Published in final edited form as:

Biometrics. 2009 December ; 65(4): 1296–1305. doi:10.1111/j.1541-0420.2009.01226.x.

A Latent Class Model with Hidden Markov Dependence for Array CGH Data

Stacia M. DeSantis^{1,*}, E. Andrés Houseman², Brent A. Coull³, David N. Louis⁴, Gayatri Mohapatra⁵, and Rebecca A. Betensky³

¹Department of Biostatistics, Bioinformatics, and Epidemiology, Medical University of South Carolina, 135 Cannon Street, Suite 303, Charleston, South Carolina 29403, U.S.A.

²Center for Environmental Health and Technology, The Warren Alpert Medical School of Brown University and Department of Biostatistics, Harvard School of Public Health, 121 South Main Street, Room 217, Providence, Rhode Island 02903, U.S.A.

³Department of Biostatistics, Harvard School of Public Health, 655 Huntington Avenue, Boston, Massachusetts 02115, U.S.A.

⁴Department of Pathology, Massachusetts General Hospital and Harvard Medical School, Building 149 13th Street, Charlestown, Massachusetts 02129, U.S.A.

⁵Department of Pathology, Massachusetts General Hospital, Building 149 13th Street, Room 7015, Charlestown, Massachusetts 02129, U.S.A.

Summary

Array CGH is a high-throughput technique designed to detect genomic alterations linked to the development and progression of cancer. The technique yields fluorescence ratios that characterize DNA copy number change in tumor versus healthy cells. Classification of tumors based on aCGH profiles is of scientific interest but the analysis of these data is complicated by the large number of highly correlated measures. In this article, we develop a supervised Bayesian latent class approach for classification that relies on a hidden Markov model to account for the dependence in the intensity ratios. Supervision means that classification is guided by a clinical endpoint. Posterior inferences are made about class-specific copy number gains and losses. We demonstrate our technique on a study of brain tumors, for which our approach is capable of identifying subsets of tumors with different genomic profiles, and differentiates classes by survival much better than unsupervised methods.

Keywords

Array CGH; Hidden Markov Model; Latent class

1. Introduction

DNA sequence copy number changes have been implicated in the development and progression of cancer (e.g., Kallioniemi et al., 1994; Tirkkonen et al., 1998; Forozan et al.,

Supplementary Materials

Web Figures and text referenced in Sections 1, 4, and 5 are available under the Paper Information link at the *Biometrics* website <http://www.biometrics.tibs.org>.

2000). Allelic copy number changes are typically manifest as gains or losses of chromosomal regions. Chromosomal gains are often associated with the overexpression of oncogenes, while losses are classically associated with the underexpression of tumor suppressor genes. Detection of patterns of genomic gains and losses help to refine cancer diagnosis and management. At the individual level, genetic subtyping can predict response to chemotherapies and inform patient prognosis (e.g., Cairncross et al., 1998). At the population level, the detection and mapping of gains and losses is necessary to identify critical genes associated with certain cancers (Kallioniemi et al., 1994). The detection of chromosomal losses and gains, and the identification of copy number profiles are the focus of this article. Furthermore, where survival data are available, we use these data to supervise the classification into copy number profiles.

Array comparative genomic hybridization (aCGH) is a powerful, high-throughput tool used to assess DNA copy number variation within a given DNA sample in a whole genome analysis. It is a global screening approach that entails the simultaneous hybridization of differentially labeled test DNA (obtained from tumors) and reference DNA (obtained from lymphocytes of healthy individuals). The reference DNA is often a collection from multiple individuals, to avoid the problem of polymorphisms. Two different fluorescent dyes, usually Cy3 and Cy5, are used to label normal and tumor DNA fragments, respectively. These fragments, also called probes, are simultaneously hybridized to a microarray spotted, in our case, with bacterial artificial chromosomes (BAC) clones. Each spot on the microarray represents an individual sequence and a unique chromosomal locus. After hybridization, a fluorescence intensity ratio is calculated at each location on a chromosome, where thousands of clones are usually represented on a microarray. The resulting measures are highly correlated $\log_2$ fluorescence ratios comparing tumor to healthy DNA fragments. The initial goal of statistical analysis is to find locations of mean $\log_2$ ratio change, which may encompass whole or partial chromosomes. A subsequent goal is to identify a few common aCGH profiles.

The analysis of these data presents a statistical challenge for several reasons. First, the observed fluorescence ratios are very noisy and require correction for nonbiological sources of error (Brown, Goodwin, and Sorger, 2001). A second feature of aCGH data, which presents an analytic challenge, and which many current statistical methods do not exploit, is the dependence between clones ordered along the chromosomes. More specifically, for certain parts of the chromosome, gain (loss) at a region often results in an increased probability of gain (loss) at a neighboring region. Web Figure 1 shows a heatmap of the data obtained from the divisive hierarchical clustering algorithm, which begins with every subject in one group, and, using a dissimilarity matrix divides clusters until each tumor is its own group. The tumor is listed horizontally and the clone is denoted vertically. The map shows both correlation across the genome and the general tendency for GBM and oligo tumors to exhibit different genetic profiles—but there is still much variation within diagnosis. A third challenge is that the number of variables far exceeds the number of replications of the experiment, and as a result, deducing classification assignments for the purpose of identifying robust profiles is difficult.

To meet these goals and challenges, we propose a novel latent class model for analysis of aCGH data, assuming tumors arise from subpopulations, or latent classes, that share similar patterns of alteration across the genome. Our proposed methodology exploits this correlation in order to mitigate the number of clones. To account for the correlation among neighboring clones, within each latent class, we assume a heterogeneous hidden Markov dependence structure. Specifically, at each clone, we assume an underlying copy number state of gain, loss, or no change, that gave rise to the observed data, which is measured with error, and with dependence parameters that potentially differ across latent classes. Computation time

depends on the number of clones, tumors, and particular model specifications. For our dataset obtained using a custom CGH array (1100 clones measured on 46 tumors), our method takes between 30 and 70 hours to run 50,000 iterations after an equal MCMC burn-in. While the approach may not be computationally feasible for large-scale commercial arrays, it is feasible for our group's custom array, which is necessary for accurate assessment of formalin-fixed paraffin-embedded tissue, such as is widely available for brain tumors. In Section 2 we describe the aCGH data obtained from brain tumor samples obtained from Massachusetts General Hospital in Boston, Massachusetts. In Section 3 we review some methods found in the literature used to analyze aCGH data, with a focus on hidden Markov models (HMMs), as well as statistical methods proposed for the classification of subjects based on the analysis of other types of functional data, namely correlated, high-dimensional, and/or time-dependent data. In Section 4 we present a novel HMM and discuss prior assumptions and posterior inference. In Section 5 we present results obtained from applying the methodology to the brain tumor data and compare these with unsupervised techniques. We then conclude with a discussion in Section 6.

2. Brain Tumor aCGH Data

The characterization of human gliomas based on molecular genetic alterations has received much recent attention (Cairncross et al., 1998; Ino et al., 2001). The two subtypes in our study, oligodendroglioma (oligo) and glioblastoma (GBM), are associated with different genomic profiles. For example, several studies have shown that allelic loss on chromosomes 1p and 19q is associated with the diagnosis of oligodendroglioma, as well as longer survival after chemotherapy (Cairncross et al., 1998). Other genome-wide studies imply that multiple copy gains of portions of chromosome 7, notably EGFR amplification, often occur in GBM. While histological diagnosis of GBM versus oligodendroglioma is straightforward for the so-called "classical" cases, it is not so for the remaining cases. Furthermore, within the oligos, in particular, there is a wide range of survival. For these reasons it is of current interest to discover novel genetic alterations to enable more accurate diagnoses and prognoses for these highly heterogeneous histopathological groupings.

Until recently, aCGH could only reproducibly be performed on frozen tissue samples and with significant tissue amounts. For brain tumors, however, paraffin-embedded tissue blocks from small stereotactic biopsies may be the only tissue routinely available. Methods have therefore recently been developed to analyze formalin-fixed, paraffin-embedded (FFPE) tissue, and were used to construct a custom BAC array representing chromosomes 1, 7, 19, and X (Mohapatra et al., 2006). Such a targeted and size-limited array is likely to be the platform that will be used in the future in diagnostic clinical settings, once prognostic alterations are discovered. Mohapatra et al. (2006) demonstrated the new aCGH technology on a subset of tissues available in the current study, which aimed to establish this technology for future discovery of novel genetic alterations, and to confirm detection of known genetic alterations. Array CGH was performed on 46 frozen glioma tissue samples: 24 oligos and 22 GBMs. To rule out germ line polymorphism, a second hybridization was performed where the germ line DNA from the patient is hybridized against the tumor DNA. If the DNA copy number change normalizes to the baseline, then this is a polymorphism. Survival endpoints were available for all patients.

3. Review of Existing Methods

There is a sizeable body of literature addressing the analysis of aCGH data (e.g., Engler et al., 2006). Signal normalization must first be performed to correct for systematic spatial or intensity biases in the ratios (Hupe et al., 2004). Statistical techniques used for normalization are not the focus of this article; we assume that the data have been suitably normalized.

There are a few papers in the literature that account for spatial dependence among clones using HMMs. Some analytical methods specify, a priori, the existence of three or four underlying, unobserved copy number states corresponding to allelic loss, no change, single copy gain, and multiple copy gain. The HMM approach accounts for inherent dependencies in the intensity ratios and provides posterior distributions that are used to make inferences about gains and losses in copy number, which are defined as locations of $\log_2$ ratio change. Considering the correlation between neighboring clones, Bröet and Richardson (2006) proposed a spatially correlated three-state mixture model that allows mixture weights to be correlated for neighboring clones. Their model is estimated using Bayesian methods, which give rise to posterior allocation probabilities that conveniently allow sequences to be classified into three states corresponding to loss, no change, or amplification of copy number. Engler et al. (2006) assumed a Gaussian mixture model with a hidden Markov dependence structure, with estimation based on a pseudolikelihood approach constructed as the product of overlapping triples of $\log_2$ ratios. This method borrows strength across clones by considering the data in triplets, and results in assessments of the likelihood of genetic alterations at each clone, along with graphical displays. Guha, Li, and Neuberg (2006) considered a four-state homogeneous HMM that implements a Metropolis-within-Gibbs algorithm for estimation.

Other methods, called segmentation methods, do not assume a fixed number of states, but instead focus on identifying segments of common $\log_2$ ratio mean. Fridlyand et al. (2004) proposed a discrete-state homogeneous HMM where underlying states are considered segments of a common mean, but the number of hidden groups is not prespecified. The procedure seeks to identify copy number transitions, whole chromosome aberrations, and focal aberrations (a single clone whose state differs from both neighboring clones). Marioni et al. (2006) extended this approach by considering a heterogeneous HMM in which the probability of transitioning between states depends on the distance between adjacent clones. In their method, transition probabilities are allowed to depend on known distance from neighboring clones, as well as vectors of observed covariates, with the goal of segmenting ratios into unspecified states.

More generally, several classification methods exist for functional data (e.g., James, 2002; Muller and Stadtmüller, 2005; Ray and Mallick, 2006), of which aCGH is an example. However, they do not apply to the type of dependence observed in aCGH data.

4. Bayesian Hidden Markov Model for aCGH

4.1 Unsupervised Model

In order to identify a few subsets of tumors with different genomic profiles, it is helpful to have a scheme that allows for the assignment of tumors to unobservable groups. Latent class analysis allows us to achieve this because it results in a probabilistic class assignment; tumors are assigned to a latent class based on their highest estimated posterior probabilities of class membership. We consider a heterogeneous HMM that allows clones to move through the state-space according to clone- and latent class-dependent transition probabilities. A fully heterogeneous HMM allows the transition probabilities to depend on clone, e.g., Marioni et al. (2006). As our estimation procedure is complicated by the latent class structure, we adopt a more computationally feasible heterogeneous transition matrix at every 50 markers. We will show that this more parsimonious version of the model fits the aCGH data adequately. We describe further in Section 6 how this model could be extended to incorporate fully heterogeneous transition probabilities without increasing computational demand.

For a given tumor we let Y_{lm} denote the observed $\log_2$ ratio for the m th mapped clone, $m = 1, \dots, M_l$ on chromosome l . We let s_{lm} represent the unknown copy number state associated with clone m on chromosome l . The unobserved hidden state variable, s_{lm} , can take on three discrete values, $\{L, O, G\}$, which correspond to copy number loss, no copy number alteration, and copy number gain, respectively. We let the parameters, (μ_L, μ_O, μ_G) represent the expected $\log_2$ ratios, and $(\sigma_L^2, \sigma_O^2, \sigma_G^2)$ represent the variances of the $\log_2$ ratios associated with the three states. Although the theoretical value for one copy number gain is $\log_2(2/1) = 0.58$, for loss is $\log_2(1/2) = -1$, and for no change is $\log_2(1/1) = 0$, the theoretical values for L , O , and G are rarely observed in practice, as they are subject to measurement error arising from many sources, such as contamination of the tumor sample with normal cells. Our model assumes that the $\log_2$ ratios are normally distributed and conditionally independent given the hidden state variable,

$$Y_{lm} | s_{lm} \sim N(\mu_{s_{lm}}, \sigma_{s_{lm}}^2). \quad (1)$$

We assume that clones move through the state space according to a Markov chain with class-dependent heterogeneous transition probabilities. Letting the superscript, k , denote latent class, and m clone, then for a given chromosome l ,

$$P_{ab}^{km} = \text{pr}\{(s_{lm+1}=a | s_{lm}=b, K=k)\},$$

denotes the element of the 3×3 matrix of transition probabilities that represents the probability of transitioning from state b to state a at each locus, conditional on latent class, and $k \in 1, \dots, K$. Finally, η_k denotes the prevalence of latent class k .

4.2 Supervised Model

In cancer management, clinicians are interested in assigning patients to diagnostic groups that correlate with prognosis. To accommodate this, we consider latent class estimation supervised by survival, and assume a proportional hazards model for the effect of latent class membership on survival as described in Larsen (2004) and DeSantis et al. (2008). In the presence of a survival outcome, the observed data for chromosome l for tumor i consist of $\{y_{il}, t_i, \delta_i\}$, where $y_{il} = \{y_{il1}, \dots, y_{ilM_l}\}$, t_i is the minimum of survival and censoring times, and δ_i is the failure indicator. We assume that conditional on the tumor's latent class, K_i , (t_i, δ_i) and y_{il} are independent. Censoring is assumed to be uninformative. We let v_k denote the relative log hazard of membership in class k versus the reference class 1. For the estimation of the baseline hazard rate, we assume a piecewise constant exponential hazard. This approach is robust in that it is approximately nonparametric, and easy to model in a Bayesian setting. In particular, the event times are ordered and the time axis is split into R intervals, $(0, u_1], (u_1, u_2], \dots, (u_{R-1}, u_R]$, where, for each interval, there is a constant baseline hazard, $\lambda_0(t) = \lambda_r$ for $t \in I_r = (u_{r-1}, u_r]$. To properly define the density over the R intervals, we let $d_{ir} = 1$ if subject i has an event or is censored in interval r , and 0 otherwise. The density of the event time is thus:

$$P(t_i, \delta_i | k_i) = \prod_{r=1}^R \{\lambda_r \exp(v_{k_i})\}^{d_{ir}\delta_i} \times \exp \left\{ -d_{ir} \left[\lambda_r(t_i - u_{r-1}) + \sum_{g=1}^{r-1} \lambda_g(u_g - u_{g-1}) \right] \exp(v_{k_i}) \right\}. \quad (2)$$

The optimal number of intervals depends on the number of subjects and events; we suggest allowing approximately 10 events in each interval.

4.3 Estimation

We adopt a Bayesian approach to model fitting, which requires specifying prior distributions for all unknown parameters. The posterior distribution of the parameters cannot be computed analytically, so we use a Gibbs sampler to simulate random draws from the posterior distribution, as implemented in WinBugs (Spiegelhalter et al., 1994, 2003; Lunn, 2003). Following a burn-in of 15,000 iterations, we monitor posterior distributions over 15,000 further iterations of the chain, saving every 15th sample, both to reduce autocorrelation in the MCMC simulation, and to reduce storage requirements. The length of burn-in and monitoring was determined by assessing trace and autocorrelation plots (Kass et al., 1998), which are presented for select parameters in Supplementary Material.

4.4 Prior Specification

It is natural to specify informative priors for the vector of $\log_2$ ratio means since theoretically, these means have an underlying biological meaning (Guha et al., 2006; Bröet and Richardson, 2006). Adopting the idea of informative normal priors from Guha et al.

(2006), we specify $\mu_L \sim N(-1, \tau_L^2)$, $\mu_O \sim N(0, \tau_O^2)$, $\mu_G \sim N(0.58, \tau_G^2)$; that is, the normal priors are centered at the theoretical mean for that particular state. Guha et al. (2006) used truncated normal priors ensuring that $\mu_L < \mu_O < \mu_G$, and tested several different values for truncation; for our data, truncation was unnecessary as inference was the same (i.e., label-switching was not a problem as the class assignment remained stable over a range of priors).

As results were not sensitive to choice of prior for the τ_j^2 in the tested range of 1 to 100, we set $\tau_L^2 = \tau_O^2 = \tau_G^2 = 1$. For the measurement error variances from expression (1), $\sigma_L^2, \sigma_O^2, \sigma_G^2$, we specified a noninformative gamma prior, $\text{Inv-gamma}(0.01, 0.01)$, and test the sensitivity by comparing to results obtained under a $U(0, 1)$ prior. We chose the priors for the rows of the heterogeneous transition matrix, p^{km} , to reflect the assumption that consecutive $\log_2$ ratios on a chromosome are generally constant, with an unknown number of change points.

Accordingly, there is a tendency for there to be very few transitions between states. We therefore assume independent Dirichlet priors for the rows and select row-specific parameters of the Dirichlet distribution that allow a higher probability of remaining in the same state for adjacent clones. Letting j denote the row of the transition matrix, we specify

the priors, $p_j^{km} \sim \text{Dirichlet}(\alpha_{kmj})$ where $\alpha_{km1} = (2, 1, 1)$, $\alpha_{km2} = (1, 2, 1)$, and $\alpha_{km3} = (1, 1, 2)$, for $k = 1, \dots, K$ and $m = 1, \dots, M_l$ for chromosome l , but results were not sensitive to this choice of prior within one order of magnitude. Our choice of initial values is explained in the Supplementary Materials.

4.5 Posterior Inference

For each tumor, we ascertained regions of gain and loss along the chromosome by calculating posterior probabilities of hidden state membership, $p(s_{lm} = L | D)$, $p(s_{lm} = O | D)$, and $p(s_{lm} = G | D)$, where D consists of y for unsupervised classification and (t, δ, y) for supervised classification using survival. For each tumor, i , we calculate posterior probabilities of latent class membership, $p(K_i | D)$, and assign tumors to the latent class for which they have the highest posterior probability. This allows us to visualize the genomic profiles stratified by latent class, and to discern whether the classes estimated by the model have distinct profiles. With our methodology, it is possible to locate the change points at which there are likely transitions between states by inspection of plots similar to those presented by Engler et al. (2006). All figures in the article have two vertical axes: one on the left that denotes the posterior probability of state membership for each of the three states, L , O , and G , and one on the right that denotes the posterior mean $\log_2$ expression ratios. The posterior probabilities of loss are represented by the red lines extending from the bottom of the figure upward. The posterior probabilities of gain are represented by the blue lines

extending from the top of the graph downward. The posterior probabilities of nochange are represented by the white lines, and are calculated as one minus the probabilities of loss and gain. Posterior mean $\log_2$ ratios are presented as black dots. The vertical dashed lines demarcate the end of the respective chromosomal arms. The natural simplicity of the visualization approach in identifying regions of copy number alteration makes our method feasible for clinicians not trained in Bayesian statistics. To assess whether our Bayesian model is consistent with the observed data, we performed posterior predictive checks of the model as outlined in Supplementary Materials.

5. Application to Brain Tumor Data

We applied the latent class model with heterogeneous Markov dependence to the aCGH brain tumor data described in Section 2. The goal is to discover a few classes defined by aCGH profiles. While the histologic distinction of GBM versus oligo-dendroglioma is not difficult for the “classic” cases, it can be challenging for the remaining cases. Furthermore, there is considerable heterogeneity in the survival of patients with oligodendroglioma, and so we turn to survival to help guide the construction of novel classes. We certainly expect an aCGH-based classification to correlate with the histological classification, but not perfectly, as it will bring to bear the complete genomic profiles and when supervised, its development will also be informed by survival. We fit 2- and 3-class models to the data, as we were mainly interested in uncovering just a few subpopulations, and were limited by a small sample size.

Figure 1 shows the class-specific posterior probabilities of state membership obtained from fitting the unsupervised 2-class model, averaged across the tumors in each class. Class 2, with a prevalence of 75%, is characterized by a higher probability of copy number gain on chromosome 7 known as EGFR amplification, and like latent class 1, moderate probabilities of loss at chromosomes 1p and 19q. The methodology is unsuccessful in separating patients by survival as median survival in class 1 versus 2 is 847 versus 850 days. A post hoc survival analysis with latent class assignment as a predictor implies that latent class is not a predictor of survival ($p = 0.40$).

In contrast, Figure 2 shows results from fitting the 2-class supervised model. The two notable differences are the virtually zero posterior probabilities of gain on chromosome 7p and larger posterior probability of loss on chromosome 19q for patients in latent class 1. Both polymorphisms are known to result in a more favorable prognosis. Of the nine tumors assigned to latent class 1, which is characterized by no gain at chromosome 7p (lack of EGFR amplification) and which has a prevalence of 20%, eight are oligodendrogliomas. Median survival time for patients in this class is 1718 versus 585 days (for class 2) and the posterior mean estimate of v , the log hazard rate comparing class 2 to 1, is 27.51, 95% credible interval = [5.84, 64.28]. Subsequent inspection of the raw data (not shown) showed that the majority of tumors in class 2 exhibited EGFR amplification characterized by multiple copy gains on chromosome 7p. Twenty-one of the tumors in this class were diagnosed as GBM and 16 as oligo. These latent classes are survival driven and also better clarify diagnosis than the unsupervised 2-class model, which did not correlate with diagnosis. Importantly, the analysis demonstrates that EGFR amplification is a marker of significantly poorer prognosis, which is information that we cannot glean from existing unsupervised aCGH clustering methods. In Section 5.2, we show that this supervised approach also outperforms other unsupervised clustering approaches.

We further examined the differential tumor classification resulting from fitting our 2-class supervised and unsupervised models. Eight tumors originally classified into class 2 without supervision moved into class 1 (denoted the “good prognosis” class) with supervision

(median survival of these eight tumors = 1800.5 days). Nine tumors classified into class 1 without survival were moved to class 2 (poor prognosis class) when survival was included (median survival of these nine tumors = 615 days). These consisted of five poor prognosis oligo tumors as well as four GBMs. We therefore conclude that the supervised model provides clarification of class assignment based on prognosis, but not necessarily diagnosis. Thus classification with supervision offers more than diagnosis alone. We also note that posterior probabilities of class assignment were all near 0 or 1.

Figure 3 shows results from fitting the 3-class unsupervised model to the data. Median survival times are 595, 910, and 539 days, respectively. All subjects assigned to class 3 have GBM diagnosis, and this class is characterized by poorest survival, high posterior probability of gain on chromosome 7, and low posterior probability of loss at 1p and 19q compared with the other latent classes. Figure 4 shows results from fitting the 3-class supervised model. Posterior estimates [95% Credible Intervals] of the hazard ratio for class 2 versus class 1 is 1.55 [0.76, 3.51], and for class 3 versus class 1 is 5.26 [1.30, 27.54]. This result is consistent with biology of glioma tumors as latent class 1 has very low posterior probability of gain at 7p, while class 2 and even more so class 3, exhibit higher probabilities of posterior gain on chromosome 7p, all consistent with the increasing estimated hazard rates. As the posterior distribution of the hazard ratios overlap considerably, we feel that the 2-class supervised model in Figure 2 provides the best fit to the data and is most informative in distinguishing glioma tumors according to both copy number profile and prognosis. Posterior predictive checks of the model (Supplementary Materials), show no evidence of lack of fit of the model.

5.1 Sensitivity to Prior Specification

The 2- and 3-class supervised models were assessed for sensitivity to specification of the prior distribution for the variance of the $\log_2$ ratios as there is a great deal of prior information about the feasible values of these parameters from other studies. Ours and other aCGH studies show that the variance of the $\log_2$ ratios is generally quite small—in our study the variance ranged from 0.01 to 0.66 for the 46 tumors. Thus, rather than a relatively noninformative gamma prior for this parameter, a more informative prior is a $U(0,1)$, i.e., one that allows for double the observed variance (Gelman, 2006). We assessed sensitivity to this hyperprior and found that it did not impact inference, where a change in inference was measured by differential classification into latent classes based on posterior probabilities. This result speaks to the reproducibility of the Bayesian estimation method even in the complete absence of prior information.

Along similar lines, if the known molecular genetics of the tumors in question suggest the existence of a subgroup of patients with a particular genetic profile, one could easily incorporate this knowledge by using more informative Dirichlet prior distributions on the rows of the transition matrices. The parameters of the Dirichlet distribution, $(\alpha_{km1}, \alpha_{km2}, \alpha_{km3})$ would be altered to put more or less weight on the probability of transitioning between particular states, as dictated by this prior knowledge. For the glioma data, we have found that results are not sensitive within a reasonable range of the Dirichlet parameters but we were able to drive the posterior frequency of transitioning by increasing the appropriate parameter by one to two orders of magnitude. For example, setting $\alpha_{km2} = (1, 100, 1)$ for all m clones alters posterior inference by resulting in a latent class characterized by no change across the genome, as the prior likelihood of transitioning from the no change state is extremely small.

5.2 Comparison with Other Clustering Methods

We compared our method with two unsupervised clustering techniques: K-means and hierarchical clustering. These methods do not take into account dependence between neighboring clones, do not identify regions of interest on the genome, and do not provide an estimate of the effect of survival on classification, except in a post hoc analysis. The first split of the hierarchical clustering algorithm (Web Figure 1) classifies 5 GBMs into one group and the remainder of tumors into another. These 5 GBM tumors appear to exhibit EGFR amplification on chromosome 7p but the method misses four tumors that also clearly exhibit multiple copy gain. In a second stage analysis, these estimated clusters are associated with survival (median survival = 375 versus 927 days, log rank p-value = 0.004). A three-cluster split in the hierarchical analysis moves only one subject into its own class, and therefore does not further clarify group-specific genomic alterations or survival differences (a K-means analysis with two and three specified clusters resulted in the exact same split as hierarchical clustering and therefore the same inference). Recall in comparison that the supervised latent class approach resulted in median survival times of 585 versus 1718 for the poor versus good prognosis classes, respectively, along with a posterior estimate and credible interval for the hazard ratio in the joint model. Although hierarchical clustering and K-means are also capable of separating groups by survival, this can only be demonstrated by a post hoc analysis, and in addition, the split is less compelling than that resulting from our supervised approach.

The utility of model-based classification approaches versus hierarchical clustering is cited in the literature. In genomic contexts where there is high correlation between loci, Siegmund et al. (2004) remark that mixture model approaches to classification typically have lowest cross-validation and classification errors when compared to hierarchical clustering methods. This remains unexplored in the copy number literature. In addition, model-based clustering allows us to incorporate biological structure as part of the model.

6. Discussion

We have presented a novel latent class model that can be used for the classification of tumors based on $\log_2$ ratios of relative copy number obtained from high-throughput arraybased comparative genomic hybridization. Using this model, we detected and characterized genome-wide copy number alterations on more than one thousand clones arising from several chromosomes, and classified patients into subgroups that were highly correlated with glioma diagnoses and survival. We also showed that the Bayesian estimation procedure gives rise to posterior predictive distributions that are consistent with the observed data. We demonstrated in detail how supervision improves the ability to predict prognosis. Another important contribution of our method is that we can impose informative prior knowledge about known regions of genomic alteration.

It is possible to introduce variability to the means at the chromosome level although this would add a layer of complexity onto the model that we were able to avoid. Engler et al. (2006) recognized this issue and addressed it through the introduction of random effects for these mean levels at the chromosomal level. However, this complicates the computational burden of the method, and in several examples, they found the ultimate calls of gain and loss to be similar with and without the use of random effects. That is, the distinction between levels of loss, no-change, and gain is greater than the variability across chromosomes within any one of these states.

Although no formal model selection criteria were applied in this example, supervised posterior parameter estimates of hazard ratios provided a natural way to compare supervised and unsupervised models, as well as supervised models with different numbers of latent

classes. For more formal model comparison, DeSantis et al. (2008) considered the use of the deviance information criteria (DIC) for model selection in a high-dimensional latent class setting, but this and other work showed it to perform poorly in latent class settings (Celeux et al., 2006). It is of interest to pursue a criterion such as the DIC so that we could perform formal model selection in the current setting.

While it is theoretically possible to fit a fully heterogeneous model to our data, due in part to the large number of clones as well as the effective loss of number of independent pieces of information as a result of the imposed latent class structure, it was time consuming to implement in practice; we therefore settled for specifying heterogeneous transition matrices at every 50 markers. Marioni et al. (2006) considered a heterogeneous HMM for the analysis of aCGH data under the assumption that the transition probabilities depend, in some structured way, on the distance between clones—our approach is similar in that we used a fixed number of markers as a representation of distance. Posterior predictive checks suggest that our simplified model is adequate for these data. Certainly there are applications in cancer for which a fully heterogeneous model might be more suitable. Further research might use measures of clonal distance within a model that incorporates smoothing of clone-dependent transition probabilities. For example, an autoregressive structure similar to that proposed by Cargnoni, Muller, and West (1997) in the context of time series data might be imposed on the transition probabilities to reduce the number of parameters.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

REFERENCES

- Bröet PE, Richardson S. Detection of copy number changes in CGH microarrays using a spatially correlated mixture model. *Bioinformatics* 2006;22:911–918. [PubMed: 16455750]
- Brown CS, Goodwin PC, Sorger PK. Image metrics in the statistical analysis of DNA microarray data. *Proceedings of the National Academy of Sciences USA* 2001;98:8944–8949.
- Cairncross JG, Ueki K, Zlatescu M, Lisle DK, Finkelstein DM, Hammond RR, Silver JS, Stark PC, Macdonald DR, Ino Y, Ramsay DR, Louis DN. Specific genetic predictors of chemotherapeutic response and survival in patients with anaplastic oligodendrogliomas. *Journal of the National Cancer Institute* 1998;90:1473–1479. [PubMed: 9776413]
- Cargnoni C, Muller P, West M. Bayesian forecasting of multinomial time series through conditional Gaussian dynamic models. *Journal of the American Statistical Association* 1997;92:640–647.
- Celeux G, Forbes F, Robert CP, Titterton DM. Deviance information criteria for missing data models. *Bayesian Analysis* 2006;1:1–23.
- DeSantis, SM.; Houseman, EA.; Coull, BA.; Betensky, RA. Supervised latent class models for high-dimensional data. Technical Report. 2008.
- Engler DA, Mohapatra G, Louis DN, Betensky RA. A pseudolikelihood approach for simultaneous analysis of array comparative genomic hybridizations. *Biostatistics* 2006;7:391–421.
- Fridlyand J, Snijders AM, Pinkel D, Albertson DG, Jain AN. Hidden Markov models approach to the analysis of array CGH data. *Journal of Multivariate Analysis* 2004;90:511–522.
- Forozan F, Mahlamaki EH, Monni O, Chen YD, Veldman R, Jiang Y, Gooden GC, Ethier SP, Kallioniemi A, Kallioniemi OP. Comparative genomic hybridization analysis of 38 breast cancer cell lines: A basis for interpreting complementary DNA microarray data. *Cancer Research* 2000;60:4519–4525. [PubMed: 10969801]
- Gelman A. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis* 2006;1:515–534.
- Guha S, Li Y, Neuberg D. Bayesian hidden Markov modeling of array CGH data. *Bepress, Working Paper* 24. 2006

- Hupe P, Stranksy N, Thiery JP, Radvanyi F, Barillot E. Analysis of array CGH data from signal ratio to gain and loss of DNA regions. *Bioinformatics* 2004;20:3413–3422. [PubMed: 15381628]
- Ino Y, Betensky RA, Zlatescu MC, Sasaki H, Macdonald DR, Stemmer-Rachamimov AO, Ramsay DA, Cairncross JG, Louis DN. Molecular subtypes of anaplastic oligodendroglioma: Implications for patient management and diagnosis. *Clinical Cancer Research* 2001;7:839–845. [PubMed: 11309331]
- James GM. Generalized linear models with functional predictors. *Journal of the Royal Statistical Society, Series B* 2002;64:411–432.
- Kallioniemi A, Kallioniemi OP, Piper J, Tanner M, Stokke T, Chen L, Smith HS, Pinkel D, Gray JW, Waldman FM. Detection and mapping of amplified DNA sequences in breast cancer by comparative genomics hybridization. *Proceedings of the National Academy of Sciences of the United States of America* 1994;91:2156–2160. [PubMed: 8134364]
- Kass RE, Carlin BP, Gelman A, Neal RM. Markov chain Monte Carlo in practice, a roundtable discussion. *American Statistician* 1998;52:93–100.
- Larsen K. Joint analysis of time-to-event and multiple binary indicators of latent classes. *Biometrics* 2004;60:85–92. [PubMed: 15032777]
- Lunn DJ. WinBUGS development interface WDev. *ISBA Bulletin* 2003;10:10–11.
- Marioni JC, Thorne NP, Tavaré S, Radvanyi F. BioHMM: A heterogeneous hidden Markov model for segmenting array CGH data. *Bioinformatics* 2006;22:1144–1146. [PubMed: 16533818]
- Mohapatra G, Betensky RA, Miller ER, Carey B, Gaumont DA, Louis DN. Glioma test array for use with formalin-fixed paraffin-embedded tissue. *Journal of Molecular Diagnostics* 2006;8:268–276. [PubMed: 16645215]
- Muller HG, Stadtmüller U. Generalized functional linear models. *Annals of Statistics* 2005;33:774–805.
- Ray S, Mallick BK. Functional clustering by Bayesian wavelet methods. *Journal of the Royal Statistical Society, Series B* 2006;68:305–332.
- Siegmund KD, Laird PW, Laird-Offringa IA. A comparison of cluster analysis methods using DNA methylation data. *Bioinformatics* 2004;20:1896–1904. [PubMed: 15044245]
- Spiegelhalter, DJ.; Thomas, A.; Best, NG.; Gilks, WR.; Lunn, D. Bayesian inference using Gibbs sampling. Cambridge, UK: MRC Biostatistics Unit; (1994, 2003). www.mrc-bsu.cam.ac.uk/bugs/
- Tirkkonen M, Tanner M, Karhu R, Kallioniemi A, Isola J, Kallioniemi OP. The molecular cytogenetics of primary breast cancer by CGH. *Genes Chromosomes & Cancer* 1998;21:177–184. [PubMed: 9523192]

Figure 1.

Unsupervised class-specific posterior probabilities of hidden state (left axis) and posterior mean $\log_2$ ratios (right axis) obtained from fitting the 2-class heterogeneous latent class model with hidden Markov dependence to the brain tumor data. This figure appears in color in the electronic version of this article.

Figure 2.

Supervised class-specific posterior probabilities of hidden state (left axis) and posterior mean $\log_2$ ratios (right axis) obtained from fitting the supervised 2-class heterogeneous latent class model with hidden Markov dependence to the brain tumor data. This figure appears in color in the electronic version of this article.

Figure 3. Unsupervised class-specific posterior probabilities of hidden state (left axis) and posterior mean $\log_2$ ratios (right axis) obtained from fitting the 3-class heterogeneous latent class model with hidden Markov dependence to the brain tumor data. This figure appears in color in the electronic version of this article.

Figure 4.

Supervised class-specific posterior probabilities of hidden state (left axis) and posterior mean $\log_2$ ratios (right axis) obtained from fitting the 3-class heterogeneous latent class model with hidden Markov dependence to the brain tumor data. This figure appears in color in the electronic version of this article.