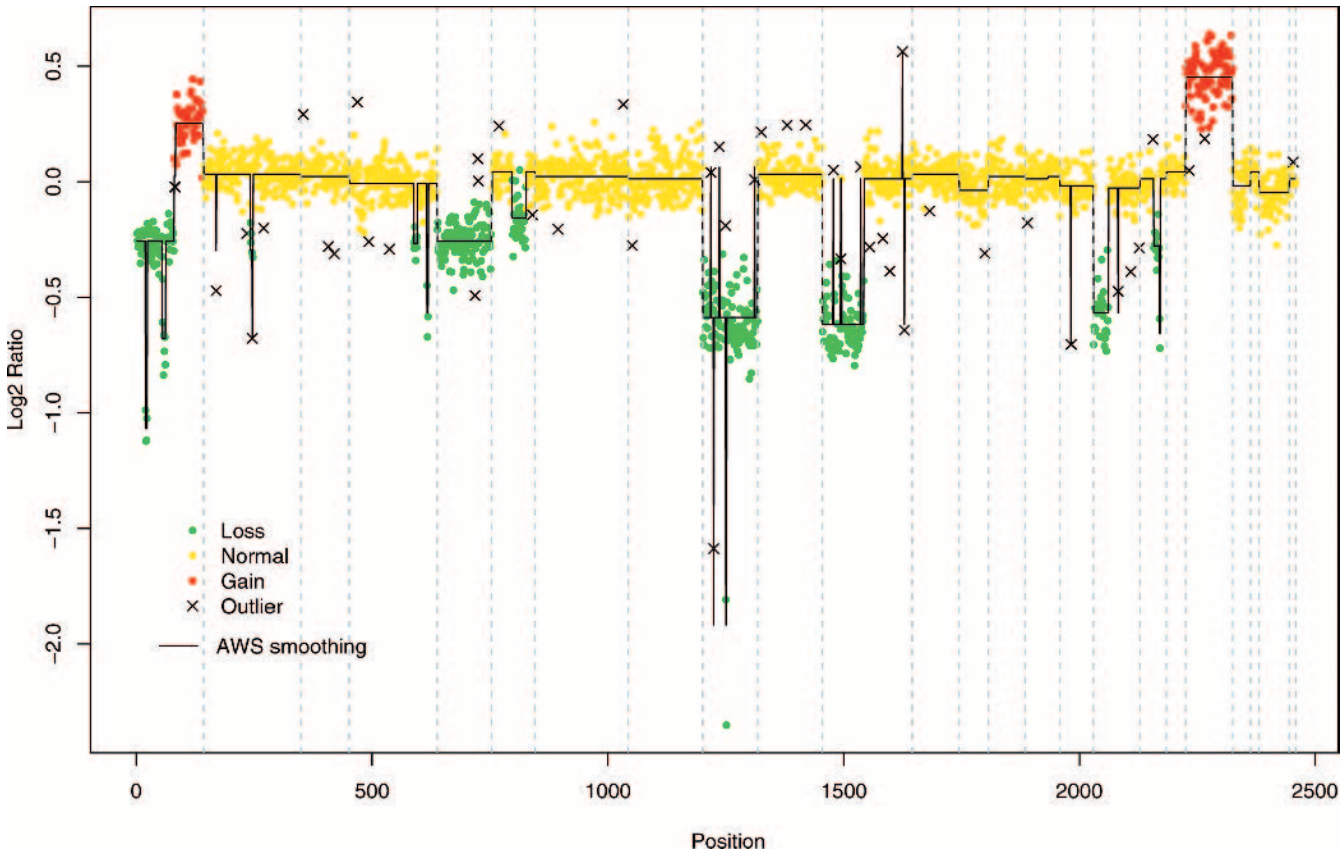

Fig. 1. Genomic profile of bladder tumor 824 (T1G2) according to our methodology: the BREAKPOINT DETECTION step makes it possible to calculate the piecewise constant function, in black, and to detect outliers; during the REGION ASSIGNMENT step, a two-step clustering process groups together regions of same status and then assigns a label (gain, normal or loss) to each region. The vertical gray dashed lines indicate the separation between chromosomes. The horizontal axis shows the rank position of each BAC along the genome and the vertical axis shows the tumor/normal $\log_2$-ratios after median centering.

DISCUSSION

We present here a new methodology for breakpoint detection and status assignment to each BAC in a array CGH experiment. Our algorithm is highly efficient with both simulated and manually analyzed data. For real data, our results are similar to those obtained by a human expert. On a public dataset, our algorithm outperformed the method described by Olshen and Vankatraman (2002). Simulated data are also correctly analyzed by our method: in most cases, missed breakpoints or outliers were not detected properly simply because the randomization procedure gave them signal ratios far from the expected ratios of their class of origin. In such cases, the available information is insufficient for the correct detection of breakpoints or outliers from the data, whatever the algorithm used.

The AWS procedure correctly detects large regions but also accurately fits data for both fine structures and small local effects. Local effects have been already reported by Olshen and Vankatraman (2002), consisting of regions of the genome showing a recurrent bias in the signal ratios confirmed on normal/normal hybridization (N/N) (data not shown): in our data, the strongest local effects were observed on chromosome 6 and chromosome 13. Both involved a shifting down of the tumor signal with respect to the normal signal. With homogeneous biological samples (e.g. cell lines), a local effect generally induces a much smaller shift than a gain or a loss of DNA. However, tumor biopsy samples are generally a mixture of normal and tumor cells and thus, heterogeneity reduces shifts, making it difficult to distinguish a biological effect from a local effect, leading to the identification of false-positive breakpoints. This suggests that an *ad hoc* procedure should be developed to eliminate such biases.

From our experience with normal/normal hybridizations, this local effect and other sources of variation, such as a BAC effect, appear to be array-dependent, rather than systematic. One solution is to flag the regions or BACs subject to such biases and to consider them with caution. More generally, this problem raises the question of array CGH data normalization and shows that breakpoint detection and label assignment are closely linked to the normalization step. Our

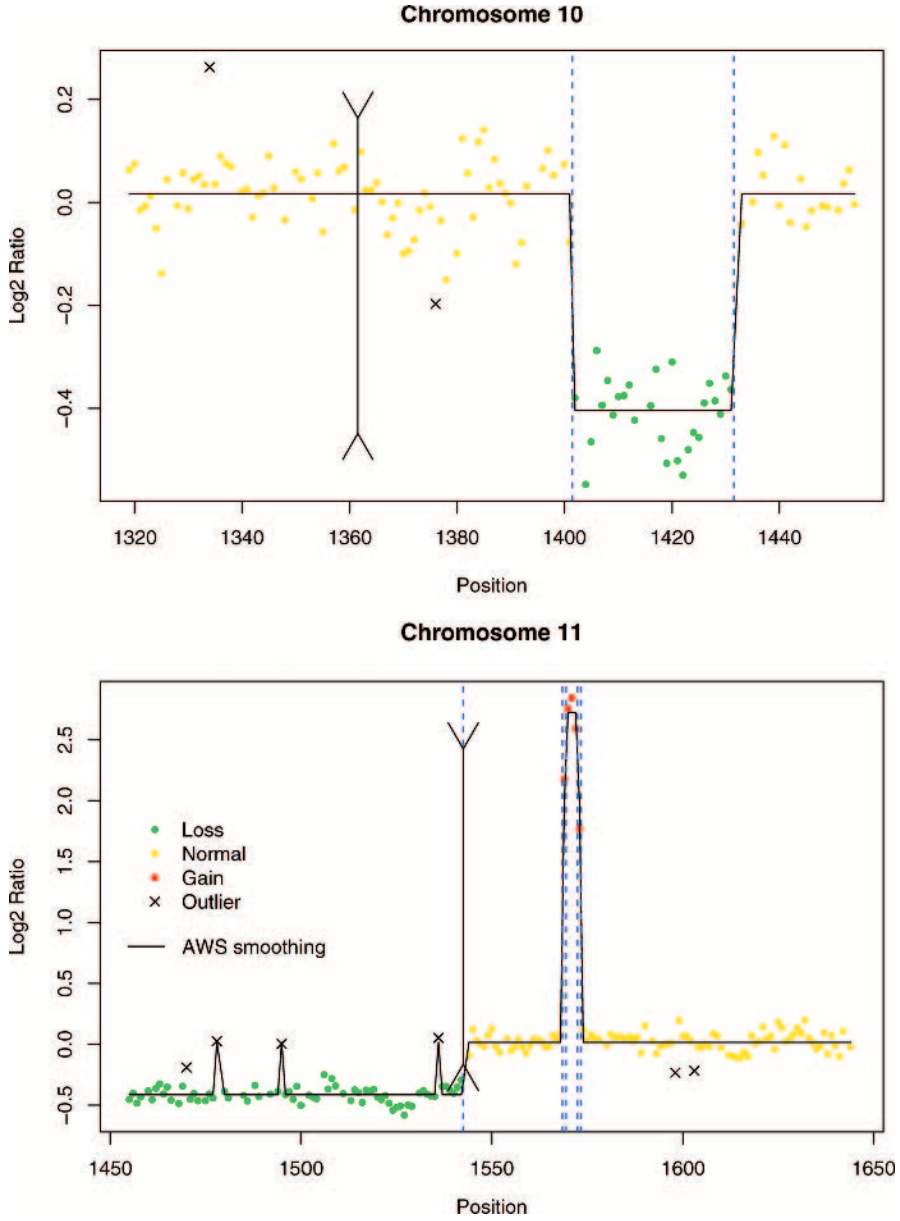

Fig. 2. Profiles for chromosomes 10 and 11 for the bladder tumor 1533e (T4G3). The vertical red dashed lines indicate the breakpoints and the vertical black double arrow indicates the centromere. The horizontal axis indicates the rank position of each BAC along the genome and the vertical axis indicates the $\log_2$-ratios after median centering.

findings also show that normalization should be carried out with an adaptive (array-dependent) algorithm. In this study, we simply applied a filter based on spot quality control criteria and removed abnormally high $\log_2$-ratios measured in some areas of the array, referred to as spatial biases (generally an edge or corner effect). Further improvements to normalization are envisaged and will form the subject of another publication. The biological significance of the outliers detected must be considered carefully for several reasons: first, natural polymorphisms may result in outliers, as shown in some cases on normal/normal hybridizations. These particular clones must therefore be flagged (such polymorphisms have been observed in our data). Second, some BACs may systematically display aberrant behavior. Finally, some BACs may have been mislocated on the genome: between two consecutive versions of the draft sequence, some BACs may be transferred from one position on a chromosome to another.

When using our algorithm, several parameters must be set: the main parameters are the statistical penalty λ for the AWS procedure, the λ' value for optimization of the number of breakpoints and λ^* in the two-step clustering step. We have set these values empirically based on our own data, but when

applying our method to arrays CGH obtained on another platform, it may be necessary to modify these parameters and a model selection step may be required (array replicates and normal/normal arrays are particularly useful at this stage).

Although breakpoint and outlier detection are entirely satisfactory with our method, label assignment is much more difficult. Several phenomena make it difficult to classify regions correctly into three classes (loss, normal and gain), not to mention to assign a number of DNA copies to a region. We have already raised the problem of sample heterogeneity. In cases of polyploidy, a single loss results in mathematically smaller shifts. In situations in which label assignment is problematic, the use of other sources of biological knowledge, such as genotyping, is likely to improve performance.

Although our methodology requires further improvement, it already provides new materials for the large-scale analysis of array CGH profiles and makes it possible to envisage further analysis. Indeed, the segmentation of CGH profiles and the assignment of statuses to BACs are required for more advanced transverse analysis in sets of patients: detection of regions recurrently lost or gained, unsupervised and supervised classification based on the CGH profile, integration of the genome and transcriptome profiles for the identification of new genes involved in tumorigenesis and/or tumor progression. This work should lead to new insight valuable for clinical research and cancer treatment. Our work was driven by and applied to cancer array CGH analysis but can also be applied to any genetic disease involving deletion or amplification in genomic DNA.

ACKNOWLEDGEMENTS

This work was supported by the Centre National de la Recherche Scientifique, the Institut Curie, the Comité de Paris Ligue Nationale contre le Cancer (Laboratoire Associé) and the IST program from the European Commission through the HKIS project (IST-2001-38153). Data processing was managed by the ™ Amadea software from ISoft (Gif sur Yvette, France).

REFERENCES

Albertson,D.G., Collins,C., McCormick,F. and Gray,J.W. (2003) Chromosome aberrations in solid tumors. *Nat. Genet.*, **34**, 369–376.

Albertson,D.G. and Pinkel,D. (2003) Genomic microarrays in human genetic disease and cancer. *Hum. Mol. Genet.*, **12**, R145–R152.

Axon Instruments (2003) *GenePix Pro 5.0 User's Guide*. ©Axon Instruments, Inc.

Billerey,C., Chopin,D., Aubriot-Lorton,M.H., Ricol,D., de Medina,S.G.D., Rhijn,B.V., Bralet,M.P., Lefrere-Belda,M.A., Lahaye,J.B., Abbou,C.C., *et al*. (2001) Frequent FGFR3 mutations in papillary non-invasive bladder (pTa) tumors. *Am. J. Pathol.*, **158**, 955–1959.

Cappellen,D., Gil Diez de Medina,S., Chopin,D., Thiery,J.P. and Radvanyi,F. (1997) Frequent loss of heterozygosity on chromosome 10q in muscle-invasive transitional cell carcinomas of the bladder. *Oncogene*, **14**, 3059–3066.

Ishkanian,A.S., Malloff,C.A., Watson,S.K., DeLeeuw,R.J., Chi,B., Coe,B.P., Snijders,A., Albertson,D.G., Pinkel,D., Marra,M.A. (2004) A tiling resolution DNA microarray with complete coverage of the human genome. *Nat. Genet.*, **36**, 299–303.

Jain,A.N., Tokuyasu,T.A., Snijders,A.M., Segraves,R., Albertson,D.G., and Pinkel,D. (2002) Fully automatic quantification of microarray image data. *Genome Res.*, **12**, 325–332.

Jong,K., Marchiori,E., van der Vaart,A., Ylstra,B., Weiss,M. and Meijer,G. (2003). Chromosomal breakpoint detection in human cancer. In Raidl,G.R., Cagnoni,S., Cardalda,J.J.R., Corne,D.W., Gottlieb,J., Guillot,A., Hart,E., Johnson,C.G., Marchiori,E., Meyer,J.-A. and Middendorf,M. (eds), *Applications of Evolutionary Computing, EvoWorkshops2003: EvoBIO, EvoCOP, EvoIASP, EvoMUSART, EvoROB, EvoSTIM*, Volume 2611 of *LNCS*, University of Essex, England, UK. Springer-Verlag, Berlin, pp. 54–65.

Olshen,A.B. and Vankatraman,E.S. (2002) Change-point analysis or array-based comparative genomic hybridization data. *Proceedings of the Joint Statistical Meetings*, New York, August 11–15, 2530–2535.

Pinkel,D., Segraves,R., Sudar,D., Clark,S., Poole,I., Kowbel,D., Collins,C., Kuo,W.L., Chen,C., Zhai,Y. (1998) High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat. Genet.*, **20**, 207–211.

Polzehl,J. and Spokoiny,S. (2000) Adaptive weights smoothing with applications to image restoration. *J. R. Stat. Soc., Ser. B*, **62**(2), 335–354.

Polzehl,J. and Spokoiny,S. (2002) Local likelihood modelling by adaptive weights smoothing. WIAS-Preprint 787.

Sen,A. and Srivastava,M.S. (1975) On tests for detecting a change in mean. *Ann. Stat.*, **3**, 98–108.

Shaw-Smith,C., Redon,R., Rickman,L., Rio,M., Willatt,L., Fiegler,H., Firth,H., Sanlaville,D., Winter,R., Colleaux,L., Bobrow,M. and Carter,N.P. (2004) Microarray based comparative genomic hybridisation (array-CGH) detects submicroscopic chromosomal deletions and duplications in patients with learning disability/mental retardation and dysmorphic features. *J. Med. Genet.*, **41**, 241–248.

Snijders,A.M., Nowak,N., Segraves,R., Blackwood,S., Brown,N., Conroy,J., Hamilton,G., Hindle,A.K., Huey,B., Kimura,K. (2001) Assembly of microarrays for genome-wide measurement of DNA copy number. *Nat. Genet.*, **29**, 263–264.

Solinas-Toldo,S., Lampel,S., Stilgenbauer,S., Nickolenko,J., Benner,A., Dohner,H., Cremer,T., and Lichter,P. (1997) Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. *Genes Chromosomes Cancer*, **20**, 399–407.

Vogelstein,B., Fearon,E.R., Kern,S.E., Hamilton,S.R., Preisinger,A.C., Nakamura,Y. and White,R. (1989) Allelotype of colorectal carcinomas. *Science*, **244**, 207–211.

Yang,Y., Dudoit,S. and Speed,T. (2001) Normalization for cDNA microarray data. *SPIE BiOS 2001*, San Jose, CA, January 2001.