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


UMR INA P-G/ENGREF/INRA MIA 518
16 rue Claude Bernard, 75231 Paris cedex 05.

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 sharing 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. 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, model selection.

Résumé

L’objectif des expériences de microarrays CGH est de détecter et de cartographier des déséquilibres chromosomiques. Des cibles d’ADN génomique provenant d’un échantillon test et d’un échantillon de référence sont hybridées à des séquences immobilisées sur une lame de verre. Un profil CGH peut être interprété comme une succession de segments qui représentent des régions homogènes sur le génome partageant le même nombre relatif de copies en moyenne. Les méthodes de segmentation constituent un cadre naturel pour l’analyse, mais ne permettent pas d’assigner un statut biologique aux segments détectés. Nous proposons un nouveau modèle pour ce problème de segmentation/classification, en combinant un modèle de segmentation à un modèle de mélange. Nous présentons un algorithme hybride pour estimer les paramètres du modèle par maximum de vraisemblance. Nous proposons également d’estimer le nombre de segments quand le nombre de groupes est fixé. Un exemple de notre procédure est présenté, à partir de données publiques.

mots clés: méthodes de segmentation, modèles de mélange, sélection de modèle.
Introduction

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. Two samples of genomic DNA (referred as the reference and the test DNA) are labelled with fluorescent dyes and hybridized to known mapped sequences (referred as BACs) that are immobilized on a slide. 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 \( \log_2(\text{ratio}) = 0 \) 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. Segmentation methods seem to be a natural framework for the analysis [3, 5] as they provide a partition of the data into segments, each segment being characterized by its mean and variance in the Gaussian case. Nevertheless, even if the data are intrinsically 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. We propose to handle this segmentation-clustering problem combining a segmentation model and a mixture model to assign a biological status to segments. 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.

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 in [4], that has been successfully applied to array CGH data [5]. 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 $\mu_k$ and $\sigma_k^2$, the mean and the variance of the process are constant between two break-points.

We assume that the partitionned data $\{Y^1, \ldots, Y^K\}$ are structured into $P$ clusters with weights $\pi_p$ ($\sum_p \pi_p = 1$). We introduce a sequence of independent hidden random variables, $Z^k = \{Z^k_1, \ldots, Z^k_P\}$ 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: $\pi^k_p = \Pr \{Z^k_p = 1 | Y^k = y^k\}$.

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

$$Y^k | Z^k_p = 1 \sim \mathcal{N}(m_p \| n_k, \sigma_p^2 I_{n_k}).$$

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 to alternatively estimate the break-point coordinates $T$ with dynamic programming and the mixture parameters $\psi$ with the EM algorithm. It requires the knowledge of the number of segments $K$ and the number of populations $P$ whose whose choice will be discussed in a later section.

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 $C_{n-1}^{K-1}$, and because of the additivity in $K$ of the log-likelihood, we use a dynamic programming approach to reduce the computational load from $O(n^K)$ to $O(n^2)$ [1].

Let $\hat{C}_{k+1,P}(i, j; \psi)$ be the maximum log-likelihood obtained by the best partition of the data $Y^{ij} = \{Y_i, Y_{i+1}, \ldots, Y_j\}$ into $k + 1$ segments, when the mixture parameters $\psi$ 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.

When the break-point coordinates are known, we dispose of a partition of the data into \( K \) segments \( \{Y^1, \ldots, Y^K\} \). 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 \( \log L_{KP}(T, \psi) \) according to \( \psi \) using the EM algorithm in the complete-data framework [2]. 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 \( \psi \).

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

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

- **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 \( \psi \) to give the updated estimate \( \psi^{(h+1)} \):

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

### 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 $\hat{L}_{KP}$ can be viewed as a quality measurement of the fit to the data of the model with $K$ segments that should be maximal when the number of segments equals the number of data points. Nevertheless, as our model also considers the clustered nature of segments, the quality of fit of the model is not always increasing with the number of segments, as shown in Figure 1. 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 adjustment 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.

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

$$\hat{K}_P = \arg \max_K \left( \log \hat{L}_{KP} - \beta_P pen(K) \right),$$

with $pen(K)$ a penalty function that increases with the number of segments, and $\beta_P$ a penalty constant. Recently, [4] proposed to use an adaptive procedure to estimate the penalty constant, that has been successfully applied to array CGH data [5]. The principle of this procedure is to find the number of segments for which the log-likelihood ceases to increase significantly (see [4] for further details). A result of our procedure is shown in Figure 1. For a number of clusters $P = 3$, the adaptive 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 1.

4 Discussion

We introduced a statistical methodology for the analysis of CGH microarray data, that combines segmentation methods and clustering techniques. 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 compromise between a reasonable 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.
Figure 1: Left: Incomplete-data log-likelihood according to the number of segments $K$ for different number of clusters ($P = 2, 3, 4$). Right: Segmentation-clustering for $P = 3$ clusters and an estimated number of segments $\hat{K}_3 = 10$. These data concern chromosome 1 of breast cancer cell lines Bt474.

References


