A statistical approach for CGH microarray data analysis

Franck Picard, Stéphane Robin, Marc Lavielle, Christian Vaisse, Gilles Celeux, Jean-Jacques Daudin

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

A statistical approach for CGH microarray data analysis

Franck Picard , Stéphane Robin , Marc Lavielle , Christian Vaisse , Gilles Celeux ,

Jean-Jacques Daudin

N° 5139

Mars 2004

THÈME 3

[1]**Institut National Agronomique Paris-Grignon**, Département OMIP, Paris, France.
[2]**Université Paris Sud**, Laboratoire de mathématique, équipe Probabilités, Statistiques et modélisation, Paris, France
[3]**University of California San Francisco**, Diabetes Center, San Francisco, USA
[4]**Institut National de Recherche en Informatiuque et en Automatique**

ISSN 0249-6399 ISRN INRIA/RR--5139--FR+ENG

INRIA FUTURS

A statistical approach for CGH microarray data analysis

Franck Picard [*], Stéphane Robin [*], Marc Lavielle [†], Christian Vaisse [‡], Gilles Celeux [§], Jean-Jacques Daudin [*]

Thème 3 — Simulation et optimisation
de systèmes complexes

Projet SELECT

Rapport de recherche n° 5139 — Mars 2004 — 13 pages

Abstract: Microarray-CGH experiments aim at detecting and mapping chromosomal imbalances, by hybridizing targets of genomic DNA from a test and a reference sample. Probes are constituted by sequences of genomic DNA (BACs) that are mapped on the genome. For this reason, the signal has a spatial coherence that has to be handled by specific statistical tools. Process segmentation seems to be a natural framework for this purpose. A CGH profile can be viewed as a succession of segments that represent homogeneous regions in the genome whose BACs share the same relative copy number in average. We model a CGH profile by a random gaussian process whose distribution parameters are affected by abrupt changes at unknown coordinates. Two major problem arise: the estimation of the break-points coordinates and the estimation of the number of segments. A dynamic programming algorithm is used to partition the data into a finite number of segments. A model selection approach is used to determine the number of segments in the profile, using an adaptative method. We explain why classical penalized criteria can not be used in the context of break-points detection and show the potentialities of our methodology, using publicly available data sets. We detect previously mapped chromosomal aberrations and disuss the performance of our methodology on noisier data concerning breast cancer cell lines.

Key-words: array CGH, process segmentation, dynamic programming, model selection

Unité de recherche INRIA Futurs
Domaine de Voluceau, Rocquencourt, BP 105, 78153 Le Chesnay Cedex (France)
Téléphone : +33 1 39 63 55 11 — Télécopie : +33 1 39 63 53 30

Une approche statistique pour l'analyse des données de microarrays CGH

Résumé : L'objectif des expériences de microarrays CGH est de détecter et de cartographier les défauts chromosomiques, en hybridant des cibles d'ADN génomiques entre un génome test et un génome de référence. Les sondes sur la puce sont des séquences d'ADN génomique (BACs) qui sont ancrées sur le génome. Un profil CGH peut être interprété comme une succession de segments qui représentent des régions homogènes sur le génome dont les BACs présentent le même nombre de copies moyen. Nous modelisons un profil CGH par un processus aléatoire de distribution gaussienne dont les paramètres sont affectés par des changements à des coordonnées inconnues. Deux problèmes se posent: l'estimation des coodonnées de rupture, et l'estimation du nombre de segments. Un algorithme de programmation dynamique est utilisé pour partitionner les données en un nombre fini de segments. Une approche par sélection de modèle permet de déterminer le nombre de segments dans le profil, à partir d'une méthode adapatative. Nous expliquons pourquoi les critères pénalisés classiques ne peuvent pas être utilisés dans le contexte de la détection de ruptures, et nous montrons les potentialités de notre méthodologie, à partir de données disponibles sur le web. Nous détectons des aberrations chromosomiques déjà référencées, et discutons les performances de notre méthodologie sur des données de cancer du sein.

Mots-clés : array CGH, détection de ruptures, programmation dynamique, sélection de modèle

Introduction

Chromosomal aberrations often occur in solid tumors : tumor suppressor genes may be inactivated by physical deletion, and oncogenes may be activated via duplication in the genome. Gene dosage effect has become particularly important in the understanding of human solid tumor genesis and progression, and has also been associated with other diseases, such as mental retardation (Albertson *et al.* (2003), Albertson and Pinkel (2003)). Chromosomal aberrations can be studied using many different techniques, such as Comparative Genomic Hybridization (CGH), Fluorescence in Situ Hybridization (FISH), and Representational Difference Analysis (RDA). Although chromosome CGH has become a standard method for cytogenetic studies, technical limitations restrict its usefulness as a comprehensive screening tool (Beheshti *et al.* (2002)). Recently, the resolution of Comparative Genomic Hybridizations has been greatly improved using microarray technology (Solinas-Toldo *et al.* (1997), Pinkel *et al.* (1998)).

The purpose of array-based Comparative Genomic Hybridization (array CGH) is to detect and map chromosomal aberrations, on a genomic scale, in a single experiment. Since chromosomal copy numbers can not be measured directly, two samples of genomic DNA (referred as the reference and the test DNA) are differentially labelled with fluorescent dyes and competitively hybridized to known mapped sequences (referred as BACs) that are immobilized on a slide. Subsequently, the ratio of the intensities of the two fluorochromes is computed and its changes indicate either gain or loss of sequences in the reference DNA compared with the test DNA. The reference DNA typically contains 2 copies for each sequence in a diploid organism, whereas chromosomal copy number may be equal to 0 or 1 for double and single deletions, or 3 and higher for single and higher level of amplification. The basic principle of array-based CGH is that the ratio of the fluorescence intensities computed for each BAC is proportional to the ratio of the BAC copy number in the test sample compared with the reference sample (Pinkel *et al.* (1998)).

Microarrays described in Snijders *et al.* (2001) consist of 2460 human BACs and P1 clones, representing approximatively 7,500 spots on the arrays. Each single BAC is mapped on the genome, and put together, they offer a coverage of the 22 autosomal chromosomes and of the 2 sex chromosomes, with an average of one clone every 1.4Mb. A CGH profile is constituted for each chromosome when fluorescence ratios (or their log_2) are ranked and plotted according to the physical position of their corresponding BACs on the genome. Each profile can be viewed as a succession of "segments" that represent homogeneous regions in the genome whose BACs share the same relative copy number in average. Array CGH data are normalized with median set to $log_2(ratio) = 0$ for regions of no change, segments with positive means represent duplicated regions in the test sample genome, and segments with negative means represent deleted regions.

Microarray technology is well known and widely used to study gene expression profiles. Since array-CGH technology is very similar, the same technical and experimental biases may arise. Technical variability could be handled with known statistical tools, with adapation to the specific case of CGH microarrays. In this regard, CGH microarray data normalization will have to be carefully considered. Nevertheless, our assumption will be that our data do not need to be normalized, in order to focus on the very specificity of array-CGH : the spatial coherence of the data on the genome.

The purpose of a statistical analysis is to quantify chromosomal aberrations, and to localize them on the genome. Process segmentation seems to be a natural framework for this purpose, and has been used in other studies (Autio *et al.* (2003), Jong *et al.* (2003), Olshen and Venkatraman (2002)). Two major issues arise in break-points detection studies, a computational problem: the localization of the segments on the genome, and a statistical problem: the estimation of the number of segments.

We propose to estimate the segments coordinates using dynamic programming. This approach has two main advantages: it provides a global optimum compared with the recursive approach proposed by Olshen and Venkatraman (2002), in a reasonable computational time. Autio *et al.* (2003) also used dynamic programming, but did not specifically assess the problem of the estimation of the number of segments, whereas it requires careful statistical consideration. We use a penalized version of the likelihood for this purpose. Classical criterions are the Akaike Information Criterion (AIC), and the Bayes Information Criterion (BIC). They use a constant coefficient of penalty and often lead to an overestimation of the number of segments. For this reason, Jong *et al.* (2003) propose to define an arbitrary penalty constant, in order to select a lower number of segments in the profile. We explain why classical penalized criteria can not be used in the context of break-points detection, and we propose to use the procedure developped by Lavielle (2003), that chooses the penalty coefficient adaptatively to the data. We explain the construction of such penalty, and theoretical results can be found in Lavielle (2003).

The article is organized as follows: in a first section, we model array CGH data using a segmented gaussian process. The parameters estimation is presented in section 2, section 3 presents the procedure to estimate the segments coordinates, using dynamic programming, and section 4 the procedure to estimate the number of segments. An application of our methodology is presented in section 5, using publicly available data sets.

1 Model and likelihood

Let us consider a CGH profile, and note $Y(x)$, the $log_2 ratio$ of the intensities for the BAC at position x on the genome: BACs are the basic units in our model. We suppose that the $log_2 ratio$ are the realizations of independent random variables $\{Y(x)\}_{x=1\dots n}$, with gaussian distributions $\mathcal{N}(\mu_x, \sigma_x^2)$. We assume that $K-1$ changes affect the parameters of the distribution of the Ys, at unknown coordinates $(t_0, t_1, t_2, ..., t_{K-1}, t_K)$ with convention $t_0 = 1$ and $t_K = n$. Our assumption is that the parameters of the Ys distributions are constant between two changes:

$$\forall x \in]t_{k-1}, t_k], \qquad Y(x) = \mu_k + \varepsilon(x) \qquad \text{with} \quad \varepsilon(x) \sim \mathcal{N}(0, \sigma_k^2),$$

where μ_k and σ_k^2 are the mean and variance of the k^{th} segment. Since BACs are supposed to be independent, the log-likelihood can be decomposed into a sum of "local" likelihoods, calculated on each segments:

$$\mathcal{L}_K = \sum_{k=1}^{K} \ell_k(y(x)_{t_{k-1}<x\leq t_k}; \mu_k, \sigma_k^2), \tag{1}$$

$$\text{where} \quad \ell_k(y(x)_{t_{k-1}<x\leq t_k}; \mu_k, \sigma_k^2) = -\frac{1}{2} \sum_{x=t_{k-1}+1}^{t_k} \left\{ \log(2\pi \times \sigma_k^2) + \left[\frac{y(x)-\mu_k}{\sigma_k}\right]^2 \right\}.$$

2 Estimation of the segments mean and variance

Given the number of segments K and the segments coordinates $(t_0, t_1, t_2, ..., t_{K-1}, t_K)$, the problem is to estimate the mean and variance for each segment. The maximum likelihood estimators are:

$$\hat{\mu}_k = \frac{1}{t_k - t_{k-1}} \sum_{x=t_{k-1}+1}^{t_k} y(x), \quad \text{and} \quad \hat{\sigma}_k^2 = \frac{1}{t_k - t_{k-1}} \sum_{x=t_{k-1}+1}^{t_k} [y(x) - \hat{\mu}_k]^2 .$$

Notice that when the segments coordinates are known, the estimation of the mean and variance for each segment is straightforward. Then, the key problem is to estimate K and $(t_0, t_1, t_2, ..., t_{K-1}, t_K)$. We will proceed in two steps: in the first step, we will consider that the number of segments is known, and the problem will be to estimate the t_ks, that is to find the best partition of a set of n individuals into K segments. In the second step, we will estimate the number of segments in the profile, using a penalized version of the likelihood.

3 A segmentation algorithm when the number of segments is known

When the number of segments K is known, the problem is to find the best partition of $\{1, ..., n\}$ according to the likelihood, where n is the size of the sample. An exhaustive search is impossible since the number of partitions of a set with n elements into K segments is $\mathcal{C}_{n-1}^{K-1}$. To reduce the computational load, we use a dynamic programming approach (programs are coded in MATLAB language and are available upon request). Let $\hat{\mathcal{L}}_{k+1}(i,j)$ be the maximum log-likelihood obtained by the best partition of the data $\{Y(i), Y(i+1), ..., Y(j)\}$ into $k+1$ segments, with k change-points, and let note $\hat{J}_{k+1}(i,j) = -2\hat{\mathcal{L}}_{k+1}(i,j)$. The algorithm is as follows:

$$\begin{aligned} k = 0, \quad \forall 0 \leq i < j \leq n \quad \hat{J}_1(i,j) &= \sum_{x=i+1}^{j} \left\{ \log(2\pi \times \hat{\sigma}_1^2) + \left[\frac{y(x) - \hat{\mu}_1}{\hat{\sigma}_1} \right]^2 \right\} \\ \forall k \in [1, K_{max}] \quad \hat{J}_{k+1}(1,j) &= \min_h \left\{ \hat{J}_k(1,h) + \hat{J}_1(h+1,j) \right\} \end{aligned}$$

Dynamic programming takes advantage of the additivity of the log-likelihood described in (1), considering that a partition of the data into $k+1$ segments is a union of a partition into k segments and a set containing 1 segment. This approach presents two main advantages: it provides an exact solution for the global optimum of the likelihood (Auger and Lawrence 1989), and reduces the computational load from $\mathcal{O}(n^K)$ to $\mathcal{O}(n^2)$ for a given K (the algorithm only requires the storage of an upper $n \times n$ triangular matrix). At the end of the procedure, the quantities $\hat{J}_1(1,n), ..., \hat{J}_{K_{max}}(1,n)$ are stored and will be used in the next step.

Notice that this problem of partitioning is analogous to the search for the shortest path to travel from one point to another, where $\hat{J}_{k+1}(1,n)$ represents the total length of a $(k+1)$-step-path connecting the point with coordinate 1 to the point with coordinate n.

4 Estimating the number of segments via penalized likelihood

In real situations, the number of segments K is unknown and should be estimated. We adopt a model selection approach, using a penalized version of the likelihood.

Let us recall that $\hat{\mathcal{L}}_K$ is the log-likelihood for the best model with K segments. This quantity can be viewed as a quality measurement of the fit of the best model with K segments to the data, that increases with the number of segments. Nevertheless , the number of parameters to estimate is proportional to the number of segments. Therefore, a too large number of segments implies a large estimation error and a risk of overfitting. A penalized likelihood is used as a trade-off between a good adjustement and a reasonnable number of parameters to estimate. It is noted

$$\tilde{\mathcal{L}}_K = \hat{\mathcal{L}}_K - \beta \times pen(K),$$

where $pen(K)$ is a penalty function that increases with the number of segments, and β is a coefficient of penalization. The estimated number of segments is such as:

$$\hat{K} = \underset{K}{Argmax} \left(\tilde{\mathcal{L}}_K \right).$$

4.1 Choice of the penalty function and coefficient

Classical penalized likelihoods use the number of independent continuous parameters to be estimated as a penalty function. Even though those criteria are widely used in the context of model selection, Lebarbier (2003) points out that they are not appropriated in the context of an exhaustive search for abrupt changes.

Let us focus on the penalty function in a first step. For classical information criteria, such as the Akaike Information Criterion and the Bayes Information Criterion, the penalty function equals to $2K$ (table 1) for an heteroscedastic model with K segments. These criteria consider that K means and K variances have to be estimated in this model. Jong *et al.* (2003) also use a penalized criterion to estimate the number of segments. They implicitly consider that the break-points coordinates are also continuous parameters, leading to a new penalty function $pen(K) = 3K - 1$, with $K - 1$ additional parameters to estimate, which correspond to the $K - 1$ break-points coordinates. Nevertheless, the characteristic of break-points detection models lies in the mixture of continuous parameters (the mean and variance of each segment) and discrete parameters (the break-points coordinates). Lebarbier (2003) specifies that the discrete parameters can not be counted as continuous parameters, since the number of possible configurations for K segments is finite and equals $\mathcal{C}_{n-1}^{K-1}$. This leads to the definition of a new penalty function adapted to the special context of the exhaustive search of abrupt changes. This function (table 1) is proportional to the number of continuous parameters, but is also proportional to a new term in $\log(\frac{n}{K})$ that takes the complexity of the visited configurations into account. It is written $pen(K) = 2K(c_1 + c_2 log(\frac{n}{K}))$, where c_1 and c_2 are constant coefficients that have to be calibrated using numerical simulations. Since AIC and BIC do not consider the complexity of the visited models, they select a too high number of segments.

criterion	β	$pen(K)$
AIC	1	$2K$
BIC	$\log(n)$	$2K$
Jong *et al.* (2003)	10	$3K-1$
Lebarbier (2003)	adaptative	$2K(c_1 + c_2 \log(\frac{n}{K}))$
Lavielle (2003)	adaptative	$2K$

Table 1: Constant and penalty function for different penalized criteria, in an heteroscedastic model with K segments.

The second term of the penalty is the penalty coefficient. This term is constant in the case of AIC and BIC ($\beta = 1$, $\beta = \log(n)$ respectively), and contributes to the oversegmentation as mentioned before. For this reason, Jong *et al.* (2003) arbitrarily choose a penalty coefficient $\beta = 10$ for their procedure to select a reasonnable number of segments (without any statistical criterion). Instead of an arbitrary choice for this coefficient, Lebarbier (2003) and Lavielle (2003) propose to choose it adaptatively to the data. Furthermore, Lavielle (2003) points out that when the number of segments is small with respect to the number of data points (which is the case in CGH data analysis), the log-term can be considered as a constant. The author rather suggests to use the penalty function $pen(K) = 2K$ and to define an automatic procedure to choose the coefficient of penalization β adaptively. We will present briefly the principle of this procedure. Theoretical developpements can be found in Lavielle (2003).

4.2 An adaptative method to estimate the penalty coefficient

If we consider that the likelihood $\hat{\mathcal{L}}_K$ measures the adjustment of a model with K segments to the data, we aim at selecting the dimension for which $\hat{\mathcal{L}}_K$ ceases to increase significatively. For this purpose, let define a decreasing sequence (β) such as $\beta_0 = \infty$ and

$$\forall i \geq 1 \qquad \beta_i = \frac{\hat{\mathcal{L}}_{K_i} - \hat{\mathcal{L}}_{K_{i+1}}}{pen(K_{i+1}) - pen(K_i)}.$$

If we represent the curve $(pen(K), \hat{\mathcal{L}}_K)$, the sequence of β_i represents the slopes between points $(pen(K_{i+1}), \hat{\mathcal{L}}_{K_{i+1}})$ and $(pen(K_i), \hat{\mathcal{L}}_{K_i})$, where the subset $\{(pen(K_i), \hat{\mathcal{L}}_{K_i}), i \geq 1\})$ is the convex hull of the set $\{(pen(K), \hat{\mathcal{L}}_K)\}$, as shown in figure 1.

Since we aim at selecting the dimension for which $\hat{\mathcal{L}}_K$ ceases to increase significantly, we look for breaks in the slope of the curve. We define l_i, the variation of the slope, that exactly corresponds to the length of the interval $]\beta_i, \beta_{i-1}]$ and we select the highest number of segments K_i such that $l_i >> l_j$ for $i < j$. Other procedures have been developped to automatically locate the break in the slope of the likelihood. Nevertheless, the criterion we use can be interpreted geometrically and is easy to implement.

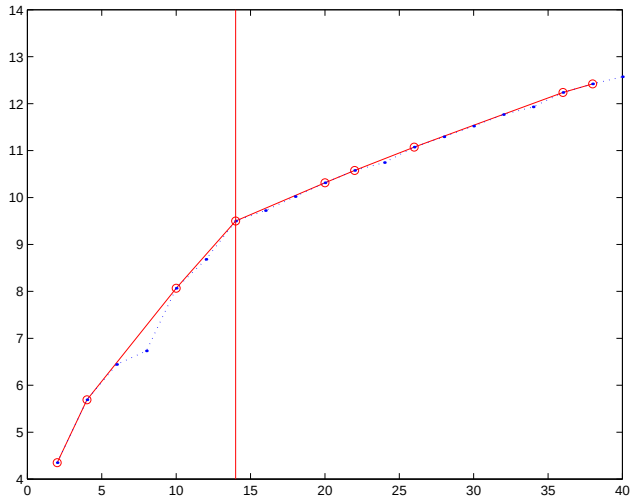


Figure 1: Illustration of the model selection procedure. The log-likelihood is plotted vs the penalty function $pen(K) = 2K$ for an heteroscedastic model. Red circles represent the convex hull of the set $\{(pen(K), \hat{\mathcal{L}}_K)\}$. The selected number of segments is $K = 7$ in this example, that corresponds to the major break in the slope of the log-likelihood.

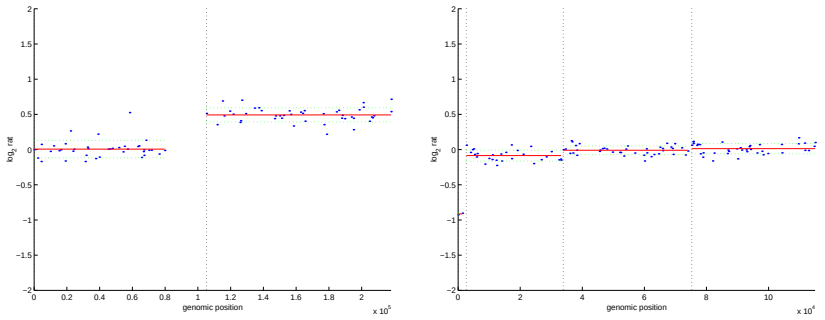
5 Results

Our procedure is applied to real data sets, described in (Snijders *et al.* 2001). We first test the ability of our procedure to detect known chromosomal aberrations on pure diploid cell lines. We take advantage of the simplicity of these data to demonstrate the importance of the model selection criterion on the estimated number of segments. Then we discuss the performance of our method on noisier data where chromosomal aberrations can not be identified manually.

5.1 Coriell Cell lines

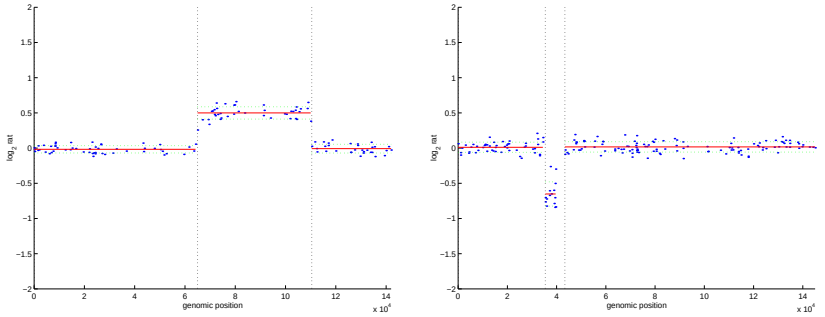
The data consist of a single experiment on fibroblast cell lines whose chromosomal aberrations have been previously mapped. Those defaults concern partial or whole chromosome aneuploidy. Figure 2 shows the graphical result of our procedure. This data have been previously used by other authors (Olshen and Venkatraman (2002)), and we also perfectly identify known cytogenetically mapped aberrations.

We would like to take advantage of the simplicity of the data to show how critical is the choice of the model selection criterion. Table 2 show the number of segments estimated with different penalized criteria. AIC and BIC lead to an overestimation of the number of segments, and figure 3 shows the segmented profile when AIC is used to estimate the number of segments. This example show the practical consequence of the use of theoretically unappropriated criteria. Let us notice that the procedures proposed by Lebarbier (2003) and Jong *et al.* (2003) give approximatively the same number of segments compared to the method we use (discussed below).



GM03563 chromosome 3
K=2

GM03563 chromosome 9
K=4



GM05296 chromosome 10
K=3

GM05296 chromosome 11
K=3

Figure 2: Result of the segmentation procedure for Coriell cell lines. Red lines represent the estimated mean of each segment, green lines the estimated mean plus or minus one standard deviation. Dots represent the $log_2 ratio$ of the intensities for each BAC plotted vs their genomic position ($10^5 kb$).

	GM03563 cell line		GM05296 cell line	
	chromosome 3	chromosome 9	chromosome 10	chromosome 11
AIC	37	48	57	77
BIC	31	29	38	47
Jong *et al.* (2003)	2	4	4	6
Lebarbier (2003)	5	7	4	5
Lavielle (2003)	2	4	3	3

Table 2: Estimated number of segments with different penalized criteria for Coriell cell lines.

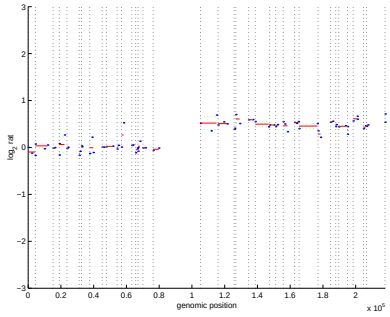
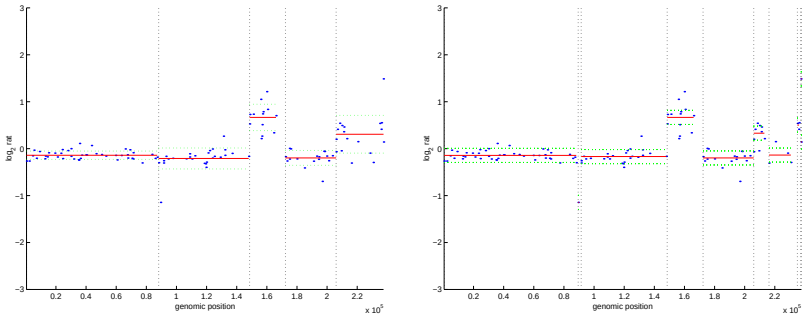


Figure 3: Result of the segmentation procedure for Coriell cell lines : GM03563 chromosome 3, using AIC to estimate the number of segments.



heteroscedastic model homoscedastic model

Figure 4: Result of the segmentation procedure for breast cancer cell lines Bt474, chromosome 1. The estimated number of segments is K=5 for model $\mathcal{M}_1$ and 10 for model $\mathcal{M}_2$, using the criterion proposed by Lavielle. BAC with position 0.9 $10^5 kb$ is identified as an outlier by model $\mathcal{M}_2$.

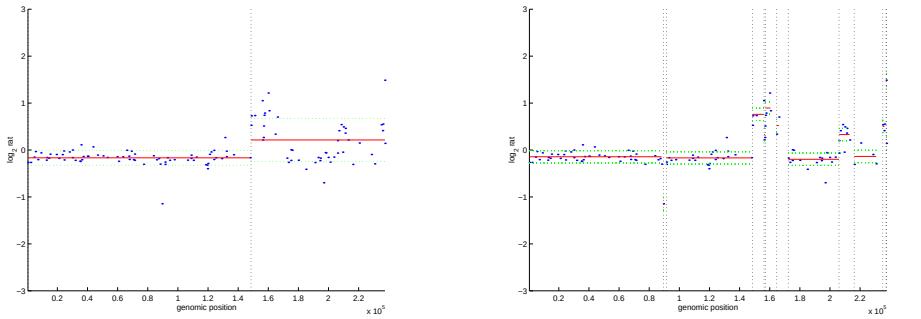
5.2 Breast Cancer Cell lines

A test genome from breast cancer cell line Bt474 is compared to a normal reference male genome. DNA are differentially labelled and hybridized to microarrays described in in (Snijders *et al.* 2001). Figure 4 represents the CGH profiles for chromosome 1. Data represent the average over three independent replicates. We consider two models :

$$\begin{array}{lllllll} \mathcal{M}_1 : & \forall x \in]t_{k-1}, t_k], & Y(x) & = & \mu_k + \varepsilon(x) & \text{with} & \varepsilon(x) \sim \mathcal{N}(0, \sigma_k^2), \\ \mathcal{M}_2 : & \forall x \in]t_{k-1}, t_k], & Y(x) & = & \mu_k + \varepsilon(x) & \text{with} & \varepsilon(x) \sim \mathcal{N}(0, \sigma^2). \end{array}$$

In model $\mathcal{M}_1$, each segment has a specific mean and variance (comparable to the model of Jong *et al.* (2003)), but in model $\mathcal{M}_2$, the variance is common between segments (this model is used in the procedure proposed by Autio *et al.* (2003)). We aim at comparing the performances of our procedure with those two models.

In figure 4, the selected number of segments is $K = 5$ for model $\mathcal{M}_1$ and $K = 10$ for model $\mathcal{M}_2$ using the estimation method proposed by Lavielle (2003) . Interestingly, the major break-points coordinates seem to be conserved between those



heteroscedastic model/Lebarbier's criterion homoscedastic model/Jong's criterion

Figure 5: Result of the segmentation procedure for breast cancer cell lines Bt474, chromosome 1. Left: number of segments estimated via Lebarbier's method in an heteroscedastic model. Right: number of segments estimated via Jong's method in an homoscedastic model.

	Bt474 cell line chromosome 1	
	model $\mathcal{M}_1$	model $\mathcal{M}_2$
Jong *et al.* (2003)	5	13
Lebarbier (2003)	2	4
Lavielle (2003)	5	10

Table 3: Estimated number of segments with different penalized criteria for Coriell cell lines.

two models, with additional segments in the case of model $\mathcal{M}_2$. This behaviour of model $\mathcal{M}_2$ could be interpreted as a trend to divide large segments into smaller parts, in order to maintain the variance σ^2 homogeneous between segments. This leads to a more segmented profile, maybe more precise, but that may be more difficult to interpret in terms of relative copy numbers. Nevertheless, as model $\mathcal{M}_2$ allows the exploration of segments with one observation, it will be more efficient for the identification of outliers, as pointed out in figure 4. Let us notice that the criteria proposed by Lebarbier (2003) and by Jong *et al.* (2003) accentuate these trends of the models (table 3 and figure 5).

Outliers seem to be a major concern in microarray CGH data analysis. For instance, if only one BAC is altered whereas its neighbors are not, the conclusion could be either that it is biologically relevant, or that the signal is due to technical artefacts. Replications are crucial in this situation, as well as secondary validations. An other possibility could be that the BAC is misannotated: if the ratio is plotted at the wrong coordinate on the genome, it will appear as an outlier, whereas it is not. To that extend, clones positions need to be updated, using newly available data from public databases.

Conclusion

Microarrays CGH currently constitute the most powerful method to detect gain or loss of genetic material on a genomic scale. To date, applications have been mainly

restricted to cancer research, but the emerging potentialities of this technique have been applied to the study of congenital and acquired diseases. As expression profile experiments require careful statistical analysis before any biological expertise, CGH microarray experiments will require specific statistical tools to handle experimental variability, and to consider the specificity of the the studied biological phenomena. We introduced a statistical method for the analysis of CGH microarray data, that models the abrupt changes in the relative copy number ratio between a test DNA and a reference DNA. Interestingly, the segmentation algorithm is not the main issue in our context, since dynamic programming provides an optimal solution in a reasonnable computational time. The power of this approach has been illustrated in many other contexts, and process segmentation is only one example among all the possibilities of dynamic programming, especially in bioinformatics (Giegerich (2000)). The main theoretical issue is rather in the estimation of the number of segments that requires the definition of appropriate penality function and coefficient. We use a new procedure proposed in Lavielle (2003), for simplicity of implementation and interpretation, but other methods have been proposed see Lebarbier (2003) for instance. Assessing the number of segments in a model is theoretically complex, and requires the definition of a precise model of inference. To that extend, microarray CGH analysis not only requires computational approaches, but also a careful statistical methodology.

Acknowledgements

The authors want to thank Prs D. Pinkel and D. G. Albertson, and Dr E. Lebarbier for helpful discussion and comments.

References

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

Albertson, D.G. and Dan Pinkel (2003). Genomic microarrays in human genetic disease and cancer. *Human Molecular Genetics* **12**, 145–152.

Auger, I.E. and C.E. Lawrence (1989). Algorithms for the optimal identification of segments neighborhoods. *Bull. Math. Biol.* **51**, 39–54.

Autio, R., S. Hautaniemi, P. Kauraniemi, O. Yli-Harja, J. Astola, M. Wolf, and A. Kallioniemi (2003). CGH-plotter: MATLAB toolbox for cgh-data analysis. *Bioinformatics* **13**, 1714–1715.

Beheshti, B., P.C. Park, I. Braude, and J.A. Squire (2002). *Molecular Cytogenetics: Protocols and Applications*. Humana Press.

Giegerich, R. (2000). A systematic approach to dynamic programming in bioinformatics. *Bioinformatics 16*(8), 665–677.

Jong, K., E. Marchiori, A. van der Vaart, B. Ylstra, M. Weiss, and G. Meijer (2003). *Applications of Evolutionary Computing: EvoWorkshops 2003: Proceedings*, Volume 2611, Chapter chromosomal breakpoint detection in human cancer, pp. 54–65. Springer-Verlag Heidelberg.

Lavielle, M. (2003). On the use of penalized contrasts for solving inverse problems. application to the change-point problem. *submitted*.

Lebarbier, E. (2003). Detecting multiple change-points in the mean of gaussian process by model selection. Technical Report RR-4740, Institut National de Recherche en Informatique et en Automatique, Rhône-Alpes.

Olshen, A. B. and E. S. Venkatraman (2002). Change-point analysis of array-based comparative genomic hybridization data. Technical report, Memorial Sloan-Kettering Cancer Center, http://www.mskcc.org/mskcc/html/14044.cfm.

Pinkel, D., R. Segraves, D. Sudar, S. Clark, I. Poole, D. Kowbel, C. Collins, W. Kuo, C. Chen, Y. Zhai, S.H. Dairkee, B. Ljung, and J.W. Gray (1998). High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nature Genetics* **20**, 207–211.

Snijders, A. M., N. Nowak, R. Segraves, S. Blakwood, N. Brown, J. Conroy, G. Hamilton, A. K. Hindle, B. Huey, K. Kimura, S. Law, K. Myambo, J. Palmer, B. Ylstra, J. P. Yue, J. W. Gray, A.N. Jain, D. Pinkel, and D. G. Albertson (2001). Assembly of microarrays for genome-wide measurement of DNA copy number. *Nature Genetics* **29**, 263–264.

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

Unité de recherche INRIA Futurs
Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex (France)

Unité de recherche INRIA Lorraine : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex (France)
Unité de recherche INRIA Rennes : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)
Unité de recherche INRIA Rhône-Alpes : 655, avenue de l'Europe - 38330 Montbonnot-St-Martin (France)
Unité de recherche INRIA Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex (France)
Unité de recherche INRIA Sophia Antipolis : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399