A Segmentation-Clustering problem for the analysis of array CGH data

F. Picard[1], S. Robin[1], E. Lebarbier[1], and J-J. Daudin[1]

Institut National Agronomique Paris-Grignon
UMR INA P-G/ENGREF/INRA MIA 518
16 rue Claude Bernard,75231 Paris cedex 05.
(e-mail: picard@inapg.fr)

Abstract. Microarray-CGH experiments are used to detect and map chromosomal imbalances, by hybridizing targets of genomic DNA from a test and a reference sample to sequences immobilized on a slide. A CGH profile can be viewed as a succession of segments that represent homogeneous regions in the genome whose representative sequences (or BACs) share the same relative copy number on average. Segmentation methods constitute a natural framework for the analysis, but they do not assess a biological status to the detected segments. We propose a new model for this segmentation-clustering problem, combining a segmentation model with a mixture model. We present an hybrid algorithm to estimate the parameters of the model by maximum likelihood. This algorithm is based on dynamic programming and on the EM algorithm. We also propose to adaptively estimate the number of segments when the number of clusters is fixed. An example of our procedure is presented, based on publicly available data sets.

Keywords: Segmentation methods, Mixture Models, Dynamic Programming, EM algorithm, Model Selection.

Introduction

Chromosomal aberrations often occur in solid tumors: tumor suppressor genes may be inactivated by physical deletion, and oncogenes activated via duplication in the genome. 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 a CGH profile is constituted for each chromosome when the $\log_2$ of fluorescence ratios 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 on average. Array CGH data are normalized with a 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. It has to be noted that even if the underlying biological process is discrete (counting of relative copy numbers of DNA sequences), the signal under study is viewed as being continuous, because the quantification is based on fluorescence measurements, and because the possible values for chromosomal copy numbers in the test sample may vary considerably, especially in the case of clinical tumor samples that present mixtures of tissues of different natures.

Segmentation methods seem to be a natural framework to handle the spatial coherence on the genome that is a specificity of array CGH data [Autio *et al.*, 2003, Jong *et al.*, 2003]. These methods provide a partition of the data into segments, each segment being characterized by its mean and variance μ_k and σ_k^2 in the Gaussian case. Nevertheless, even if the data are instrinsically segmented, they are also structured into clusters which have a biological interpretation: we can define a group of deleted segments, a group of unaltered segments, and many groups of amplified segments for instance. This refinement means that the mean and variance of each segment should be restricted to a finite set such that $\mu_k \in \{m_1, \ldots, m_P\}$ and $\sigma_k^2 \in \{s_1^2, \ldots, s_P^2\}$ if the segments are structured into P clusters.

We propose to handle this segmentation-clustering problem combining a segmentation model and a mixture model to assign a biological status to segments. Section 1 is devoted to the precise definition of such model. In Section 2 we propose an hybrid algorithm combining dynamic programming and the EM algorithm to alternatively estimate the break-point coordinates and the parameters of the mixture. The convergence properties of this algorithm are presented.

Once the parameters of the model have been estimated, a key issue is the estimation of the number of segments and of the number of clusters. We propose to estimate the number of segments when the number of groups is fixed, using a penalized version of the likelihood. We propose to apply the procedure defined by [Lavielle, 2005], that has been successfully applied to array CGH data [Picard *et al.*, 2005]. An example of our method is provided in Section 3, using publicly available data sets.

1 A new model for the segmentation-clustering problem

Let y_t represent the $\log_2$ ratio of the t^{th} BAC on the genome and $y = \{y_1 \ldots, y_n\}$ the entire CGH profile constituted by n data points. We suppose that y is the realization of a Gaussian process Y whose mean and variance are affected by $K+1$ abrupt changes at unknown coordinates $T = \{t_0, t_1, \ldots, t_K\}$ with the convention $t_0 = 1$ and $t_K = n$. This defines a partition of the data into K segments of length n_k. We write Y as $\{Y^1, \ldots, Y^K\}$, where

$Y^k = \{Y_t, t \in I_k\}$, with $I_k = \{t, t \in]t_{k-1}, t_k]\}$. We suppose that the mean and the variance of the process are constant between two break-points and they are noted μ_k and σ_k^2.

More than classical segmentation models, we assume that the mean and variance of the segment Y^k can only take a limited number of values with $\mu_k \in \{m_1, \ldots, m_P\}$, and $\sigma_k^2 \in \{s_1^2, \ldots, s_P^2\}$. In addition to the spatial organization of the data, via the partition T, there exists a secondary structure of the process into P clusters, and we adopt a mixture model approach to handle this problem.

We assume that the partitionned data $\{Y^1, \ldots, Y^K\}$ are structured into P clusters with weights π_p ($\sum_p \pi_p = 1$). We introduce a sequence of independent hidden random variables, $Z^k = \{Z_1^k, \ldots, Z_P^k\}$ such that Z^k is distributed according to a multinomial distribution consisting of one draw on P categories with probabilities $\pi_1, \ldots, \pi_P$. The mixing proportions $\pi_1, \ldots, \pi_P$ then represent the *prior* probability for segment Y^k to belong to the p^{th} component, while the *posterior* probability of membership to the p^{th} component with y^k having been observed is: $\tau_p^k = \Pr\left\{Z_p^k = 1 | Y^k = y^k\right\}$. Contrary to classical mixture models, where the indicator variables provide informations about the labelling of individual data points (which would be Y_t in our case), our model focuses on the belonging of the segments Y^k to different clusters.

We focus on the case where the data are supposed to be drawn from a mixture of Gaussian densities, with parameters $\theta_p = (m_p, s_p^2)$. If we suppose the indepence of individual data points Y_t within a segment, the model can be formulated as follows:

$$Y^k | Z_p^k = 1 \sim \mathcal{N}(m_p \mathbb{1}_{n_k}, s_p^2 I_{n_k}).$$

We note $\psi = \{\pi_1, \ldots, \pi_{P-1}, \theta_1, \ldots, \theta_P\}$ the vector of unknown independent parameters of the mixture, and the log-likelihood of the model is:

$$\log \mathcal{L}_{KP}(T, \psi) = \sum_{k=1}^{K} \log \left\{ \sum_{p=1}^{P} \pi_p f(y^k; \theta_p) \right\}.$$

$f(y^k; \theta_p)$ represents the conditional density of a vector of size n_k. Our purpose is to optimize this likelihood to estimate the parameters of the model using an hybrid algorithm.

2 An hybrid algorithm combining the EM algorithm and Dynamic Programming

The principle of our algorithm is simple: when the break-point coordinates T are known, the EM algorithm is used to estimate the mixture parameters ψ, and once ψ has been estimated, the break-point coordinates are computed using dynamic programming. This algorithm requires the *prior* knowledge of both the number of segments K and the number of populations P. The choice for these components of the model will be discussed in a later section.

2.1 Estimating the break-point coordinates when the mixture parameters are known

When the number of segments K and the parameters of the mixture are known, the problem is to find the best K-dimensional partition of the data according to the log-likelihood $\log \mathcal{L}_{KP}(T,\psi)$. Since the number of of partitions of a set with n elements into K segments is $\mathcal{C}_{n-1}^{K-1}$, and because of the additivity in K of the log-likelihood, we use a dynamic programming approach to reduce the computational load from $\mathcal{O}(n^K)$ to $\mathcal{O}(n^2)$, as suggested by [Auger and Lawrence, 1989].

Let $\hat{C}_{k+1,P}(i,j;\psi)$ be the maximum log-likelihood obtained by the best partition of the data $Y^{ij} = \{Y_i, Y_{i+1}, ..., Y_j\}$ into $k+1$ segments, when the mixture parameters ψ are known. The algorithm starts as follows: for $k=0$ and for $(i,j) \in [1,n]^2$, with $i<j$, calculate:

$$\hat{C}_{1,P}(i,j;\psi) = \log\left\{\sum_{p=1}^{P} \pi_p f(y^{ij};\theta_p)\right\} = \log\left\{\sum_{p=1}^{P} \pi_p \prod_{t=i+1}^{j} f(y_t;\theta_p)\right\}.$$

$\hat{C}_1(i,j;\psi)$ represents the local log-likelihood for segment Y^{ij}. Then the algorithm is run as follows:

$$\forall k \in [1, K_{max}] \quad \hat{C}_{k+1,P}(1,j;\psi) = \max_h \left\{\hat{C}_{k,P}(1,h;\psi) + \hat{C}_{1,P}(h+1,j;\psi)\right\}$$

Dynamic programming considers 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. More than a reduction in the computational load, this approach provides an exact solution for the global optimum of the likelihood, that will be central for downstream model selection procedures.

2.2 Estimate the mixture model parameters when the break-point coordinates are known

When the break-point coordinates are known, we dispose of a partition of the data into K segments $\{Y^1, \dots, Y^K\}$. This partition defines the statistical units of a mixture model whose parameters have to be estimated. The purpose is then to maximize the log-likelihood of the model $\log \mathcal{L}_{KP}(T,\psi)$ according to ψ. As it is the case in classical mixture models, the direct optimization of the likelihood is impossible, but can be handled using the EM algorithm in the complete-data framework [Dempster *et al.*, 1977]. Let us define the complete-data log-likelihood:

$$\log \mathcal{L}_{KP}^{c}(T,\psi) = \sum_{k=1}^{K}\sum_{p=1}^{P} z_p^k \log\left\{\pi_p f(y^k;\theta_p)\right\}.$$

The EM algorithm is as follows:

- E-**step**: compute the conditional expectation of the complete-data log-likelihood, given the observed data Y, using the current fit $\psi^{(h)}$ for ψ.

$$Q_{KP}(\psi|\psi^{(h)};T) = \sum_{k=1}^{K}\sum_{p=1}^{P} \tau_p^{k(h)} \log\left\{\pi_p f(y^k;\theta_p)\right\},$$

with

$$\tau_p^{k(h+1)} = \frac{\pi_p^{(h)} f(y^k;\theta_p^{(h)})}{\sum_{\ell=1}^{P} \pi_\ell^{(h)} f(y^k;\theta_\ell^{(h)})}.$$

- M-**step**: The M-step on the $(h+1)^{th}$ iteration requires the global maximization of $Q_{KP}(\psi|\psi^{(h)};T)$ with respect to ψ to give the updated estimate $\psi^{(h+1)}$:

$$\psi^{(h+1)} = \underset{\psi}{\text{Argmax}}\left\{Q_{KP}(\psi|\psi^{(h)};T)\right\}.$$

2.3 Convergence properties of the hybrid algorithm

The proof of the convergence of our algorithm is based on the properties of both dynamic programming and EM. It can be seen that both algorithms are linked through the likelihood they alternatively optimize: the incomplete-data likelihood of the mixture of segments.

Dynamic programming globally optimizes the likelihood with respect to T. At iteration (ℓ) we have:

$$\log \mathcal{L}_{KP}\left(T^{(\ell+1)};\psi^{(\ell)}\right) \geq \log \mathcal{L}_{KP}\left(T^{(\ell)},\psi^{(\ell)}\right).$$

On the other hand, the key convergence property of the EM algorithm is the increase of the incomplete-data log-likelihood at each step [Dempster *et al.*, 1977]:

$$\log \mathcal{L}_{KP}\left(T^{(\ell)},\psi^{(\ell+1)}\right) \geq \log \mathcal{L}_{KP}\left(T^{(\ell)},\psi^{(\ell)}\right).$$

Put together, our algorithm generates a sequence $\left(T^{(\ell)},\psi^{(\ell)}\right)_{\ell\geq 0}$ that increases the incomplete-data log-likelihood such as:

$$\log \mathcal{L}_{KP}\left(T^{(\ell+1)},\psi^{(\ell+1)}\right) \geq \log \mathcal{L}_{KP}\left(T^{(\ell)},\psi^{(\ell)}\right).$$

3 Estimating the number of segments K when the number of clusters P is fixed.

Once the parameters of the model have been estimated (for a fixed K and a fixed P), the next question is the estimation of the number of segments and of the number of clusters. Since the principal objective of biologists is rather the detection of biological events on the genome rather than the clustering of those events into groups, we choose to focus on the estimation of the number of segments when the number of groups is fixed.

The maximum of the log-likelihood $\log \hat{\mathcal{L}}_{KP} = \log \mathcal{L}_{KP}(\hat{T}, \hat{\psi})$ can be viewed as a quality measurement of the fit to the data of the model with K segments. In classical segmentation models, this quantity is maximal when the number of segments equals the number of data points. Nevertheless, as our model also considers the clustered nature of segments, it appears that the quality of fit of the model is not always increasing with the number of segments, as shown in Figure 1. For $P = 2$ the incomplete-data log-likelihood is decreasing for a number of segments $K \geq 12$ for instance. This behavior of the model can be interpreted as follows: since the segmentation-clustering model is under the constraint $P \leq K$, the addition of new segments can lead to contiguous segments affected to the same cluster. This configuration leads to an increase in the number of parameters (one additional break-point) without any gain for the fit of the mixture model. These considerations imply that there will be a number of segments above which the addition of a new segment will not increase the log-likelihood.

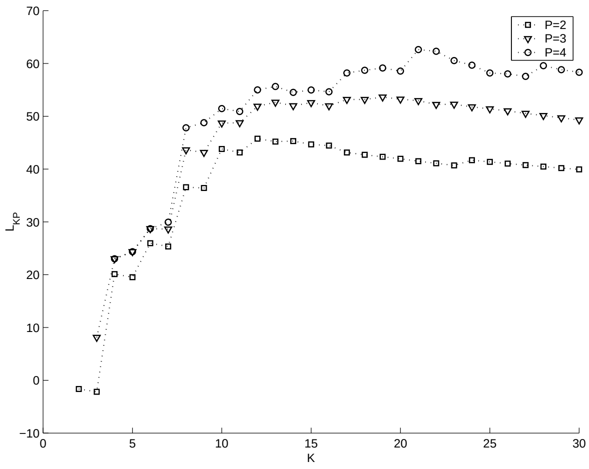

Fig. 1. Evolution of the incomplete-data log-likelihood $\log \hat{\mathcal{L}}_{KP}$ with the number of segments K for different number of clusters ($P = 2, 3, 4$).

A penalized version of the likelihood is used as a trade-off between a good adjustement and a reasonnable number of break-points. The estimated number of segments is such as:

$$\hat{K}_P = \underset{K}{Argmax} \left(\hat{\mathcal{L}}_{KP} - \beta_P pen(K) \right),$$

with $pen(K)$ a penalty function that increases with the number of segments, and β_P a penalty constant. The definition of an appropriate penalty function and constant has lead to theoretical developments in the context of breakpoint detection models. Recently, [Lavielle, 2005] proposed to use an adaptive procedure to estimate the penalty constant, that has been successfully applied to array CGH data [Picard *et al.*, 2005]. The principle of this procedure is to find the number of segments for which the log-likelihood ceases to increase significantly. It is geometrically linked to the finding of the number of segments for which the second derivative of the log-likelihood function is maximal (see [Lavielle, 2005] for further details). A result of our procedure is shown in Figure 2. For a number of clusters $P = 3$, the adpative procedure estimates a number of segments $\hat{K}_3 = 10$. This leads to a profile which presents three types of segments that can be interpreted in terms of biological groups, as shown in Figure 2.

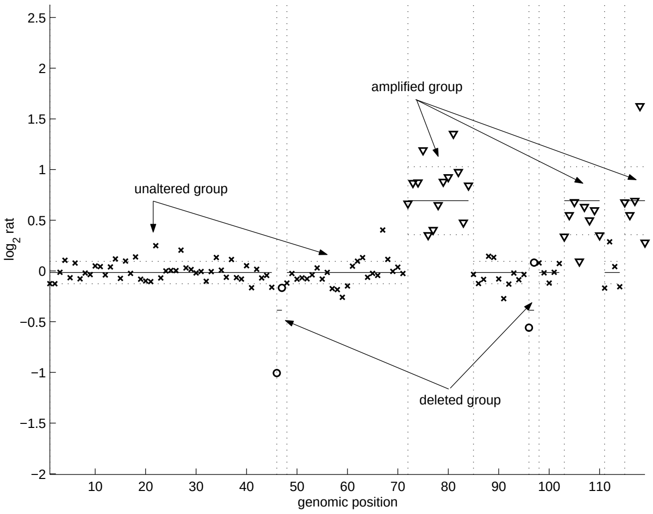

Fig. 2. Result of the segmentation-clustering procedure for a fixed number of clusters $P = 3$ and an estimated number of segments $\hat{K}_3 = 10$. These data concern chromosome 1 of breast cancer cell lines Bt474.

4 Discussion

Microarray CGH currently constitutes the most powerful method to detect gain or loss of genetic material on a genomic scale. We introduced a statistical methodology for the analysis of CGH microarray data, that combines segmentation methods and clustering techniques. It terms of modeling, the discovery of homogeneous regions clustered into groups could have been handled using Hidden Markov Models, as in [Fridlyand *et al.*, 2004]. In those models, the segmented structure of the data is recovered using the *posterior* probability of membership of individual data points into a fixed number of hidden groups, whereas our method focuses on the labelling of segments to hidden groups. Moreover, a property of Hidden Markov Models is that the distance between two 'break-points' is dependent on the probability distribution of the hidden sequence: the within-class sojourn time is geometrically distributed. Our approach is free from those constraints, since break-point coordinates are 'real' parameters of the model that are not randomly distributed.

The definition of this new model leads to unusual statistical considerations: it appears that the statistical units of the mixture model (when the segmentation is known) are segments of different size. Since the partition of the data is random, the individuals of the mixture model themselves are random. This explains the difficulty of the joint estimation of K the number of segments, and P the number of clusters, since classical model selection procedures are based on a compromize between a reasonnable number of parameters to estimate given a fixed number of statistical units. To these extents, this problem of model selection for two components remains an open question.

References

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

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

[Dempster *et al.*, 1977]A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B*, 39:1–38, 1977.

[Fridlyand *et al.*, 2004]J. Fridlyand, A. Snijders, D. Pinkel, D.G. Albertson, and A.N. Jain. Hidden markov models approach to the analysis of array CGH data. *Journal of Multivariate Analysis*, 90(1):132–1533, 2004.

[Jong *et al.*, 2003]K. Jong, E. Marchiori, A. van der Vaart, B. Ylstra, M. Weiss, and G. Meijer. *Applications of Evolutionary Computing: EvoWorkshops 2003: Proceedings*, volume 2611, chapter chromosomal breakpoint detection in human cancer, pages 54–65. Springer-Verlag Heidelberg, 2003.

[Lavielle, 2005]M. Lavielle. Using penalized contrasts for the change-point problem. *(to appear in) Signal Processing*, 2005.

[Picard *et al.*, 2005]F. Picard, S. Robin, M. Lavielle, C. Vaisse, and J-J. Daudin. A statistical approach for CGH microarray data analysis. *BMC Bioinformatics*, 6:27, 2005.