

A flexible statistical method for detecting
genomic copy-number changes using Hidden
Markov Models with Reversible Jump MCMC

Oscar M. Rueda and Ramón Díaz-Uriarte
Statistical Computing Team
Structural and Computational Biology Programme
Spanish National Cancer Centre (CNIO)
Melchor Fernández Almagro 3, 28029 Madrid, Spain
omrueda@cnio.es, rdiaz02@gmail.com

August 24, 2006

Abstract

We have developed a statistical method for the analysis of array-based CGH data to detect genomic DNA copy number changes. Our method allows us to answer the biologically relevant questions (what is the probability that a given gene or region has increased or decreased copy number changes) in a clear and simple way, within a rigorous statistical framework. We use a non-homogeneous Hidden Markov Model that incorporates distance between genes, a crucial requirement to analyze data from platforms where distances between probes is highly variable. As the true number of hidden states (states of copy number changes) is not known in advance in biological samples, we do not fix the number of hidden states of the model, but use Reversible Jump Markov Chain Monte Carlo for inference. We can therefore investigate the likely number of hidden states in the data and, more importantly, provide posterior probabilities that a gene or a set of genes is in a given state. To summarize results, we employ Bayesian Model Averaging, averaging over models with different states, and thus incorporating model uncertainty. Our method can be used to analyze data from each chromosome independently or all chromosomes together, offering both flexibility in the biological phenomena studied and increased statistical precision. Thus, our method provides a rigorous statistical foundation for locating genes and chromosomal regions with altered copy number and potentially related to cancer and other complex diseases.

1 Introduction

Alterations in the number of copies (gains, losses) of genomic DNA have been associated with several hereditary anomalies and are involved in human cancers (reviews and examples in [PA05, LCCL06, UKS⁺06, MPN⁺05a, ABB⁺04, SLT⁺04, FMM⁺00]). For example, amplification of some genes, especially oncogenes, is one well known mechanism for tumor activation [HBC⁺00, HKS⁺04] and it is involved in the deregulation of cellular control [VFP⁺03, VK04]. Copy number alterations has been associated with tumoral grade, metastasis development, and patient survival [PA05, LCCL06, UKS⁺06, MPN⁺05a, ABB⁺04, SLT⁺04, FMM⁺00], and studies about copy number changes have been instrumental for identifying relevant genes for cancer development [PA05, LCCL06, PSP⁺02].

A widely used technique to identify copy number changes in genomic DNA is array-based Comparative Genomic Hybridization (aCGH). Two DNA samples (e.g., problem and control) are differentially labeled (often with fluorescent dyes) and competitively hybridized to chromosomal DNA targets. After hybridization, emission from each of the two fluorescent dyes is measured, and the signal intensity ratios are indicative of the relative copy number of the two samples (see reviews in [PA05, LCCL06]). Therefore, a key step in any study of the relationship between altered copy numbers and disease is using the fluorescence ratio data to identify genes and contiguous chromosomal regions with altered copy numbers.

The analysis of aCGH data should use a rigorous statistical method that is an appropriate model for the underlying biological phenomenon and allows the biologist to answer important questions in a clear and simple way. First, this method should incorporate the distance between probes [MTT06, FSPA04, LJKP05, BR06, HWLZ05]: widely used aCGH platforms like those based on cDNA microarrays and ROMA lead to variable coverage across chromosomes [LCCL06], with very unequal distances between probes (i.e., some regions have probes that are very close to each other, whereas in other regions probes are very far apart). As copy number changes involve chromosome segments, contiguous loci will have the same copy number, unless there is an abrupt change to another copy number [PA05, DRO⁺04]; the further apart two loci are, the more likely it is that a copy number event will have taken place in between them. Thus, in densely covered regions the copy number of a probe is a good predictor of the copy number of the neighboring probe. In contrast, in poorly covered regions, contiguous probes or loci might be many thousands of kilobases apart, making it more likely that at least one copy number change has taken place, and consequently a probe provides less information about the likely state of its neighboring probes. Therefore, unless we use a platform where all probes are equally spaced, we need to use the distance between probes (and not just the order), so that the information that consecutive probes provide is adequately accounted for.

Second, a flexible method, applicable over different diseases and platforms, must not require the researcher to specify in advance the likely number of copy number levels for two main reasons (see [DRO⁺04, PRL⁺05, PA05, LCCL06]). How many levels a statistical method can detect can be strongly limited (or enhanced) by the aCGH technology and the number of copy number levels can be a characteristic of the disease under study. Whereas for some combinations of disease/platform a few copy number levels might suffice (e.g., loss, no-change, gain), for other cases a finer gradation might be required for appropriate characterization of the syndrome. Consequently, in the absence of overwhelming previous information about the true number of copy number alterations, we will want a method that does not impose in advance a fixed number of states.

Third, we will want to be able to analyze the data either chromosome by chromosome, or genome-wide [SXD⁺06, BR06, EMLB06]. Analysis at the chromosome level are ideal to detect alterations in copy number of loci relative to the rest of the loci in that same chromosome, regardless of that chromosome's ploidy (a trivial example would be detection of copy number changes in loci of the human Y chromosome in an otherwise diploid genome). On the other hand, detection of copy number changes that affect most of a chromosome often require genome-wide analysis (in chromosome-wide analysis, as the mean or median chromosome level is used as the reference, detection of such changes is virtually impossible). Moreover, the use of genome-wide analysis can offer statistical advantages (e.g., reduced variance of estimation), if we are willing to assume that certain features are similar over chromosomes (e.g., the mean log ratio corresponding to one deletion). Often, both types of analysis can offer complementary information, as they focus on different biological phenomena

(e.g., chromosomal gains/losses vs. gains of loci within chromosome) and, thus, we would like to use a method that offers these two approaches.

Finally, we will want a method that returns results that can be used in contexts that cover from basic research to clinical applications [PA05, LCCL06]. Probabilities of alteration, both of individual positions and contiguous sets of genes (segments) [EMLB06, BR06, GLN06] (without underestimating the uncertainty in model building [SXD⁺06]) can be immediately adapted to the context of the study, so that a researcher can choose to consider genes/regions with just moderate probability of alteration (e.g., > 0.5) for further study, whereas a clinician might want to require high certainty of alteration of a specific gene before more invasive procedures.

We have developed a method to fulfill the above requirements. We use a non-homogeneous HMM which provides a natural model for these type of data and allows us to incorporate the distance between clones/genes. As we will not fix the number of hidden states in advance, we can consider a family of models with varying number of states and let the data indicate the likely number of hidden states. To obtain our final estimates, we incorporate the uncertainty in model selection by using Bayesian Model Averaging. Our method allows fitting either individual chromosomes or all chromosomes jointly. Finally, we can obtain posterior estimates of the probability that each clone/gene has an altered copy number and we can also return the posterior probability of the most likely state of complete segments.

1.1 Statistical model: overview

For a given chromosome or genome, the copy numbers of genomic DNA (e.g., 0, 1, 2 copies, ...) of different genes or segments are an unknown finite number. Thus, genes or segments could be classified into several groups with respect to their (unknown) copy number. We expect that the copy number of a gene will be similar to the copy number of its closest neighbors, with that expected similarity decreasing when genes are further apart (see above).

We cannot observe directly the true copy number but, rather, with aCGH, we measure the fluorescence ratios between two samples. For a given alteration of copy number, the fluorescence ratios should be centered around a \log_2 value, with some random noise, and we want to use the observed log-ratios to estimate the underlying copy numbers and/or identify regions with altered copy number. The above features (a finite number of unknown or hidden states that are indirectly measured, with states of close elements likely to be similar) are often modelled with a Hidden Markov Model (HMM) [Rab90, CMR05]. As distances between genes are not constant, and we want to incorporate distance between genes in our model, we use a non-homogeneous HMM [CMR05].

We will fit the above model in a Bayesian framework, so that we can answer the biological question of what is the probability that a given gene or region has increased or decreased copy number changes. For computation, we will use Markov Chain Monte Carlo (MCMC) [GCSR03]. Since we do not know the true number of hidden states, we will fit models with varying number of hidden states

and, to allow for transdimensional moves between models with different numbers of states, we will use Reversible Jump [Gre95]. After running a large number of MCMC iterations, we can summarize the posterior probabilities. First, we will obtain posterior probabilities for the number of states. Conditional on a given number of states, each model will provide posterior distributions of the parameters of interest (e.g., means, variances, transition matrices). From these, we can obtain posterior probabilities that a gene is gained or lost. Finally, we will use Bayesian Model Averaging [HMRV99] to obtain estimates, weighted by posterior probability of each number of states, for the probabilities of genes being gained or lost.

2 Example

We have analyzed the well known nine cell lines from Snijders et al. [SNS⁺01] available from http://www.nature.com/ng/journal/v29/n3/supinfo/ng754_S1.html and we have compared the results from our method with the known ploidy, as provided by Snijders et al. We ran four parallel chains, each with 20,000 iterations of which the first 10,000 were discarded as burn-in. The parameters of the distributions of the candidates were selected automatically by a heuristic approach that, within model, leads to an acceptance probability near 0.23 [GCSR03]. The parameters of the jumps between models were taken as the mean of the within model parameters.

Figure 1 shows the results of the analysis for the complete genome of the cell line gm03563. Panel a) indicates a large posterior probability of a model with four hidden states; two of the states of the four-state model, however, are extremely close to each other (panel b) and, because of their posterior means (panel b) and variances (panel c) we consider them to represent the same biological state of no change in copy number. The other two states are well separated, with posterior means clearly negative or positive, so we regard them as biological states of loss and gain of copy number. Note that the component that represents the hidden state of loss is assigned to only two genes (panel e, green dots), exactly the same two genes whose true state is loss (see also table 1) [SNS⁺01]. Panel d) shows that the probability of remaining on the same state decreases as distance increases, eventually becoming $0.25 (= \frac{1}{\text{Number hidden states}})$. Finally, panel e) shows the results from the Bayesian Model Averaging. This is a particularly clear-cut model, as the posterior probabilities that each gene belongs to the state with highest posterior is very high (the lower blue line is > 0.9 for almost all genes). Table 1 shows the comparison of the predictions from our method with the true ploidy provided by Snijders et al. The classification error is very low (0.009), lower than that obtained with the Circular Binary Segmentation procedure [OVLW04], considered one of the currently best performing methods [WF05, LJKP05].

3 Discussion

We have developed a method to analyze aCGH for copy number changes that incorporates distance between genes, does not fix in advance the number of hidden states, accounts for model selection uncertainty, and allows to analyze one or more chromosomes simultaneously. Our method provides clear answers to biological questions, returning posterior probabilities that a gene or chromosomal region presents gains or losses of genomic DNA. Many of the currently available methods have some of these features, but none of the existing methods combines all these features. Many smoothing techniques (e.g., [HST⁺04, OVLW04, PRM⁺05, HSG⁺05, LBL⁺05, HWLZ05, PRL⁺05]), do not allow the incorporation of gene distance nor provide posterior estimates of the likely state of each gene/clone thus making interpretation harder. Moreover, data from each chromosome are analyzed independently of each other (except [LBL⁺05, PRM⁺05] can be modified to use information from the complete genome). In addition, most smoothing techniques treat segmentation and classification into underlying states (loss, no-change, etc) as separate steps, introducing the need for ad-hoc procedures, and not allowing to use the class labels in the segmentation itself [EMLB06, SXD⁺06, WF05] (see also below).

HMMs and related techniques offer potentially easier biological interpretation of results. However, some implementations [FSPA04, SXD⁺06, GLN06, BR06], do not allow to incorporate distance between genes. The approach of Marioni et al. [MTT06] does use, like ours, a non-homogeneous HMM to incorporate distance between genes, but Marioni et al. [MTT06], as Fridlyand et al. [FSPA04], analyze the data chromosome by chromosome. More importantly, both Marioni et al. and Fridlyand et al. include a post-HMM step where a model (number of states) is selected using AIC or BIC criteria and clustering is used for further merging of neighboring states. The use of AIC and BIC with HMMs has not been theoretically justified [CMR05], does not provide a probability of the likely number of states, and selecting a single model underestimates the true variability in the data (in contrast to the Bayesian Model Averaging we use); moreover, the final clustering step introduces several ad-hoc decisions and render a model that is not a natural extension of the original HMM.

Together with Engler et al., [EMLB06], Shah et al. [SX⁺06], Bret and Richardson [BR06], and Guha et al. [GLN06], our method is one of the few to provide quantitative measures of the likely state of each clone and segment, which we return via posterior probabilities. However, these approaches differ from ours in that they fix in advance the number of hidden states: four states in [SX⁺06] and [GLN06], three states in [BR06] and [EMLB06]. Prespecification of the number of states, often with the consequence of lumping together all changes involving multiple gains into a single state with a common mean, seems biologically questionable [DRO⁺04]. Instead, our approach provides posterior estimates of the probability of models with different number of states; using such a model over different experiments will tell whether four- or three-state models are a reasonable simplification.

Daruwala et al. [DRO⁺04] have developed a Bayesian approach that returns

the maximum a posteriori segmentation solution for a given data. Like our approach, and in contrast to [EMLB06, SxD⁺06, BR06, GLN06], Daruwala et al. do not assume that all gains can be modeled by a single Gaussian. In terms of assigning genes and regions to the gained/lost/no-change categories, both Daruwala et al.'s and our approach can be used; from the perspective of the practitioner, the main difference are the parameters that need to be set: p_r, p_b in Daruwala et al. vs. the priors in eq. 5 and 7. Their approach, however, is not suitable for answering questions regarding the number of hidden states, or for breaking the data into more categories than gained/lost/no-change.

Recent advances in aCGH methodology [RSH⁺06, DEG⁺06] are focusing on the identification of regions that show consistent alterations across samples. The framework we have developed offers a direct approach to this problem: we provide posterior probabilities of gain/loss/no-change for individual genes and, more importantly, for contiguous genes (segments) from the Viterbi algorithm (to obtain the probability that a given sequence is altered we cannot simply multiply the marginal probabilities, since genes states are not independent). These posterior probabilities allow us to identify regions with consistent alterations across samples in a statistically rigorous way, including control of False Discovery Rate (FDR) and detection of subgroups of samples according to recurrence patterns [MPN⁺05b].

In conclusion, our method is based both on plausible biological assumptions and on a sound statistical model. It allows the biologist to answer in an objective way questions about the probability of a gene or region having an altered copy number.

4 Material and methods

4.1 Model

We use a non-homogeneous Hidden Markov Model with Gaussian emissions. We can either fit one model to all the chromosomes of an array or we can fit a different model for each chromosome of an array. Let n be the number of genes, and K the number of different copy numbers in the collection of genes. Let S_i be the true state (copy number) of the gene i : $S_i = \{1, \dots, K\}_{i=1, \dots, n}$. Let Y_i be the relative copy number of the gene i , that is the log ratio of fluorescence intensities between tumor and control samples. Let X_i be the distance in bases between gene i and gene $i + 1$ (we normalize these distances between 0 and 1 to increase numerical stability). How distance is measured depends on the platform: distance can be the distance from the end of the spot to the start of the next, if the length of the spots is proportional to the length of the gene (so we have the same information for every gene), or the distance between the midpoint of the spots, if the length of the spots is not proportional to the length of the gene.

We assume that $\{S_i\}$ follows a non-homogeneous 1st order Markov process, as: $P(S_i = s_i | S_{i-1} = s_{i-1}, X_{i-1} = x_{i-1}) = Q_{s_i, s_{i-1}, x_{i-1}}$. Biologically, we

expect that $Q(S_i = r, S_{i-1} = r, X_{i-1})$, the probability of staying in the same hidden state, is a decreasing function of X_{i-1} , so the dependence of the state of a gene onto the next one is lower the further the genes are. We also expect that when the distance between two genes is maximal, the state of a gene should be independent from the state of its predecessor. Thus, we model the transition probabilities functions as:

$$Q_{i,j,x} = \frac{\exp\{-\beta_{i,j} + \beta_{i,j}x\}}{\sum_{p=1}^k \exp\{-\beta_p + \beta_p x\}} \quad (1)$$

Where β has the form:

$$\beta = \begin{pmatrix} 0 & \beta_1 & \dots & \beta_{k-1} \\ \beta_k & 0 & \dots & \beta_{2k-1} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{(k-1)(k-1)-(k-1)} & \beta_{(k-1)(k-1)-k} & \dots & 0 \end{pmatrix} \quad (2)$$

With all $\beta_i \geq 0 \quad \forall i$.

Finally, conditioned on $\{S_i\}$, $\{Y_i\}$ follows a Gaussian process: $(Y_i|S_i = s_i) \sim N(\mu_{s_i}, \sigma_{s_i}^2)$.

This model can be fitted by maximum likelihood using a modified version of the EM algorithm [HGC99] or directly by numerical methods [MTT06]. For computational reasons and modeling flexibility, we opted for Bayesian methods using Markov Chain Monte Carlo. To fit models with varying number of hidden states we will use Reversible Jump. Suppose that we have a collection of K HMM models, and each of them has a number of k hidden states, from $k = \{1, \dots, K\}$. Let $\theta(k)$ be the HMM associated to k , that is $\theta(k) = \{\mu(k), \sigma^2(k), \beta(k)\}$. The prior distributions for the model are the usual ones in mixture problems [RG97, MP00]: $p(k)$ is the prior for the number of hidden states with $p(k) \sim U(1, k)$, $p(\theta(k)/k)$ is the prior of the HMM conditioned to k , the number of hidden states with $\mu(k) \sim N(\alpha, \varrho^2)$, where α and ϱ are the median and range of Y_i ; $\sigma^2(k) \sim IG(ka, g)$, where ka is 2 and g is $\varrho^2(Y_i)/50$; $\beta(k) \sim \Gamma(1, 1)$.

The likelihood of the model, $L(y; k, \theta(k))$ can be computed by Forward Filtering [Rab90], so the joint distribution is $p(k)p(\theta(k)/k)L(y; k, \theta(k))$.

4.2 Estimation and fitting

We can draw samples from the posterior distribution through a Reversible Jump Markov Chain Monte Carlo (RJMCMC) algorithm [Gre95]. In typical Markov Chain Monte Carlo, we explore the posterior distribution of a model drawing samples from a Markov Chain whose stationary distribution is that posterior [GCSR03].

In RJMCMC, we explore the posterior distribution of possible models, jumping not only within a model but also between models with a different number of parameters. To match the difference between degrees of freedom, some random

numbers u with density $P(u)$ are generated, so if we are in state x , the new one is proposed in a deterministic way $x'(x, u)$. The reverse move is the inverse of that function: $x(x', u')$. This way, the usual Metropolis-Hastings acceptance probability can be computed (notation from [RG97]):

$$\min \left\{ 1, \frac{L(y/x)p(x')p(u'/x')}{L(y/x)p(x)p(u/x)} |J| \right\} \quad (3)$$

where $L(y/x)$ is the likelihood, $p(x)$ are the priors, $p(u/x)$ are the densities of the candidates, and $J = \frac{\partial x'}{\partial(x,u)}$, the Jacobian of the change of variable.

We combine several Metropolis steps in a sweep, as in [CMR05, RRT00]:

1. Update HMM of a model using a series of Metropolis-Hastings moves. (We do not use Gibbs Sampler to avoid the hidden state sequence from becoming part of the state space of the sampler, so dimensionality is reduced and reaching convergence is easier).
2. Update model (birth/death). When we have r states, a birth/death move is chosen with probabilities $p_{birth}(r)$ and $p_{death}(r)$ (these are 1/2 except in the cases when no movement of that type can be made, e.g. a death move when there is only one state). If a birth move is selected a new one is created from the prior distributions and accepted with probability

$$\min \left\{ 1, \frac{p(k = r + 1)L(y; r + 1, \theta(r + 1))p_{death}(r + 1)}{p(k = r)L(y; r, \theta(r))p_{birth}(r)} \right\} \quad (4)$$

If a death move is chosen, a random state is deleted with a probability inverse to eq.[4].

3. Update model (split/combine). A split/combine move is attempted with probabilities $p_{split}(r)$ and $p_{combine}(r)$ (again, 1/2 except when a move can not be made). If a split move is selected, an existing state i_0 is split into two, i_1, i_2 :

$$\mu_{i_1} = \mu_{i_0} - \epsilon_\mu, \quad \mu_{i_2} = \mu_{i_0} + \epsilon_\mu, \quad \epsilon_\mu \sim N(0, \tau_\mu) \quad (5)$$

$$\sigma_{i_1}^2 = \sigma_{i_0}^2 \epsilon_\sigma, \quad \sigma_{i_2}^2 = \sigma_{i_0}^2 (1 - \epsilon_\sigma), \quad \epsilon_\sigma \sim \beta(2, 2) \quad (6)$$

Split column

$$i_0 : \quad \beta_{i, i_1} = \beta_{i, i_0} \epsilon_\beta, \quad \beta_{i, i_2} = \beta_{i, i_0} / \epsilon_\beta, \\ \epsilon_\beta \sim LN(0, \tau_\beta) \quad \text{for } i \neq i_0 \quad (7)$$

Split row

$$i_0 : \quad \beta_{i_1, j} = \beta_{i_0, j} U_j, \quad \beta_{i_2, j} = \beta_{i_0, j} (1 - U_j), \\ \text{where } U_j \sim \beta(2, 2) \quad \text{for } j \neq i_0 \\ \beta_{i_1, i_2} \sim \Gamma(1, 1) \quad (8)$$

This move is accepted with probability

$$\begin{aligned}
& \min\{1, p\} \quad \text{where} \\
p &= \frac{1}{2P(\epsilon_\mu)P(\epsilon_\sigma) \prod P(\epsilon_\beta) \prod P(U_j)} J_{split} \\
& \times \frac{P(k=r+1)P(\theta(r+1))L(y; r+1, \theta(r+1))(r+1)}{P(k=r)P(\theta(r))L(y; r, \theta(r))} \\
& \quad \text{and} \quad J_{split} = 2^r \sigma_{i_0}^2 \prod_{j \neq i_0} \beta_{i_0, j} \prod_{i \neq i_0} \frac{\beta_{i, i_0}}{\epsilon_\beta}
\end{aligned} \tag{9}$$

The split move must follow the adjacency condition in [RG97] (the resulting states must be closer between them than to any other of the existing ones). If a combine step is selected, the symmetric move is performed and the inverse probability of acceptance is computed.

The combination of birth and split moves makes it possible not only to visit models with different number of parameters, but also to explore more thoroughly the posterior probability in the case of a parameter with a multi-modal density.

These moves are similar to [CMR05, RRT00], but we have change several aspects of their design to improve the probability of acceptance, which is the most difficult step in Reversible Jump [CMR05, Gre95, RRT00]. We constraint the variance of every state so that it can not be greater than the variance of the whole data and we have added the adjacency condition mentioned before, and used centring proposals [BGR03].

An important problem in mixture models is label-switching of states [Ste00], which arises because the likelihood is invariant under permutations of the states. We have tried the methods in [Ste00], but found that in our case, as there is a natural ordering in the means of the states, the results are similar and much more time consuming than simple alternatives such as ordering the states according to means after every iteration of the sweep, as in [RG97].

4.3 Inference

We run the former algorithm a large number of times (e.g., 20000) and, after discarding the first iterations as burn-in, we keep the samples as observations from the joint distribution, so we can make inferences from it. For every model that has been visited we obtain the posterior probabilities of the mean copy number of every state, the variance of the copy number of every state, and the function of transitions between hidden states. By counting the number of times that each model has been visited we obtain an estimate of the posterior probability of each model (i.e., we avoid using BIC or AIC). Then, applying the Viterbi algorithm [CMR05] to every sample obtained from the MCMC, and as this sample is a function of the HMM, we can obtain its posterior probability, something that usual Viterbi can not. From the Viterbi paths for all the samples,

we can then compute the posterior probability that a gene belongs to every state or the probability that a sequence of genes is in a given state.

When obtaining posterior probabilities of copy number change, we use Bayesian Model Averaging [HMRV99] over all models visited. Let S_i be the lost, gained, no-change status of gene i , K the set of the models considered (in our case, that would be HMMs with $1, \dots, K$ number of states), M_k the model with k number of states and S_i/M_k the state of gene i according to model k . We compute the unconditional probability for the gene i as:

$$p(S_i = s_i) = \sum_{k \in K} p(M_k|y)p(S_i = s_i|M_k, y) \quad (10)$$

When analyzing multiple arrays, it is straightforward to use our approach to identify genes that show consistent copy number alterations across samples as $p(S_i = s_i) = \sum_{j=1}^N \sum_{k \in K} p(M_k|y_j)p(S_i = s_i|M_k, y_j)p(y_j)$ where y_j are the data from array j and we have N arrays. If we have information about the reliability/representativeness of an array, that can be incorporated via $p(y_j)$; otherwise we set $p(y_j) = 1/N$.

4.4 Checking convergence and influence of priors

As in any MCMC approach, it is crucial to assess convergence of the sampler. We have followed the usual [BG98] approach using several chains run in parallel. The convergence of the sampler depends strongly on the distribution of the candidates in Metropolis-Hastings. That is, every iteration a new value for the parameters is proposed from a distribution centered in their current values. The standard deviation of that distribution must be chosen in a way that samples explore all the space parameters. These standard deviations are not parameters of the model in the sense that different values give different fits, but values that can speed up convergence of the algorithm. Sometimes it must be necessary some trial and error initial fits to choose parameters that achieve convergence. The convergence of the posterior probability of the number of hidden states is reached when a large enough number of transdimensional moves is made. This number need not to be great if the likelihood is substantially higher in a particular model and data size is big enough. The birth and death moves only depend on the priors, but the split and combine moves depend also on their own design and the values of τ_μ and τ_β (see eq. [5] and eq. [7]).

The priors chosen have been extensively tested in mixture models [RG97].

In addition, the priors and rest of the parameters have very little effects: even small CGH arrays contain thousands of points so that the likelihood from the data dominates any prior.

4.5 Implementation

We have implemented this model using C (for the sweep algorithm) and R [RD06]. The code is available from CRAN

(<http://cran.r-project.org/src/contrib/Descriptions/RJaCGH.html>) and from the Asterias site (<http://www.asterias.info>).

5 Acknowledgments

Funding provided by Fundación de Investigación Médica Mutua Madrileña and Project TIC2003-09331-C02-02 of the Spanish Ministry of Education and Science (MEC). R.D.-U. partially supported by the Ramón y Cajal programme of the Spanish MEC. C. Lázaro-Perea, J. F. Poyatos, and A. Alibés provided comment on the ms.

References

- [ABB⁺04] Andrew J. Aguirre, Cameron Brennan, Gerald Bailey, Raktim Sinha, Bin Feng, Christopher Leo, Yunyu Zhang, Jean Zhang, Joseph D. Gans, Nabeel Bardeesy, Craig Cauwels, Carlos Cordon-Cardo, Mark S. Redston, Ronald A. Depinho, and Lynda Chin. High-resolution characterization of the pancreatic adenocarcinoma genome. *Proc Natl Acad Sci U S A*, 101(24):9067–9072, 2004.
- [BG98] S.P. Brooks and A. Gelman. General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7:434–455, 1998.
- [BGR03] S. P. Brooks, P. Giudici, and G. O. Roberts. Efficient construction of reversible jump markov chain monte carlo proposal distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(1):3–39, 2003.
- [BR06] P. Broët and S. Richardson. Detection of gene copy number changes in cgh microarrays using a spatially correlated mixture model. *Bioinformatics*, 22(8):911–918, April 2006.
- [CMR05] Olivier Cappé, Eric Moulines, and Tobias Ryden. *Inference in Hidden Markov Models (Springer Series in Statistics)*. Springer, August 2005.
- [DEG⁺06] Sharon J J. Diskin, Thomas Eck, Joel Greshock, Yael P P. Mosse, Tara Naylor, Christian J J. Stoeckert, Barbara L L. Weber, John M M. Maris, and Gregory R R. Grant. Stac: A method for testing the significance of dna copy number aberrations across multiple array-cgh experiments. *Genome Res*, page in press, August 2006.
- [DRO⁺04] R. S. Daruwala, A. Rudra, H. Ostrer, R. Lucito, M. Wigler, and B. Mishra. A versatile statistical analysis algorithm to detect genome copy number variation. *Proc Natl Acad Sci U S A*, 101(46):16292–16297, November 2004.

- [EMLB06] D.A. Engler, G. Mohaptra, D.N. Louis, and R. Betensky. A pseudo-likelihood approach for simultaneous analysis of array comparative genomic hybridizations. *Biostatistics*, 7(3):399–421, 2006.
- [FMM⁺00] F. Forozan, E. H. Mahlamki, O. Monni, Y. Chen, R. Veldman, Y. Jiang, G. C. Gooden, S. P. Ethier, A. Kallioniemi, and O. P. Kallioniemi. Comparative genomic hybridization analysis of 38 breast cancer cell lines: a basis for interpreting complementary dna microarray data. *Cancer Res*, 60(16):4519–4525, August 2000.
- [FSPA04] Jane Fridlyand, Antoine M. Snijders, Dan Pinkel, and Donna G. and Albertson. Hidden markov models approach to the analysis of array cgh data. *Journal of Multivariate Analysis*, 90(1):132–153, July 2004.
- [GCSR03] Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis, Second Edition*. Chapman & Hall/CRC, 2003.
- [GLN06] S. Guha, Y. Li, and D. Neuberg. Bayesian hidden markov modeling of array cgh data. *Harvard University Biostatistics Working Paper Series*, 24, 2006.
- [Gre95] P. Green. Reversible jump markov chain monte carlo computation and bayesian model determination, 1995.
- [HBC⁺00] M. A. Heiskanen, M. L. Bittner, Y. Chen, J. Khan, K. E. Adler, J. M. Trent, and P. S. Meltzer. Detection of gene amplification by genomic hybridization to cDNA microarrays. *Cancer Res*, 60(4):799–802, February 2000.
- [HGC99] P. Hughes, J., P. Guttorp, and P. Charles, S. A nonhomogeneous hidden markov model for precipitation. Technical report, NRCSE, September 1999.
- [HKS⁺04] K. Holzmann, H. Kohlhammer, C. Schwaenen, S. Wessendorf, H. A. Kestler, A. Schwoerer, B. Rau, B. Radlwimmer, H. Dhner, P. Lichter, T. Gress, and M. Bentz. Genomic dna-chip hybridization reveals a higher incidence of genomic amplifications in pancreatic cancer than conventional comparative genomic hybridization and leads to the identification of novel candidate genes. *Cancer Res*, 64(13):4428–4433, July 2004.
- [HMRV99] J.A Hoeting, H. Madigan, A.E. Raftery, and C.T Volinsky. Bayesian model averaging: a tutorial. *Statistical Science*, 14(4):382–417, 1999.
- [HSG⁺05] L. Hsu, S. G. Self, D. Grove, T. Randolph, K. Wang, J. J. Dellow, L. Loo, and P. Porter. Denoising array-based comparative

- genomic hybridization data using wavelets. *Biostatistics*, 6(2):211–226, April 2005.
- [HST⁺04] P. Hupé, N. Stransky, J. P. Thiery, F. Radvanyi, and E. Barillot. Analysis of array cgh data: from signal ratio to gain and loss of dna regions. *Bioinformatics*, 20(18):3413–3422, December 2004.
- [HWLZ05] Tao Huang, Baolin Wu, Paul Lizardi, and Hongyu Zhao. Detection of dna copy number alterations using penalized least squares regression. *Bioinformatics*, September 2005.
- [LBL⁺05] O. C. Lingjaerde, L. O. Baumbusch, K. Liestl, I. K. Glad, and A. L. Borresen-Dale. Cgh-explorer: a program for analysis of array-cgh data. *Bioinformatics*, 21(6):821–822, March 2005.
- [LCCL06] William W. Lockwood, Raj Chari, Bryan Chi, and Wan L. and Lam. Recent advances in array comparative genomic hybridization technologies and their applications in human genetics. *European Journal of Human Genetics*, 14(current):139–148, 2006.
- [LJKP05] Weil R. R. Lai, Mark D. D. Johnson, Raju Kucherlapati, and Peter J. J. Park. Comparative analysis of algorithms for identifying amplifications and deletions in array cgh data. *Bioinformatics*, 21:3763–3770, 2005.
- [MP00] Geoffrey McLachlan and David Peel. *Finite Mixture Models*. Wiley-Interscience, October 2000.
- [MPN⁺05a] A. Misra, M. Pellarin, J. Nigro, I. Smirnov, D. Moore, K. R. Lamborn, D. Pinkel, D. G. Albertson, and B. G. Feuerstein. Array comparative genomic hybridization identifies genetic subgroups in grade 4 human astrocytoma. *Clin Cancer Res*, 11(8):2907–2918, April 2005.
- [MPN⁺05b] A. Misra, M. Pellarin, J. Nigro, I. Smirnov, D. Moore, K. R. Lamborn, D. Pinkel, D. G. Albertson, and B. G. Feuerstein. Array comparative genomic hybridization identifies genetic subgroups in grade 4 human astrocytoma. *Clin Cancer Res*, 11(8):2907–2918, April 2005.
- [MTT06] J. C. Marioni, N. P. Thorne, and S. Tavaré. Biohmm: a heterogeneous hidden markov model for segmenting array cgh data. *Bioinformatics*, 22(9):1144–1146, May 2006.
- [OVLW04] A. B. Olshen, E. S. Venkatraman, R. Lucito, and M. Wigler. Circular binary segmentation for the analysis of array-based dna copy number data. *Biostatistics*, 5(4):557–572, October 2004.

- [PA05] D. Pinkel and D. G. Albertson. Array comparative genomic hybridization and its applications in cancer. *Nat Genet*, 37 Suppl:S11–S17, June 2005.
- [PRL⁺05] F. Picard, S. Robin, M. Lavielle, C. Vaisse, and J. J. Daudin. A statistical approach for array cgh data analysis. *BMC Bioinformatics*, 6:27, 2005.
- [PRM⁺05] T. S. Price, R. Regan, R. Mott, A. Hedman, B. Honey, R. J. Daniels, L. Smith, A. Greenfield, A. Tiganescu, V. Buckle, N. Ventress, H. Ayyub, A. Salhan, S. Pedraza-Diaz, J. Broxholme, J. Ragoussis, D. R. Higgs, J. Flint, and S. J. Knight. Sw-array: a dynamic programming solution for the identification of copy-number changes in genomic dna using array comparative genome hybridization data. *Nucleic Acids Res*, 33(11):3455–3464, 2005.
- [PSP⁺02] J. R. Pollack, T. Srлие, C. M. Perou, C. A. Rees, S. S. Jeffrey, P. E. Lonning, R. Tibshirani, D. Botstein, A. L. Brresen-Dale, and P. O. Brown. Microarray analysis reveals a major direct role of dna copy number alteration in the transcriptional program of human breast tumors. *Proc Natl Acad Sci U S A*, 99(20):12963–12968, October 2002.
- [R D06] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2006. ISBN 3-900051-07-0.
- [Rab90] Lawrence R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1990.
- [RG97] S. Richardson and P. J. Green. On bayesian analysis of mixtures with an unknown number of components. *Journal of the of the Royal Statistical Society Series B (Statistical Methodology)*, 59:731 – 792, 1997.
- [RRT00] C. Robert, T. Ryden, and D. Titterington. Bayesian inference in hidden markov models through reversible jump markov chain monte carlo. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(1):57–75, 2000.
- [RSH⁺06] C Rouveirol, N Stransky, Ph Hup, Ph La Rosa, E Viara, E Barillot, and F Radvanyi. Computation of recurrent minimal genomic alterations from array-cgh data. *Bioinformatics*, 22:2066–2073, January 2006.
- [SLT⁺04] Jonathan Sebat, B. Lakshmi, Jennifer Troge, Joan Alexander, Janet Young, Par Lundin, Susanne Maner, Hillary Massa, Megan Walker, Maoyen Chi, Nicholas Navin, Robert Lucito, John Healy,

- James Hicks, Kenny Ye, Andrew Reiner, Conrad C. Gilliam, Barbara Trask, Nick Patterson, Anders Zetterberg, and Michael Wigler. Large-scale copy number polymorphism in the human genome. *Science*, 305(5683):525–528, July 2004.
- [SNS⁺01] A. M. Snijders, N. Nowak, R. Seagraves, S. Blackwood, 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. Assembly of microarrays for genome-wide measurement of dna copy number. *Nat Genet*, 29(3):263–264, 2001.
- [Ste00] M. Stephens. Dealing with label switching in mixture models. *Journal of the of the Royal Statistical Society Series B (Statistical Methodology)*, 62(4):795–809, 2000.
- [SXD⁺06] S. P. Shah, X. Xuan, R. J. Deleeuw, M. Khojasteh, W. L. Lam, R. Ng, and K. P. Murphy. Integrating copy number polymorphisms into array cgh analysis using a robust hmm. *Bioinformatics*, 22(14):e431–e439, July 2006.
- [UKS⁺06] A. E. Urban, J. O. Korbel, R. Selzer, T. Richmond, A. Hacker, G. V. Popescu, J. F. Cubells, R. Green, B. S. Emanuel, M. B. Gerstein, S. M. Weissman, and M. Snyder. High-resolution mapping of dna copy alterations in human chromosome 22 using high-density tiling oligonucleotide arrays. *Proc Natl Acad Sci U S A*, 103(12):4534–4539, March 2006.
- [VFP⁺03] J. A. Veltman, J. Fridlyand, S. Pejavar, A. B. Olshen, J. E. Korkola, S. DeVries, P. Carroll, W. L. Kuo, D. Pinkel, D. Albertson, C. Cordon-Cardo, A. N. Jain, and F. M. Waldman. Array-based comparative genomic hybridization for genome-wide screening of dna copy number in bladder tumors. *Cancer Res*, 63(11):2872–2880, June 2003.
- [VK04] B. Vogelstein and K. W. Kinzler. Cancer genes and the pathways they control. *Nat Med*, 10(8):789–799, August 2004.
- [WF05] Hanni Willenbrock and Jane Fridlyand. A comparison study: applying segmentation to array cgh data for downstream analyses. *Bioinformatics*, 21:4084–4091, September 2005.

Table 1: True vs. estimated ploidy. True ploidy obtained from [SNS⁺01]; estimated ploidy from Bayesian Model Averaging, assigning every gene to the “loss”, “normal”, “gain” state with largest posterior probability. In parenthesis, under the line name, the correct classification rate.

Cell line	True ploidy	Loss	Normal	Gain
gm01524 (0.998)	Monosomy	0	0	0
	Normal	0	2244	2
	Trisomy	0	2	23
gm01750 (0.999)	Monosomy	0	0	0
	Normal	0	2233	2
	Trisomy	0	1	35
gm01535 (0.972)	Monosomy	0	1	0
	Normal	0	2200	52
	Trisomy	0	11	7
gm03134 (0.997)	Monosomy	13	3	0
	Normal	4	2251	0
	Trisomy	0	0	0
gm05296 (0.971)	Monosomy	15	0	0
	Normal	7	2151	53
	Trisomy	0	6	39
gm03563 (0.993)	Monosomy	2	0	0
	Normal	0	2208	11
	Trisomy	0	5	45
gm07081 (0.992)	Monosomy	0	0	0
	Normal	0	2185	9
	Trisomy	0	10	67
gm13031 (0.999)	Monosomy	9	0	0
	Normal	2	2260	0
	Trisomy	0	0	0
gm13330 (0.997)	Monosomy	17	2	0
	Normal	0	2198	3
	Trisomy	0	1	50

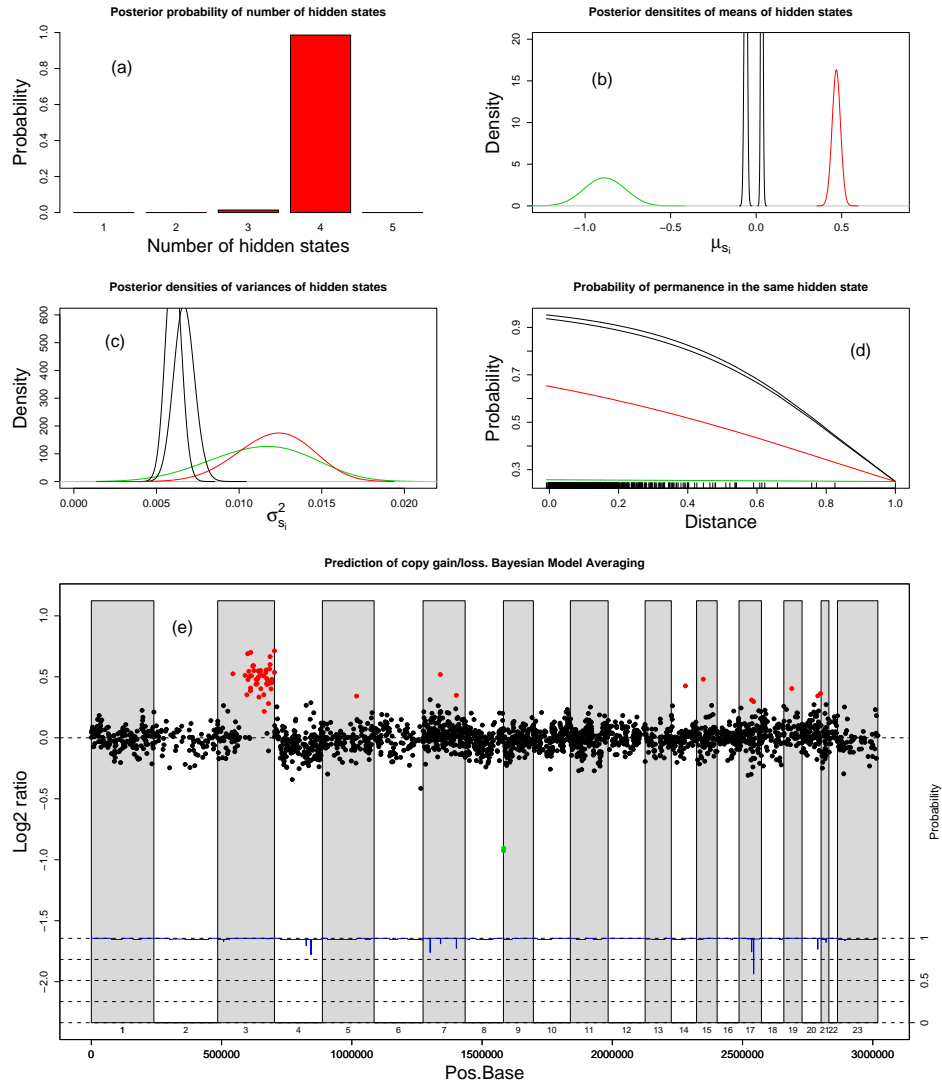


Figure 1: Results of the RJACGH analysis of gm03563 cell line from Snijders. Results shown are from four parallel chains (10000 burn-in iterations, 10000 draws from the posterior); see text for details about other parameters. The lower panel shows the results from the Bayesian Model Averaging step (see text, eq. [10]); black dots correspond to genes classified as 'normal' or non-changed, red dots to genes classified as 'gained' and green dots to genes classified as 'losses'; the lower blue line shows the posterior probability for every gene of belonging to the predicted state. The vertical alternating white and grey bars denote the different chromosomes with the chromosome number shown at bottom.