BIOINFORMATICS ORIGINAL PAPER Vol. 24 no. 6 2008, pages 751–758
doi:10.1093/bioinformatics/btn003

Genome analysis

A segmental maximum a posteriori approach to genome-wide copy number profiling

Robin Andersson[1], Carl E. G. Bruder[2], Arkadiusz Piotrowski[2], Uwe Menzel[3], Helena Nord[3], Johanna Sandgren[4], Torgeir R. Hvidsten[1], Teresita Diaz de Ståhl[3], Jan P. Dumanski[2,3] and Jan Komorowski[1,5,*]

[1]The Linnaeus Centre for Bioinformatics, Uppsala University, 751 24 Uppsala, Sweden, [2]Department of Genetics, University of Alabama at Birmingham, Birmingham AL 35294-0024, USA, [3]Department of Genetics and Pathology, Rudbeck Laboratory, Uppsala University, [4]Department of Surgical Sciences, Uppsala University Hospital, 751 85 Uppsala, Sweden and [5]Interdisciplinary Center for Mathematical and Computational Modelling, Warsaw University, 02-106 Warsaw, Poland

Received and revised on December 19, 2007; accepted on January 2, 2008

Advance Access publication January 19, 2008

Associate Editor: Alex Bateman

ABSTRACT

Motivation: Copy number profiling methods aim at assigning DNA copy numbers to chromosomal regions using measurements from microarray-based comparative genomic hybridizations. Among the proposed methods to this end, Hidden Markov Model (HMM)-based approaches seem promising since DNA copy number transitions are naturally captured in the model. Current discrete-index HMM-based approaches do not, however, take into account heterogeneous information regarding the genomic overlap between clones. Moreover, the majority of existing methods are restricted to chromosome-wise analysis.

Results: We introduce a novel Segmental Maximum A Posteriori approach, SMAP, for DNA copy number profiling. Our method is based on discrete-index Hidden Markov Modeling and incorporates genomic distance and overlap between clones. We exploit a priori information through user-controllable parameterization that enables the identification of copy number deviations of various lengths and amplitudes. The model parameters may be inferred at a genome-wide scale to avoid overfitting of model parameters often resulting from chromosome-wise model inference. We report superior performances of SMAP on synthetic data when compared with two recent methods. When applied on our new experimental data, SMAP readily recognizes already known genetic aberrations including both large-scale regions with aberrant DNA copy number and changes affecting only single features on the array. We highlight the differences between the prediction of SMAP and the compared methods and show that SMAP accurately determines copy number changes and benefits from overlap consideration.

Availability: SMAP is available from Bioconductor and within the Linnaeus Centre for Bioinformatics Data Warehouse.

Contact: Jan.Komorowski@lcb.uu.se

Supplementary information: Supplementary data are available at http://www.lcb.uu.se/papers/r_andersson/SMAP/

*To whom correspondence should be addressed.

1 INTRODUCTION

The study of human genetic variation at the level of nucleotide sequence changes constitutes a major challenge and has, therefore, received considerable attention in the genomic era. The primary type of variation explored so far has been at the level of single nucleotide polymorphisms (SNPs). Larger variations at the level of gains and deletions, also called copy number variation (CNV), have received less attention. The genome-wide detection of CNVs has been difficult due to the lack of high-resolution and high-throughput techniques.

A fundamental step towards identifying such variation has been the development of microarray-based comparative genomic hybridization (array-CGH) (Mantripragada *et al.*, 2004; Pinkel *et al.*, 1998; Solinas-Toldo *et al.*, 1997). Recently, two landmark studies have reported the presence of CNVs in the human genome using different approaches (Iafrate *et al.*, 2004; Sebat *et al.*, 2004). Both studies convincingly demonstrate the presence in normal individuals of genomic imbalances that overlap with genes and segmental duplications and may contribute to phenotypic variation and disease susceptibility. These initial findings have now been followed by a number of additional reports that further strengthen the evidence for the importance of CNVs (Redon *et al.*, 2006). The identification of DNA copy number alterations is also very important in studies of cancer, indicating that failures in the mechanisms that maintain the integrity of the genome contribute to tumor initiation/progression. Structural rearrangements (translocations, inversions) or gains may cause activation of oncogenes, whereas deletions may underlie haploinsufficiency or inactivate tumor suppressor genes. All these aberrations may also influence the expression of so-called phenotype modifier genes. Although not critical for tumor initiation as such, these genes may greatly change the clinical picture and outcome of a disease. Discovery and functional assessment of genomic regions affected by copy number alterations are thus essential for understanding the biology of cancer and for diagnostic applications.

In a typical array-CGH experiment, total genomic DNA from test and reference samples are labeled differently and hybridized to a microarray. The intensity ratio between the test and

reference signal for each spot on the microarray is, theoretically, proportional to the relative copy number of the corresponding genomic sequence. Recently developed commercial and custom-made genomic microarrays enable copy number analysis at a very high resolution, with several hundred thousand measurement points. As a consequence of the large amount of data generated from such experiments, the use of automatic procedures for the assignment of copy number profiles to test DNA, i.e. *copy number profiling*, has become an essential step in the analysis of array-CGH data. Subsequently, several methods have recently been proposed. A summary and comparison of some early methods is provided by Lai *et al*. (2005).

The majority of methods assume a model with Gaussian distributions for which the means and, in some cases, the variance change at unknown breakpoints. The task of finding such breakpoints is often referred to as *segmentation* (Willenbrock and Fridlyand, 2005). The common approach shared by these methods is to identify breakpoints in a manner that maximizes the *likelihood* probability distribution function (pdf) (Hupe *et al*., 2004; Myers *et al*., 2004; Olshen *et al*., 2004; van de Wiel *et al*., 2007). The number of breakpoints may be controlled by a penalty that is extracted from the likelihood and increases with the number of breakpoints. A summary of these methods is provided by Picard *et al*. (2005). Non-likelihood-based approaches include, e.g. smoothing methods (Eilers and de Menezes, 2005; Hsu *et al*., 2005; Tibshirani and Wang, 2007) and a clustering-based approach (Wang *et al*., 2005). Finally, Hidden Markov Models (HMMs) have been used by Engler *et al*. (2006), Fridlyand *et al*. (2004), Marioni *et al*. (2006), Rueda and Díaz-Uriarte (2007), Shah *et al*. (2006) and Stjernqvist *et al*. (2007).

HMM-based approaches seem promising since copy number changes between DNA segments are naturally modeled by transition events between hidden states in the model. Fridlyand *et al*. (2004) proposed a discrete-index HMM approach in which the optimal segmentation of clones is found by likelihood maximization using a derived number of Gaussian distributions with state-specific means and fixed variance. The number of states is selected using a penalty, based on the AIC or BIC criterion, and for each model the means are estimated using *partitioning among medoids*. The HMM parameters are reestimated using a backward–forward algorithm and the optimal state sequence, i.e. a path in the HMM, is reconstructed using the Viterbi algorithm. No information regarding the genomic distance or overlap between clones is taken into account in that method. An extension of this method was proposed by Marioni *et al*. (2006) in which the state transition probabilities depend on the distance, defined as the difference in midpoints, between genomically adjacent clones. Stjernqvist *et al*. (2007) proposed a continuous-index HMM in which state changes are determined by transition rates and may occur at any base pair. The number of states in their model is determined as in the discrete-index HMMs. The HMM parameters are estimated using a Monte Carlo EM (MCEM) approach and the realizations of the Markov chain are generated using a number of Markov Chain Monte Carlo (MCMC) simulations. Since both the MCEM and MCMC approaches are non-deterministic by nature, the predicted copy number breakpoints may differ between runs. Although Stjernqvist *et al*. (2007) may produce the finest resolution breakpoints, their computation times are infeasible. For instance, as reported by Stjernqvist *et al*. (2007), analyzing a single chromosome took about 25 CPU hours with a four-state model. Extrapolating this to a normal array-CGH project of, for example, 100 experiments with 24 chromosomes each yields an expected execution time of 60 000 CPU hours, i.e. 2500 days or 6.8 years.

The HMM-based methods described above infer the number of hidden states through model selection and perform copy number profiling/segmentation separately for each chromosome. Such approaches may easily overfit the model parameters to local effects in the chromosomes. Interpretation of results becomes questionable in cases in which inferred means and variances of the Gaussian distributions associated with a certain state differ between chromosomes. In some situations, however, one might prefer chromosome-wise models over genome-wide ones. Segmentation methods with chromosome-wise models are appropriate to detect relative copy number alterations between loci or mosaicism in the same chromosome when the actual copy number is not of interest (Rueda and Díaz-Uriarte, 2007).

A number of discrete-index HMM-based methods with genome-wide parameter estimation has been proposed to avoid overfitting the HMM parameters to chromosomal characteristics. Shah *et al*. (2006) proposed a four-state HMM in which the parameters are estimated by pooling across samples using *block Gibbs sampling*. Engler *et al*. (2006) suggested a three-state Gaussian mixture HMM in which the HMM parameters are not only considered common across chromosomes but also across samples. A slightly different approach was proposed by Rueda and Díaz-Uriarte (2007) who fitted a non-homogeneous HMM via a large number of reversible jump MCMC iterations. In contrast to the other HMM-based methods, the number of states is not explicitly set or selected using a penalty but inferred through *Bayesian model averaging*. Rueda and Díaz-Uriarte (2007) further proposed transition probabilities that converge towards equality at the (within-array) maximum interprobe distance.

2 APPROACH

In this article we propose a novel HMM-based statistical method for copy number profiling, called segmental maximum a posteriori (SMAP). In Section 3.1 we describe a heterogeneous discrete-index HMM for copy number assignments that takes into account information about genomic position and overlap between clones. The fundamentals of our method for copy number profiling based on segmental a posteriori maximization are given in Section 3.2. By adapting a maximum a posteriori (MAP) approach, we enable the incorporation of user-controllable a priori (prior) information in the profiling process. For instance, intensity ratios often deviate from expected values due to normal cell admixture in the test DNA or insufficient blocking of repetitive elements that are present within different clones on the array during hybridization experiments. Prior knowledge about such variations is thus important and can, therefore, be supplied to our method. Noise in the data may also affect the assignment of clones to copy numbers. Assuming that such errors are Gaussian distributed, we are able to model the noise (using a Gaussian distribution) for each considered copy number. Using prior knowledge about distributions for the considered copy number levels and a given set of parameters for the HMM as a start solution, the most probable copy number assignments are inferred by alternating

optimization of the copy number assignments and genome-wide optimization of the parameters.

In Section 4 we report the results of genomic profiling on both synthetic data and on a set of glioblastoma multiforme (GBM) samples and compare the predictions of SMAP with the predictions of DNACopy (Venkatraman and Olshen, 2007) and BioHMM (Marioni *et al.*, 2006).

3 METHODS

Copy number profiling aims at classifying clones to discrete copy number classes based on their intensity ratios and chromosomal positions. For a given sequence, of length T, of (genomically ordered) clones and the corresponding sequence of observed intensity ratios $O=\{o_1,\ldots,o_T\}$, start positions $Sp=\{sp_1,\ldots,sp_T\}$, end positions $Ep=\{ep_1,\ldots,ep_T\}$, and chromosome identifiers $Ch=\{ch_1,\ldots,ch_T\}$, such that $ch_t\in\{1,\ldots,C\}$ $(1\le t\le T)$, we wish to infer the most plausible sequence of copy number assignments $Q=\{q_1,\ldots,q_T\}$.

3.1 An HMM for copy number assignments

We model the copy number assignments using an HMM (Rabiner, 1990). An HMM, in our context, is a pair $\mathcal{H}=(S,\lambda)$, where $S=\{s_i\}_{i=1}^N$ is a set of N copy number states, such that $q_t\in S$ $(1\le t\le T)$, and $\lambda=(\Pi,A,\Omega)$ are parameters for the model. The $s_i\in S$ $(1\le i\le N)$ are called copy number states since the fundamental task is to associate clones with copy numbers. We advocate the use of a biologically motivated six-state model, proposed by van de Wiel *et al.* (2007), rather than the conventional three-state model, since these states will capture double deletion, single deletion, normal, gain, double gain and amplification. However, the model is not restricted to six states; a different number of states may be used if desired by the user. Using an HMM, the sequence of clones is traversed in one direction only, according to chromosomal position (chromosome and start position) of clones, simultaneously moving between the hidden copy number states in the model. The probability of starting in copy number state s_i $(1\le i\le N)$ for clone one is specified by the *initial probabilities*, $\Pi=\{\pi_i\}_{i=1}^N$. Each pair of copy number states is connected by HMM specific *transition probabilities* $A=\{a_{ij}\}_{i,j=1}^N$ that specify the probabilities of transition between states s_i and s_j $(1\le i,j\le N)$ between any two consecutive clones in the sequence. Inspired by the ideas of Marioni *et al.* (2006), Colella *et al.* (2007) and Rueda and Díaz-Uriarte (2007), we incorporate *distance-based transition probabilities*, $A^d=\{a_{ij}^d\}_{i,j=1}^N$, in the model that takes into account the genomic distance d between any two consecutive clones t and $t+1$ $(1\le t<T)$,

$$a_{ij}^d \triangleq a_{ij}-\rho\cdot\left(a_{ij}-\frac{1}{N}\right), \qquad (1)$$

$$\rho=\begin{cases}1-\exp\left(-\frac{d}{L}\right) & \text{if } d>0\\ 0 & \text{otherwise}\end{cases}, \quad d=s_{t+1}-e_t.$$

In contrast to Rueda and Díaz-Uriarte (2007), we do not assume convergence of transition probabilities at the (within-array) maximum inter-probe distance since such a value will depend solely on the array platform used. Rather, a length parameter L (specified in base pairs) is used to control the convergence of transition probabilities towards $1/N$. The following constraints restrict the HMM; $\forall i,j\in\{1,\ldots,N\}$: $\sum_{j=1}^N a_{ij}=1$, $\sum_{i=1}^N \pi_i=1$ and $0\le a_{ij},\pi_i\le 1$. It follows from Equation (1) that $\sum_{j=1}^N a_{ij}^d=1$ $(\forall i\in\{1,\ldots,N\})$ is guaranteed for any $d\in\mathcal{Z}$.

The probability of observing a specific intensity ratio for clone t $(1\le t\le T)$ given that the HMM is in state s_j, called the *emission probability*, is naturally defined to be Gaussian distributed, i.e

$$b_{s_j}(o_t)\triangleq p(o_t|q_t=s_j,\Omega),\ \text{such that} \qquad (2)$$
$$o_t|q_t=s_j,\Omega\ \sim\ \mathcal{N}(o_t|\omega_j),$$

where $\Omega=\{\omega_1=(\mu_1,\sigma_1),\ldots,\omega_N=(\mu_N,\sigma_N)\}$ are parameters, i.e. means and SDs, for Gaussian distributions associated with each state. Note that Equation (2) assumes independency between clones because the emission probability only depends on state s_j and observation o_t. This may not hold when clones on the array overlap in terms of genomic position (Stjernqvist *et al.*, 2007). The overlap of clones may introduce dependency in terms of intensity ratio between clones, e.g. aberration breakpoints may occur within a clone that partially overlaps with another, and longer clones may completely encompass shorter ones.

To deal with such dependency, we propose an extension of Equation (2) that incorporates knowledge of previous copy number assignments, i.e. assignments done before the current clone in the sequence for the clones that overlap with the current. The emission pdf is extended rather than the transition probabilities to enable the incorporation of overlap information from multiple clones rather than just the preceding clone in the observation sequence. The fraction of overlap between, for instance, the clones r and t, $o(r,t)$, is calculated as

$$o(r,t)=\begin{cases}\max\left(\frac{\min(ep_r,ep_t)-\max(sp_r,sp_t)}{ep_t-sp_t},0\right) & \text{if } ch_r=ch_t\\ 0 & \text{otherwise}\end{cases}. \qquad (3)$$

Let $olap_t$ denote the set of previous clones that overlaps with clone t and Q_{olap_t} the corresponding set of copy number states. Formally, $olap_t\triangleq\{1\le r\le t-1: o(r,t)>0\}$ and $Q_{olap_t}\triangleq\{q_r: r\in olap_t\}$. We define a new emission probability, $b'_{s_j}(o_t)$ $(1\le j\le N)$ $(1\le t\le T)$, distributed as a mixture of Gaussian distributions,

$$b'_{s_j}(o_t)\triangleq p(o_t|Q_{olap_t},q_t=s_j,\Omega)=\alpha\, b_{s_j}(o_t)+\sum_{r\in olap_t}\beta_r\, b_{q_r}(o_t) \qquad (4)$$

where α and β_r are mixing proportions that weights the influence of the states according to overlap of clones, i.e.

$$\alpha=\left(1+\sum_{r\in olap_t}o(r,t)\right)^{-1}\ \text{and}\quad \beta_r=\alpha\, o(r,t)\,. \qquad (5)$$

If no clone is overlapping with the current one then $b'_{s_j}(o_t)=b_{s_j}(o_t)$.

Note that we do not $\log_2$-transform the intensity ratios, since such a transformation does not preserve the difference between theoretical values.

3.2 Copy number profiling by segmental a posteriori maximization

Given the observed intensity ratios $O=\{o_t\}_{t=1}^T$ and chromosomal positions $P=\{Sp, Ep, Ch\}$ for a sequence of clones covering the whole or part of a genome, we can use the HMM to infer statistically the single best sequence of copy number assignments from multiple paths in the HMM; one per chromosome, $Q=\{q_t\}_{t=1}^T=\{Q_{Ch_1},\ldots,Q_{Ch_C}\}$, where $Ch_c=\{r\in\{1,\ldots,T\}: ch_r=c\}$ and each $q_t\in S$. It may sometimes be desirable to train the HMM chromosome-wise, as discussed in Section 1. In such a case, one HMM is used per chromosome. Below, we consider the genome-wide approach.

The probability of a certain sequence of copy number assignments given a sequence of observations, chromosomal positions and known parameters λ is given by the a posteriori (posterior) pdf:

$$p(Q|O,P,\lambda)=\frac{p(Q|P,\lambda)p(O|Q,P,\lambda)}{p(O|P,\lambda)}, \qquad (6)$$

where $p(Q|P,\lambda)$ is called the a priori (prior) pdf of the copy number states, $p(O|Q,P,\lambda)$ is called the *likelihood* of the observed intensity ratios and $p(O|P,\lambda)$ is a normalization constant. Hence, the most probable sequence of copy number assignments Q^* can be inferred by posterior maximization, i.e. $Q^*=\text{argmax}_Q[p(Q,O|P,\lambda)]$.

In our case λ is unknown so, inspired by the approach of Gauvain and Lee (1992), we choose to find the most plausible state sequence Q^* by maximizing the joint posterior probability of Q and λ given O and P, i.e.

$$Q^*=\underset{Q}{\text{argmax}}\ \max_{\lambda}[\,p(Q,\lambda|O,P)\,]$$
$$=\underset{Q}{\text{argmax}}\ \max_{\lambda}[\,p(Q,O|P,\lambda)p(\lambda)\,], \qquad (7)$$

where Q^* is called the *segmental MAP estimate* of Q. It can be shown that, starting with any parameter estimate $\lambda^{(m)}$, alternate maximization over Q and λ gives a sequence of estimates with non-decreasing values of $p\,(Q, \lambda|O, P)$ (Gauvain and Lee, 1992), i.e.

$$p(Q^{(m+1)}, \lambda^{(m+1)}|O, P) \geq p(Q^{(m)}, \lambda^{(m)}|O, P), \tag{8}$$

with

$$Q^{(m)} = \underset{Q}{\operatorname{argmax}}\big[\, p(Q, O|P, \lambda^{(m)}) \,\big] \text{ and} \tag{9}$$

$$\lambda^{(m+1)} = \underset{\lambda}{\operatorname{argmax}}\big[\, p(Q^{(m)}, O|P, \lambda)p(\lambda) \,\big]. \tag{10}$$

$p(\lambda)$ in Equation (10) denotes the prior pdf for the parameters λ and is defined as

$$p(\lambda) = p(\Pi)p(A)p(\Omega) = \prod_{i,j=1}^{N} p(a_{ij}) \prod_{i=1}^{N} p(\pi_i)p(\mu_i)p(\sigma_i). \tag{11}$$

The choice of priors for the parameters is important since it will influence the maximization of Equation (10). We control the variance of the state distributions by assigning a variant of Jeffrey's prior (Jaynes, 2003) to σ_i that controls the minimal allowed SD, i.e. $\sigma_i \sim \frac{\sigma_{min}}{\sigma_i}$ $(1 \leq i \leq N)$. The ability to adjust the mean of the state distributions to data is controlled by Gaussian priors for the means centered around expected copy number ratios, i.e. $\mu_i \sim \mathcal{N}(\mu_i^{\text{expected}}, \sigma_\mu)$, where μ_i^{expected} and σ_μ are parameters specified by the user. For instance, a normal cell contamination of the test sample of 40% yields an expected ratio of $(1\times0.6 + 2\times0.4)\,/\,2 = 0.7$ for heterozygous deletions (assuming reference sample is diploid). We use uniform Dirichlet priors for the transition and initial probabilities, i.e. π, $a_i \sim \mathcal{D}ir(\delta)$ $(1 \leq i \leq N)$, where $[\delta_j]_{j=1}^{N} = 1$.

Based on the previous reasoning Equations (9,10) and Gauvain and Lee (1992), we propose a SMAP algorithm for copy number profiling (see Fig. 1). Recall that the (sub) optimal copy number assignments is a combination of multiple, chromosome-wise, paths in the HMM, i.e. at iteration m we have $Q^{(m)} = \{Q_{Ch_1}^{(m)}, \ldots, Q_{Ch_C}^{(m)}\}$. Maximization of Equation (9) is done by chromosome-wise runs of a modified Viterbi algorithm (Viterbi, 1967) in which the characteristics of our model, in terms of genomic distance [Equation (1)] and overlap [Equation (4)] between clones, is taken into account. Finding the parameters λ that maximize Equation (10) is, however, harder. There does not exist any known algorithm that guarantees finding the global optimum of Equation (10). In SMAP, we choose to locally maximize Equation (10) by a gradient-based method.

The gradient descent method is a simple scheme that can be used to find the parameters λ that maximize Equation (10). We seek to minimize the function

$$f(\lambda) = -log\,[p(Q|P, \lambda)p(O|Q, P, \lambda)p(\lambda)] \tag{12}$$

by iteratively updating any parameter $\lambda_j \in \lambda$ according to the formula

$$\lambda_j^{(k+1)} = \lambda_j^{(k)} - \eta_j^{(k)} \left[\frac{\partial f}{\partial \lambda_j}\right]_{\lambda=\lambda^{(k)}} \tag{13}$$

where $\eta_j^{(k)}$ is a positive step size parameter called the *learning rate*. Formula (13) guarantees that $f(\lambda^{(k)}) \leq f(\lambda^{(k-1)})$ at any iteration k. To speed up the computation time of Equation (10) we use an individual learning rate adaptation scheme (Bagos *et al.*, 2004, ALGORITHM 1), where the individual learning rates are updated with respect to the change of partial derivatives, i.e.

$$\eta_j^{(k)} = \min\left(\eta_j^{(k-1)} a^+, \eta_{max}\right), \text{ if}$$

$$\left[\frac{\partial f}{\partial \lambda_j}\right]_{\lambda=\lambda^{(k)}} \cdot \left[\frac{\partial f}{\partial \lambda_j}\right]_{\lambda=\lambda^{(k-1)}} > 0 \text{ , and}$$

$$\eta_j^{(k)} = \max\left(\eta_j^{(k-1)} a^-, \eta_{min}\right) \text{ otherwise.}$$

In the latter case, the partial derivative is set to zero in the current iteration for smooth adaptation. a^- and a^+ control the change of learning rate; a value of 1 for a^- and a^+ means standard gradient descent. η_{min} and η_{max} control the minimum and maximum allowed learning rates, respectively. In the following analyses, we set a^-, a^+, η_{min} and η_{max} to 0.5, 1.2, 10^{-4} and 0.5, respectively, as recommended by Bagos *et al.* (2004).

During gradient descent we restrict the ability to adjust the variances of the state distributions by a weighting of the corresponding partial derivatives. The weights, γ_i $(1 \leq i \leq N)$, coupled to each copy number state, $s_i \in S$, restrict the influence of each contained clone on the corresponding state distribution, i.e. $\gamma_i = (|\{t \in \{1, \ldots, T\} : q_t = s_i\}|)^{-1}$.

SMAP$(O, P, \lambda^{(1)}, \tau)$
1 $logP^{(0)} \leftarrow -\infty$
2 $m \leftarrow 1$
3 **repeat**
4 **for** $c \leftarrow 1$ **to** C
5 **do** $Q_{Ch_c}^{(m)} \leftarrow \underset{Q_{Ch_c}}{\operatorname{argmax}} \left[\log\ p\left(Q_{Ch_c}, O_{Ch_c}|P, \lambda^{(m)}\right)\right]$
6 $logP^{(m)} \leftarrow \log\ p\left(Q^{(m)}, O|P, \lambda^{(m)}\right)$
7 $\lambda^{(m+1)} \leftarrow$
8 $\underset{\lambda}{\operatorname{argmax}} \left[\log \prod_{c=1}^{C} p\left(Q_{Ch_c}^{(m)}, O_{Ch_c}|P, \lambda\right) p\left(\lambda\right)\right]$
9 $m \leftarrow m + 1$
10 **until** $logP^{(m)} - logP^{(m-1)} \leq \tau$
11 **return** $Q^{(m)}$

Fig. 1. The SMAP algorithm. The SMAP procedure takes as input a start solution of parameters and prior information. The process of joint posterior maximization is done by alternating optimization of the copy number assignments, with fixed parameters and optimization of the parameters, with fixed assignments, until no significant improvements can be made (i.e. improvement below a given threshold τ). During optimization, we consider logarithmic probabilities due to machine precision limitations.

4 RESULTS

In order to evaluate the predictive performance of SMAP, we have profiled a synthetic data set with overlapping and unevenly positioned clones (Section 4.2). Eight configurations of SMAP were tested with the following characteristics; with or without overlap consideration, standard or distance-based transition probabilities and chromosome-wise or genome-wide HMMs. Furthermore, we report the performance of SMAP (with overlap consideration, distance-based transition probabilities and genome-wide HMM) in an array-CGH study where test and reference samples are hybridized on a tiling 32K BAC array (array design and protocols for hybridization and scanning are described in Supplementary Material). The study (Section 4.3) considers genome-wide copy number profiling of a set of glioblastoma multiforme samples. In this data set ≈87% of the clones overlap. In both evaluation studies, we have compared the predictions of SMAP with the results of DNACopy (Venkatraman and Olshen, 2007) and BioHMM (Marioni *et al.*, 2006). Profile plots illustrating the results of the various configurations of SMAP and the other two methods on both data sets are available as Supplementary Material.

4.1 Comparison setup

DNACopy is a non-parametric method based on circular binary segmentation (CBS) that identifies breakpoints of copy number segments by successive splitting of segments. A pruning algorithm is subsequently applied to control the number of

Table 1. The overall median (med) and interquartile range (IQR) of sensitivity (sens) and specificity (spec) for deletion and gain for the compared methods on the synthetic data with overlap and at windows of 700 kb around each true breakpoint

Method	Overall								Window							
	Deletion				Gain				Deletion				Gain			
	sens		spec		sens		spec		sens		spec		sens		spec	
	med	IQR	med	IQR	med	IQR	med	IQR	med	IQR	med	IQR	med	IQR	med	IQR
SMAP (o, d, g)	0.93	0.13	1.00	0.0048	0.94	0.11	1.00	0.0049	0.84	0.28	0.98	0.036	0.87	0.21	0.97	0.032
SMAP (d, g)	0.89	0.20	1.00	0.0037	0.91	0.16	1.00	0.0034	0.74	0.40	0.98	0.036	0.75	0.35	0.98	0.033
SMAP (o, g)	0.93	0.16	1.00	0.0036	0.93	0.13	1.00	0.0035	0.82	0.30	0.98	0.031	0.84	0.26	0.98	0.031
SMAP (g)	0.87	0.29	1.00	0.0027	0.89	0.20	1.00	0.0026	0.69	0.48	0.98	0.027	0.71	0.40	0.98	0.024
SMAP (o, d, c)	0.92	0.17	1.00	0.008	0.93	0.13	0.99	0.009	0.84	0.26	0.98	0.033	0.85	0.23	0.97	0.035
SMAP (d, c)	0.90	0.22	1.00	0.0026	0.92	0.15	1.00	0.0026	0.76	0.36	0.98	0.026	0.81	0.30	0.98	0.025
SMAP (o, c)	0.92	0.17	1.00	0.0067	0.93	0.14	1.00	0.0066	0.84	0.26	0.98	0.033	0.85	0.24	0.97	0.035
SMAP (c)	0.90	0.23	1.00	0.0026	0.92	0.16	1.00	0.0028	0.76	0.37	0.98	0.025	0.80	0.30	0.98	0.025
DNACopy	0.89	0.19	1.00	0.0036	0.89	0.24	1.00	0.0033	0.63	0.32	0.96	0.057	0.63	0.34	0.94	0.059
BioHMM	0.89	0.19	0.99	0.016	0.89	0.22	1.00	0.0034	0.64	0.33	0.94	0.074	0.61	0.32	0.95	0.057

Results for BioHMM and DNACopy are compiled after merging segments by *mergeLevels*. For the SMAP methods, o, d, g, and c denote overlap consideration, distance-based transition probabilities, genome-wide HMM and chromosome-wise HMMs respectively.

identified breakpoints. This method is considered state-of-the-art and has been proven to perform well on both synthetic and real data (Willenbrock and Fridlyand, 2005). BioHMM is briefly described in Section 1.

In contrast to SMAP, both DNACopy and BioHMM are segmentation methods, i.e. the model states or their equivalent correspond to segment means rather than to copy numbers and the fundamental task of these methods is to find the optimal splitting of segments. Some approaches have been suggested to transform the segment means to copy numbers or aberration classes based on thresholds (e.g. Willenbrock and Fridlyand, 2005) or by constraining the means in the segmentation process (Chen *et al.*, 2006). The latter approach seems more robust, though the method is not yet available. Segmentation approaches also come with the problem of proposing segment means too close to each other thus requiring further post-processing. To deal with this issue we applied the *mergeLevels* function (Willenbrock and Fridlyand, 2005) on the predicted profiles of both *DNACopy* and *BioHMM* prior to comparison with SMAP.

DNACopy (version 1.10.0) and *BioHMM* (version 1.4.0) were run with default parameter settings on $\log_2$ transformed intensity ratios through the *snapCGH* (Smith *et al.*, 2007) R (R Development CoreTeam, 2007) package available in Bioconductor 2.0 (Gentleman *et al.*, 2004). The protocol for preprocessing and normalization, the parameter settings used for SMAP and the raw data are available as Supplementary Material. To guarantee a fair comparison, we used the same parameter settings (default settings) for SMAP in both studies.

4.2 Evaluation of aberration detection in synthetic data

The synthetic data of Willenbrock and Fridlyand (2005) contains simulated array-CGH hybridization measurements from 500 samples, each containing 20 chromosomes of 100 measurement points. The measurements are, however, considered independent since no overlap between clones exists in the data. Furthermore, the interclone distances are evenly distributed. To test the characteristics of SMAP, we have generated new synthetic data from the same protocol (Willenbrock and Fridlyand, 2005) although with some modifications. Clone lengths and interclone distances were sampled from the 32K BAC array and copy number breakpoints were sampled to specify specific base pair locations in the chromosomes rather than at specific clones. The intensity ratio for each clone was then determined from the overlap with simulated segments. In detail, the intensity ratio for clone t, overlapping with segments S with true copy numbers C was calculated as:

$$o_t = \frac{p \cdot \sum_{s \in S} C_s \cdot o(s,t) + (1-p) \cdot 2}{2} + \epsilon \,,$$

where p is the proportion of tumor cells in the test sample and $\epsilon \sim N(0, \sigma)$. p and σ were sampled according to Willenbrock and Fridlyand (2005).

Table 1 (overall) summarizes the estimated sensitivity and specificity of predicting gains and deletions in the synthetic data set. The normal (non-aberrant) level for BioHMM and DNACopy was determined in an interval-based manner as in Willenbrock and Fridlyand (2005). Deletions and gains were then defined as everything below or above the interval categorizing normal, respectively. Sensitivity for gain is the proportion of predicted gained segments among the truly gained segments, whereas specificity for gain is the proportion of truly predicted non-gained segments among the truly non-gained segments. The same definitions hold for deletions. The results indicate similar performances between the different SMAP configurations, although slightly better when the overlap-based emission pdf is used. Both DNACopy and BioHMM perform slightly worse.

To examine whether the predictive performance of SMAP around breakpoints will benefit from the overlap model, we calculated sensitivity and specificity for gain and deletions in 700 kb large windows (two times the maximum clone length) around the true breakpoints. The results are summarized in Table 1 (window). The overlap consideration of SMAP

contribute to the highest sensitivities although still having the highest specificities among the compared methods. The worst results in this configuration are delivered by DNACopy and BioHMM.

4.3 Copy number profiling of GBM

Gliomas are common and frequently malignant primary tumors of the central nervous system that arise from glial cell lineage. GBM is the most malignant type of gliomas (histological grade IV according to the World Health Organization). Patients diagnosed with GBM have a very poor prognosis and a median survival of <1 year. From a previous study (Diaz de Ståhl *et al.*, 2005) we selected three samples for genome-wide copy number profiling on the 32K array. These samples are part of a larger genome-wide array-CGH GBM study (T. Diaz de Ståhl *et al.*, manuscript in preparation).

During profiling with SMAP, we monitored the improvement of log $p(Q, \lambda|O, P)$ (Fig. 2). The results clearly show that a single Viterbi run on the data given the initial parameters would have yielded a lower log probability than that achieved using the SMAP approach with alternating optimization of copy number assignments and HMM parameters (15 % average improvement). Table 2 summarizes the adaptation of ω_i $(1 \leq i \leq N)$ to the GBM samples. In sample *G20856*, the means of Gaussian distributions were adjusted from 0.7 and 1.3 to 0.76 and 1.23 for the HMM states associated with copy number 1 and 3, respectively. Such complementary adjustments towards 1.0 are expected as, e.g. normal cell admixture affects the observed intensity ratios of both gains and deletions. A slightly higher SD than expected can be seen for copy number states 3 and 4. The optimal SD of the Gaussian distribution associated with copy numbers ≥ 5 is more than four times the expected deviation. Higher SD for this state than the other states can be seen for all GBM samples. This is due to higher spreading of experimental data above lower copy number distributions.

A profile plot for the whole genome of sample *G24460* is given in Figure 3. The two most common alterations in GBM, trisomy 7 and monosomy 10, were identified together with partial deletions of chromosomes 6 and 9 and trisomies of chromosomes 19 and 20. Moreover, a ≈5 Mb region of homozygous deletion on 9p.21 was identified that encompasses the *CDKN2A* (MIM 600160) (cyclin-dependent kinase inhibitor 2A) gene, which is commonly deleted in a wide variety of tumors and is known to be an important tumor suppressor gene.

In order to evaluate the characteristics and performances of SMAP we compared our results on the GBM data set with the results of DNACopy and BioHMM. Following the recommendations of Willenbrock and Fridlyand (2005), we applied *mergeLevels* to the segmentation results of DNACopy and BioHMM. The number of unique state means predicted per sample was then reduced from an average of 87.7 and 53.3 to an average of 43 and 44, respectively. The average number of states per chromosome and sample was reduced from 7.7 and 3.7 to 3.3 and 3.3, respectively. In contrast to these averages, the average number of (clone-assigned) states per sample identified with SMAP was five (four, six and five for sample *G20856*, *G24460* and *G20975*, respectively). It should be stressed that the means associated with these states are equivalent across the genome of a sample, since SMAP uses the same HMM parameters for the whole genome rather than one model per chromosome, which is the case for the other two methods. Consequently, interpretation of the segment means of DNACopy and BioHMM remains questionable, making a straight forward comparison between predicted copy numbers impossible. This is apparent in Figure 4A, in which a snapshot of chromosome 12 of sample *G20856* is shown. All three methods identified two states in the chromosome. SMAP identified a ≈2 Mb region of single copy number gain (state mean 1.23) whereas DNACopy and BioHMM suggested a region with altered mean of 1.05 and 1.06, respectively, that stretches ≈11 Mb.

SMAP was able to identify subtle changes in copy number in the analyzed samples. Figure 4B shows 7 Mb contained in a

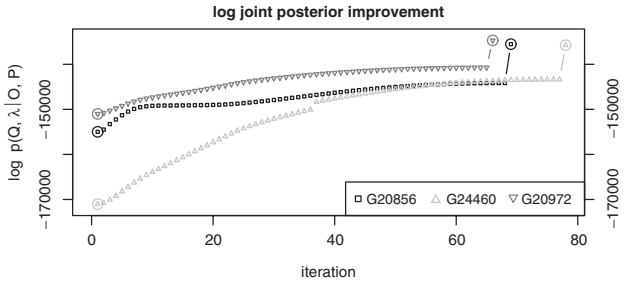

Fig. 2. Improvement of log $p(Q, \lambda|O, P)$ during execution of SMAP on the GBM samples *G20856*, *G24460* and *G20972*. Optimizations of the copy number assignments, using a modified Viterbi algorithm, are marked by larger, circled, dots and the iteration steps in between refer to iterations of the optimization of model parameters, using the gradient descent scheme.

Table 2. The initial ($\mu^{(1)}$, $\sigma^{(1)}$) and optimized values (μ^*, σ^*) of ω_i $(1 \leq i \leq N)$ on the GBM samples

Copy number			G20856		G24460		G20972	
	$\mu^{(1)}$	$\sigma^{(1)}$	μ^*	σ^*	μ^*	σ^*	μ^*	σ^*
0	0.40	0.10	0.40	0.05	0.47	0.11	0.40	0.05
1	0.70	0.10	0.76	0.11	0.77	0.12	0.77	0.12
2	1.00	0.10	1.02	0.11	1.00	0.11	1.01	0.10
3	1.30	0.10	1.23	0.14	1.23	0.14	1.23	0.14
4	1.60	0.10	1.65	0.26	1.66	0.25	1.69	0.27
≥5	4.00	0.10	4.05	0.42	4.09	1.16	4.09	0.43

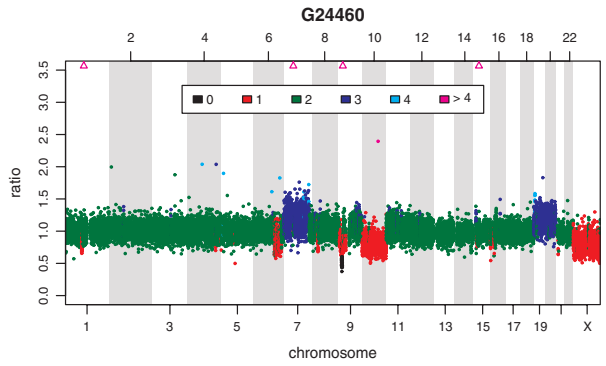

Fig. 3. Identified segments of copy number aberrations in GBM sample *G24460*.

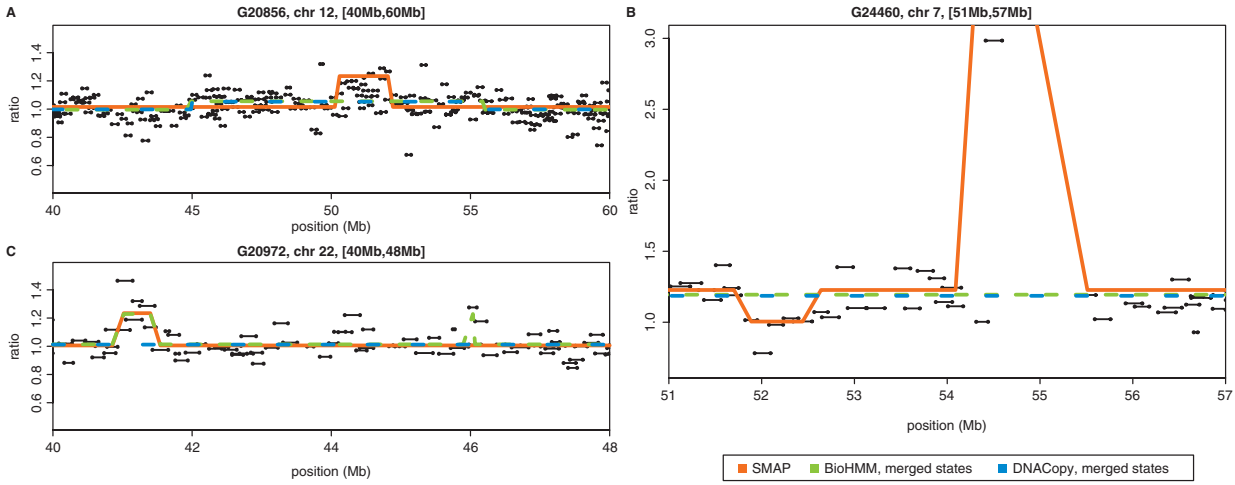

Fig. 4. Comparison of predictions made by SMAP, BioHMM and DNACopy on zoomed in regions of chromosome 12 in sample *G20856* (**A**), chromosome 7 in sample *G24460* (**B**) and chromosome 22 in sample *G20975* (**C**). In figure B, clone RP11-164O17 (≈54.5–54.8 Mb) was omitted due to size limitations (ratio ≈14.5). Horizontal black lines with dotted ends represent clones.

partial trisomy of chromosome 7 in sample *G24460*. Both BioHMM and DNACopy predicted the whole region as a trisomy (by equal segment mean of 1.19). SMAP, however, identified a diploid region of ≈0.5 Mb (state mean 1.0). Moreover, SMAP identified the amplification of the *EGFR* (MIM 131550) (epidermal growth factor receptor) gene around 54.5 Mb, commonly amplified in glioblastoma (Rasheed *et al.*, 2002), while the other two methods did not.

Although sensitive to subtle changes in intensity ratios, SMAP benefits from the overlap consideration in that segments of clones with deviations in intensity ratio are ignored if the overlap of clones at a previous state is sufficiently prominent. Such an example is given in Figure 4C, in which SMAP shares the prediction of gain at 41 Mb with BioHMM but ignores the gain at 46 Mb predicted by BioHMM. This is due to the genomic overlap of clones at this location. No gain was identified by DNACopy in this region.

On average, SMAP, DNACopy and BioHMM identified 147, 51 and 1599 breakpoints per sample, respectively. At these breakpoint locations, the average fraction of overlap between clones with non-equal state is 0.20, 0.40 and 0.61. Although it is still questionable how to determine the exact breakpoint location (in base pairs), as discussed by Stjernqvist *et al.* (2007), the overlap consideration of SMAP narrows the putative regions.

5 IMPLEMENTATION

SMAP is implemented in R (R Development Core Team, 2007) and C and is available from Bioconductor 2.1 (Gentleman *et al.*, 2004) in its current version, 1.2.0. Moreover, SMAP has been integrated as an analysis plug-in within the Linnaeus Centre for Bioinformatics Data Warehouse (Ameur *et al.*, 2006).

6 DISCUSSION AND CONCLUSIONS

In this article we present a novel method for DNA copy number profiling, called SMAP, that associates intensity ratios measured on a genomic microarray to discrete copy numbers. SMAP is based on segmental a posteriori maximization, a method in which, in our context, the most plausible assignments of DNA copy numbers to measurement points are found by maximizing the joint posterior probability of the parameters of a HMM and the HMM state sequence (copy number assignments). We propose a heterogeneous discrete-index HMM that deals with the dependency between clones due to genomic position (Section 3.1). Firstly, we include *distance-based transition probabilities* under the assumption that the probabilities of possible transition events between two clones should converge towards equality when their positions are genomically distant. Secondly, using a mixture model based on genomic overlap between clones we propose a method of dealing with the introduced dependency in terms of measurement values when analyzing tiling arrays. By using a MAP scheme, we enable the exploitation of prior knowledge about the data to provide flexibility and to enhance the ability to adapt the analysis to data sources with various characteristics. For instance, knowledge about normal cell admixture in the test DNA and the presence of noise in the data, which may cause intensity ratios to deviate from expected values, is crucial information that can be integrated in the process. A discussion regarding parameter settings of SMAP is included in the Supplementary Material. The inclusion of parameter optimization in the method further enables the adaptability of the HMM to observed data, restricted by prior probabilities and allows generation of the most plausible HMM model that describes the data.

We report superior performances of breakpoint prediction on synthetic data when compared with two other methods (Section 4.2). In a GBM study on the 32K BAC array (Section 4.3), we demonstrate that SMAP is able to identify both large-scale regions (Fig. 3) and changes affecting only single features on the array (singletons) with aberrant DNA copy number (Fig. 4B). Moreover, SMAP benefits from the overlap consideration in that transition events due to changes in intensity ratio are weighted against genomic overlap of previous clones (Fig. 4C). Furthermore, SMAP does not suffer from extensive computation time, a problem that is apparent with the continuous-index

HMM approach of Stjernqvist *et al*. (2007). Analysis of the GBM samples took 7.5 CPU seconds on average per sample on a MacBook Pro Intel Core Duo 2 GHz. Analysis of the GBM samples shows that SMAP readily recognizes already known and usually very large genetic aberrations in GBM. The vast majority of these tumor-related abnormalities, which are usually subtle, predominantly involving singleton changes, are identified by SMAP using only a single array hybridization. This is convincing evidence of SMAPs power. We have shown that, using a biologically motivated six-state model (van de Wiel *et al*., 2007) and genome-wise adaptation of model parameters, the identified distributions associated with each copy number become intuitive. This is not always the case with the compared segmentation approaches DNACopy and BioHMM, apparent in Figure 4A. These promising results provide a solid basis for unbiased identification of singletons that should be validated using an alternative methodology.

A possible future extension of SMAP would be to combine the results of profiling the observations using different orderings compatible with the overlap of clones. In its present form, SMAP traverses the observation sequence in one direction only (according to start position). This is due to the fundamental Markov chain criterion that the observations are represented as a chain of events. Although this does not restrict the mere ordering of clones, the behavior of SMAP may differ when the clones are ordered differently.

Although the analyses presented in this article concern two-channel microarray data from the 32K BAC array, this is not a restriction of SMAP. The method can easily be applied on one-channel microarray data, e.g. produced using Affymetrix, as long as a reference exists that enables the calculation of intensity ratios. We do not, however, consider allele specific copy number association, a feature which will require further extensions and may be considered in a future version of SMAP.

ACKNOWLEDGEMENTS

This work was partially supported by the Swedish Research Council, Wallenberg Consortium North, the Knut and Alice Wallenberg Foundation, the Swedish Foundation for Strategic Research, the Swedish Cancer Society and the U.S. Army Medical Research and Materiel Command, award no. W81XWH-04-1-0269. The authors would like to thank Prof. Andreas von Deimling and Christian Hartmann at the Institute for Neuropathology, Freie Universität Berlin, for providing the glioblastoma samples.

Conflict of Interest: none declared.

REFERENCES

Ameur,A. *et al*. (2006) The LCB Data Warehouse. *Bioinformatics*, **22**, 1024–1026.

Bagos,P.G. *et al*. (2004) Faster gradient descent training of hidden Markov models, using individual learning rate adaptation. In Paliouras,G. and Sakakibara,Y. (eds.) *ICGI, Lecture Notes in Computer Science*. Vol. 3264, Springer-Verlag, Berlin, Heidelberg. pp. 40–52.

Chen,W. *et al*. (2006) Array comparative genomic hybridization reveals genomic copy number changes associated with outcome in diffuse large B-cell lymphomas. *Blood*, **107**, 2477–2485.

Colella,S. *et al*. (2007) QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Res*, **35**, 2013–2025.

Diaz de Ståhl,T. *et al*. (2005) Chromosome 22 tiling-path array-CGH analysis identifies germ-line- and tumor-specific aberrations in patients with glioblastoma multiforme. *Genes Chromosomes Cancer*, **44**, 161–169.

Eilers,P.H. and de Menezes,R.X. (2005) Quantile smoothing of array CGH data. *Bioinformatics*, **21**, 1146–1153.

Engler,D.A. *et al*. (2006) A pseudolikelihood approach for simultaneous analysis of array comparative genomic hybridizations. *Biostatistics*, **7**, 399–421.

Fridlyand,J. *et al*. (2004) Hidden Markov models approach to the analysis of array CGH data. *J. Multivar. Anal.*, **90**, 132–153.

Gauvain,J.L. and Lee,C.H. (1992) MAP estimation of continuous density HMM: theory and applications. In *DARPA Sp. and Nat. Lang. Workshop*. Harriman, New York.

Gentleman,R.C. *et al*. (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, **5**, R80.

Hsu,L. *et al*. (2005) Denoising array-based comparative genomic hybridization data using wavelets. *Biostatistics*, **6**, 211–226.

Hupe,P. *et al*. (2004) Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. *Bioinformatics*, **20**, 3413–3422.

Iafrate,A.J. *et al*. (2004) Detection of large-scale variation in the human genome. *Nat. Genet.*, **36**, 949–951.

Jaynes,E.T. (2003) *Probability Theory - The Logic of Science*. Cambridge University Press, Cambridge.

Lai,W.R. *et al*. (2005) Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics*, **21**, 3763–3770.

Mantripragada,K.K. *et al*. (2004) Genomic microarrays in the spotlight. *Trends Genet.*, **20**, 87–94.

Marioni,J.C. *et al*. (2006) BioHMM: a heterogeneous hidden Markov model for segmenting array CGH data. *Bioinformatics*, **22**, 1144–1146.

Myers,C.L. *et al*. (2004) Accurate detection of aneuploidies in array CGH and gene expression microarray data. *Bioinformatics*, **20**, 3533–3543.

Olshen,A.B. *et al*. (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, **5**, 557–572.

Picard,F. *et al*. (2005) A statistical approach for array CGH data analysis. *BMC Bioinformatics*, **6**, 27–27.

Pinkel,D. *et al*. (1998) High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat. Genet.*, **20**, 207–211.

R Development Core Team (2007) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austriap. ISBN 3-900051-07-0.

Rabiner,L.R. (1990) A tutorial on hidden Markov models and selected applications in speech recognition. In Waibel,A. and Lee,K.-F. (eds.) *Readings in Speech Recognition*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 267–296.

Rasheed,A. *et al*. (2002) Molecular markers of prognosis in astrocytic tumors. *Cancer*, **94**, 2688–2697.

Redon,R. *et al*. (2006) Global variation in copy number in the human genome. *Nature*, **444**, 444–454.

Rueda,O.M. and Diaz-Uriarte,R. (2007) Flexible and accurate detection of genomic copy-number changes from aCGH. *PLoS Comput. Biol.*, **3**, e122.

Sebat,J. *et al*. (2004) Large-scale copy number polymorphism in the human genome. *Science*, **305**, 525–528.

Shah,S.P. *et al*. (2006) Integrating copy number polymorphisms into array CGH analysis using a robust HMM. *Bioinformatics*, **22**, 431–439.

Smith,M.L. *et al*. (2007) *snapCGH: segmentation, normalisation and processing of aCGH data*. R package version 1.4.0.

Solinas-Toldo,S. *et al*. (1997) Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. *Genes Chromosomes Cancer*, **20**, 399–407.

Stjernqvist,S. *et al*. (2007) Continuous-index hidden Markov modelling of array CGH copy number data. *Bioinformatics*, **23**, 1006–1014.

Tibshirani,R. and Wang,P. (2007) Spatial smoothing and hot spot detection for CGH data using the fused lasso. *Biostatistics*, **9**, 18–29.

van de Wiel,M.A. *et al*. (2007) CGHcall: calling aberrations for array CGH tumor profiles. *Bioinformatics*, **23**, 892–894.

Venkatraman,E.S. and Olshen,A.B. (2007) A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics*, **23**, 657–663.

Viterbi,A.J. (1967) Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE T. Inform. Theory*, **IT-13**, 260–269.

Wang,P. *et al*. (2005) A method for calling gains and losses in array CGH data. *Biostatistics*, **6**, 45–58.

Willenbrock,H. and Fridlyand,J. (2005) A comparison study: applying segmentation to array CGH data for downstream analyses. *Bioinformatics*, **21**, 4084–4091.