GIMscan: A New Statistical Method for Analyzing Whole-Genome Array CGH Data

Yanxin Shi[1], Fan Guo[1], Wei Wu[2], and Eric P. Xing[1]*

[1] School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, 15213
[2] Division of Pulmonary, Allergy, and Critical Care Medicine, University of Pittsburgh, Pittsburgh, PA, 15213

Abstract. Genetic instability represents an important type of biological markers for cancer and many other diseases. Array Comparative Genome Hybridization (aCGH) is a high-throughput cytogenetic technique that can efficiently detect genome-wide genetic instability events such as chromosomal gain, loss, and more complex aneuploidity, collectively known as genome imbalance (GIM). We propose a new statistical method, Genome Imbalance Scanner (GIMscan), for automatically decoding the underlying DNA dosage states from aCGH data.GIMscan captures both the intrinsic (nonrandom) spatial change of genome hybridization intensities, and the prevalent (random) measurement noise during data acquisition; and it simultaneously segments the chromosome and assigns different states to the segmented DNA. We tested the proposed method on both simulated data and real data measured from a colorectal cancer population, and we report competitive or superior performance of GIMscan in comparison with popular extant methods.

1 Introduction

A hallmark of the defective cells in precancerous lesions, transformed tumors, and metastatic tissues, is the abnormality of gene dosage caused by regional or whole chromosomal amplification and deletion in these cells [1]. Cytogenetic and molecular analysis of a wide range of cancers have suggested that amplifications of proto-oncogenes and deletions or loss of heterozygosity (LOH) of tumor suppressor genes can seriously compromise key grow-limiting functions (e.g., cell-cycle checkpoints), cell-death programs (e.g., apoptotic pathways), and self-repair abilities (e.g., DNA repair systems) of injured or transformed cells that are potentially tumorigenetic [2]. Thus DNA copy number aberrations are crucial biological markers for cancer and possibly other diseases. The development of fast and reliable technology for detecting (the presence of) and pinpointing (the location of) such aberrations has become an important subject in biomedical research, with important applications to cancer diagnosis, drug development and molecular therapy.

Array comparative genomic hybridization (array CGH, or, aCGH) assay offers a high-throughput approach to measure the DNA copy numbers across the

* Correspondence should be addressed to epxing@cs.cmu.edu.

whole genome [3]. The outcome of an array CGH assay is a collection of log-ratio (LR) values reflecting the relative DNA copy number of test (e.g., tumor cells) versus control (e.g., normal cells) samples at all examined locations in the genome. Ideally, for diploid cells, assuming no copy-number aberration in the control and perfect measurement in the assay, the LRs of clones with k copies in the test sample can be exactly computed. It is noteworthy that among all possible magnitudes of k, usually only a few need to be distinguished, such as 0, 1, 2, 3, and collectively all integers that are greater (often significantly greater) than 3. These are typical copy numbers that reflect distinct cytogenetic mechanisms of chromosome alteration and rearrangements, and hence they are commonly referred to as *gene dosage states*: deletion, loss, normal, gain, and amplification.

Manual annotation of gene dosage tedious and inaccurate due to various reasons, such as impurity of the test sample (e.g., normal cell contaminations), intrinsic inhomogeneity of copy numbers among defective cells, variations of hybridization efficiency, and measurement noises arising from the high-throughput method [4]. Numerous computational methods have been developed for efficient and automated interpretation of array CGH data. Earlier methods used value-windows defined by hard thresholds to determine gene dosage state for each clone based on noisy LR measurement (e.g. [5]). However, these methods suffer from high false positive rate and low coverage (see Sec. 3.1). Recent developments resort to more sophisticated statistical modeling and inference techniques to interpret aCGH data. Based on the underlying statistical assumptions on signal distribution adopted by these methods, they largely fall into four categories: mixture models, regression models, segmentation models and spatial dynamic models. *Mixture models* [6] assume that the LR measurements of all the clones in an aCGH assay are *independent* samples from an underlying distribution consisting of multiple components, each corresponding to a specific gene dosage state. *Regression models* [7, 8] try to fit the noisy LRs with a smooth intensity curve over the chromosome to facilitate detection of gene dosage change via visual inspection which are only suitable for data denoising and visualization, rather than explicitly predicting the discrete dosage state underlying the LR signals. *Segmentation models* [9–15] directly search for breakpoints in sequentially ordered LR signals so that the resulting LR segments have the minimum within-segment signal variations. However, this segmented clone sequences suffer from state "over-representation", in which numerous spurious states without apparent biological meanings are uncovered for the segments. *Spatial dynamic models* solve the problems of dosage-state annotation and clone-sequence segmentation under a unified model for array CGH data. Fridlyand *et al.* [4] proposed a spatial dynamic framework that models the LR sequence as the output of a hidden Markov model (HMM) that governs the distribution of the dosage-states along the chromosome. Marioni *et al.* [16] extended this model by considering the distances between adjacent clones when modeling the transition matrix in HMM. Broet and Richardson [17] developed a Bayesian HMM by allowing the mixture weights to be correlated for neighboring genomic sequences on a chromosome.

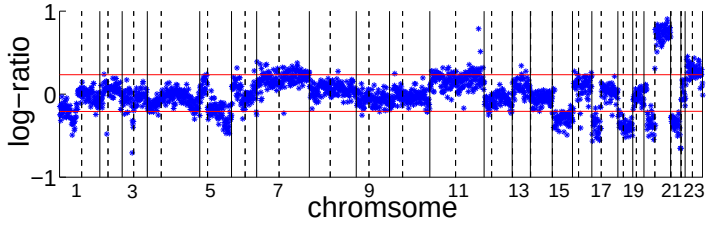

Fig. 1. The LR values (blue dots) of genome X77 from Nakao *et al.* [5]. The solid vertical lines delineate the boundaries between chromosomes; and the dashed vertical lines indicate the position of the centromere of each chromosome. The red horizontal lines indicate the thresholds used by Nakao *et al.* [5] to decide clones dosage states. Clones within these two lines are predicted to be normal state.

More recently, Shah *et al.* [18] proposed a new Bayesian HMM model that integrates prior knowledge of DNA copy number polymorphisms (CNPs).

This progress notwithstanding, the computational methods for aCGH analysis developed so far are still limited in their accuracy, robustness and flexibility for handling complex aCGH data, and are inadequate for addressing some of the deep biological and experimental issues underlying aCGH assay. Take the whole-genome aCGH data displayed in Fig. 1 as an example. Overall, the LR signals are highly fluctuating, but exhibit visible spatial auto-correlation patterns within the chromosomes. A caveat of the mixture-model-based or threshold-based methods is that they are very sensitive to such random fluctuations of the LR signals because they treat each measurement as an independent sample and ignore spatial relationships among clones. This could lead to highly frequent dosage-state switching (e.g., alternating back and forth between gain and loss, as we will show in our results) within short genetic distances, which is biologically implausible. A number of recent methods, particularly the spatial dynamic models based on HMM, have offered various ways to address this issue, which have significantly improved the performance of computational array CGH analysis.

Nevertheless, a key limitation of the HMM-based methods is that they all assume invariance of the true hybridization signal intensity alone chromosome for each dosage state, which is not always satisfied in real data. As shown in Fig. 1, an outstanding feature of the spatial pattern of the LR signals is that, within each chromosome, there exists both *segmental patterns* that are likely due to change of the copy number of the corresponding region, and *spatial drift* of the overall trend of the LR intensities along the chromosome. For example, in chromosome 4, the LR signals along the sequence of clones are not fluctuating around a baseline (presumably corresponding to a certain dosage state) that is invariant along the chromosome; instead, it is apparent that the baseline itself first has an increasing trend from left to right on 4p and into 4q, and then turns to a decreasing trend along the rest of 4q. Visually, there is not many abrupt breakage points that would signal a dosage-state alteration alone this continuously evolving sequence of LRs. But an HMM approach, which models

spatially-dependent choices among different copy-number state, each associated with an invariant distribution of LR values, can fail to capture the spatial drift of LRs over chromosome region with the same copy number as shown in Sec. 3.

Rather than reflecting the discrete change of copy numbers of the clones, the non-random spatial trend of LR signals possibly reflects a continuous change of the biophysical properties and hybridization quality along the chromosome. As discussed in [2], the intensity of the hybridization signal of each clone is affected by a number of factors such as base compositions of different probes, the proportion of repetitive content in sequence, the saturation of array, divergent sequence lengths of the clones, reassociation of double-stranded nucleic acids during hybridization, and the amount of DNA in the array element available for hybridization. These factors may further contribute random or correlated stochasticity of the LR values on top of the content-derived spatial drift. Pinkel and Albertson [2] reported that signal intensity may vary by a factor of 30 or more among array elements even if there are no copy-number changes. These complexities present in real aCGH data render extant models based on fixed state-specific LR distributions, such as an HMM, incapable of making accurate or robust state prediction.

Another problem that affects all the approaches discussed above lies in the calibration of signals across chromosomes and across individuals. As observed from Fig. 1, the mean and the variance of the LRs, and their spatial trends vary significantly from chromosome to chromosome, and more so from individual to individual (not displayed in the Figure), due to reasons possibly beyond copy number differences. This makes measurements from different individuals and/or for different chromosomes difficult to compare. Engler *et al.* [19] recently proposed a parameter sharing scheme for a Gaussian mixture model for genetic variability between and within chromosomes. In the new statistical model for aCGH data presented below, we introduce more careful treatments, which employ different parameter sharing scheme for effects shared among different chromosomes in the same individual (e.g., state baselines) and effects common to the same chromosomes in different individuals (e.g., signal dynamics).

In this paper, we introduce a new method *Genome Imbalance scanner* (GIMscan) for computational analysis of aCGH data. GIMscan employs a more powerful spatial dynamic model, known as switching Kalman filters (SKFs) [20], to jointly capture the spatial-trends of evolving LR signals along chromosomes, and spatially dependent configuration of gene dosage states along chromosomes. Unlike an HMM, which captures all the stochasticities in LRs with invariant dosage-specific distributions, an SKF breaks the accumulation of the stochasticities into two stages: 1) the *hybridization stage*, which involves physical sensory of clone-copies from the digested chromosomes, during which the spatial trend of DNA content and its biophysical properties, saturation effects, etc., can cause stochastic spatial drift of the mass of the hybridized material; 2) the *measurement stage*, which involves acquisition of the readings of fluorescence intensity of each clone, during which errors from reagents, instruments, environment, personal effects, etc., can cause another layer of random noises on top of the hybridiza-

tion signal. Under the SKF, we model the variations in the hybridization stage using dosage-state-specific continuous dynamic processes, akin to the regression approach discussed above. These hybridization intensities can be understood as the "true" *sensory signals* in an aCGH assay, which are unobservable to the examiner. We refer the sequence of hybridization intensities following such a linear dynamic model as a *hybridization trajectory*. Given the hybridization trajectory, we model the random noise from the measurement stage by a conditional Gaussian distribution whose mean is set by the sensory signal which evolves over each clone according to the trajectory. Overall, for each dosage state, we have a unique linear dynamic model for the sensory signals and a Gaussian emission model for their corresponding noising measurements. This model is known as a Kalman filter. To model changes of dosage-state along the chromosome, we follow the HMM idea to set up a hidden Markov state-transition process, but in our case not over state-specific distributions of LRs with fixed means, but over state-specific Kalman filters over both the observed LR measurements and the unobserved sensory signals for each clone.

On both simulated and experimental aCGH data, GIMscan has shown superior performance over other approaches such as HMM or mixture-model based threshold methods, being able to handle a number of complex LR patterns beyond the recognition power of reference models. We applied our methods to a whole-genome aCGH assay of 125 primary colorectal tumors [5], and constructed a high-quality genome-level gene dosage alteration map for colon cancer.

2 SKF Model and Adaptation to aCGH Analysis

For each specific gene dosage state, we model the spatial drift of its hybridization signal intensities using a hidden trajectory and model the uncertainty in LR measurements using a zero-mean Gaussian noise. This corresponds to a standard dynamic model named Kalman filter (KF). Observed LR values arise as a mixture of the outputs of state-specific Kalman filters. The mixing proportion, modeled as latent variables indicating gene dosage states, is also spatially dependent as captured by a Markov state-transition process (or switching process). Now we have multiple Kalman filters controlled by a dynamic switching process, which can be formulated as a factored switching Kalman filters (SKF). Our proposed method, GIMscan (Genome IMbalance SCANner), adopts the SKF model to whole-genome analysis of aCGH data by allowing a parameter sharing scheme among multiple chromosomes and multiple individuals which makes best use of data. In this section, we first introduce the SKF model and its parameters, then discuss the approximate inference algorithm for joint dosage-state annotation and clone-sequence segmentation. Model selection and further extension of the model are covered briefly at the end of this section.

Figure 2(a) illustrates the Kalman filter for a specific dosage state m, which is a linear chain graphical model with a backbone of hidden real-valued variables (denoted by $X_{1:T}^{(m)}$) emitting a series of real-valued observation (denoted by $Y_{1:T}^{(m)}$). The trajectory of hidden variables is linear and subject to Gaussian

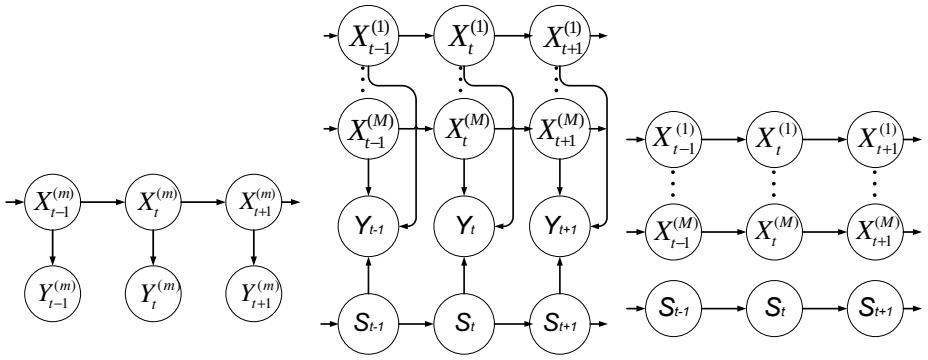

Fig. 2. (a) Graphical structure of dosage-state-specific Kalman filter for dosage state m. $X_t^{(m)}$ is the hidden variable at clone t on the trajectory, and $Y_t^{(m)}$ representing the corresponding observed variable of the Kalman filter. (b) Graphical structure of the switching Kalman filter (SKF) model. The model consists of M linear chains from Kalman filters ($X_{1:T}^{(1:M)}$), a Markov chain of switching processes ($S_{1:T}$) and a series of observed variables ($Y_{1:T}$). (c) Graphical structure of the uncoupled model which represents the tractable subfamily of distributions to approximate the true posterior of the SKF model.

noise which reflects the evolving hybridization signal intensities. The emission model imposes a Gaussian noise arising in the measurement stage on each hidden variable to generate the LR ratio at each position (clone). This model for a specific dosage state m can be formulated as $P(X_t^{(m)}|X_{t-1}^{(m)}) \sim \mathcal{N}(a^{(m)}X_{t-1}^{(m)}, b^{(m)})$, $P(Y_t^{(m)}|X_t^{(m)}) \sim \mathcal{N}(X_t^{(m)}, r)$.

The parameters $a^{(m)}, b^{(m)}, r$ are all position-invariant; r determines the degree of uncertainty in observation measurements. We also assume the initial value of the hidden trajectory, $X_1^{(m)}$, is distributed normally: $P(X_1^{(m)}) \sim \mathcal{N}(\mu^{(m)}, \sigma^{(m)})$. All the variables and parameters are univariate. The computation of posterior distributions of the hidden variables given the observation is tractable because of the conjugacy of the normal distribution to itself. This computation will be part of the inference procedure discussed later in which we decouple the SKF model to a number of tractable linear chains.

Given the dosage-state-specific Kalman filter for M dosage states, a switching Kalman filters generates the LR value at each position from one of the outputs: $Y_t = \sum_{m=1}^{M} Y_t^{(m)} S_t^{(m)}$, where S_t is the M-dimensional multinomial switching variables for clone t following $1 \times M$ binary coding scheme. The discrete switching process $S_{1:T}$ evolves according to Markov dynamics, with initial state distribution parameterized by π and state transition matrix Φ: $S_1 \sim \text{Multinomial}(1, \pi), P(S_t^{(m)} = 1|S_{t-1}^{(n)} = 1) = \phi_{mn}$. We could save the variables $Y_t^{(1:M)}$ and generate the observation directly from the M hidden lin-

ear Gaussian trajectories as $P(Y_t|X_t^{(1:M)}, S_t) \sim \mathcal{N}\left(\sum_{m=1}^{M} X_t^{(m)} S_t^{(m)}, r\right)$. The graphical structure of the SKF model is shown in Fig. 2(b).

To facilitate dosage state annotation and clone-sequence segmentation, the posterior distribution $P(S_t|Y_{1:T})$ need to be computed for $t = 1, \ldots, T$. However, exact computation of this posterior probability is intractable. We employed an algorithm [21] which approximates the posterior distribution $\mathcal{P}$ with a parameterized distribution $\mathcal{Q}(\mathbf{v})$ from some tractable subfamily of distributions. It iteratively updates the values of variational parameters $\mathbf{v}$ to minimize the KL divergence between the approximate posterior distribution and the true posterior distribution. The choice of the tractable subfamily for the SKF model is a discrete Markov chain and M uncoupled KFs (Fig. 2(c)). Two sets of variational parameters are introduced for the Markov chain and KFs respectively. Their updates can be carried out using fix-point equations [21], which maintain or increase a lower bound of log likelihood of the model and usually converge in a few iterations. Fast rate of convergence is mainly due to low data dimension.

Parameter estimation is performed under the EM framework. The E step employs the variational inference algorithm to find the best approximate posterior via iterative updates of the variational parameters. The M step reestimates the model parameters Θ to maximize the same lower bound of log-likelihood in variational inference. This reestimation can be performed exactly by zeroing the derivatives with respect to the model parameters. Parameter estimation is implemented by a coordinate ascent procedure.

Now we have introduced the SKF model and its parameters $\mu^{(m)}, \sigma^{(m)}, a^{(m)}$, $b^{(m)}, r, \pi$ and ϕ, each of which delineates one property behind the aCGH data. $\mu^{(m)}$ and $\sigma^{(m)}$ are the mean and variance of Gaussian distribution of the starting clone on hidden trajectory for dosage state m. $a^{(m)}$ and $b^{(m)}$ determine the transition model of that trajectory which dictate the spatial drift of the signal intensities. r is the variance of the Gaussian accounting for the noise introduced in the experiment stage, and is independent of the hidden dosage-state. Lastly, π and Φ are initial state parameters and transition matrix for the discrete switching process between different dosage states.

In the settings for a whole-genome analysis, the aCGH dataset are collected from experimental data of J individuals, the genome of which consists of K chromosomes, and chromosome k contains T_k clones. The LR values $Y_{1:T,j,k}$ on individual j, chromosome k are generated by an SKF model with hidden trajectory $X_{1:T,j,k}$ and switching states $S_{1:T,j,k}$.

We are now ready to describe the parameter sharing scheme in GIMscan for the analysis of whole genome aCGH data. We consider two groups of parameters. Firstly, we let $\mu^{(m)}$, $\sigma^{(m)}$, r, π and Φ be shared across all chromosomes of one particular individuals. Mainly due to the normal cell contamination, the magnitude of starting value for the trajectory of one particular state varies across different individuals. Different $\mu^{(m)}$ and $\sigma^{(m)}$ for different individuals can account for this "un-normalized" starting value of the trajectory of one state. r is also shared by chromosomes from one individual because one individual corresponds to one experiment, and different experiments may have different noise

levels. π and Φ are shared by one individual because the number of dosage states are individual-specific: different individuals may have different number of dosage states. The second group of parameters, $a^{(1:M)}$ and $b^{(1:M)}$, is assumed to be shared by one particular chromosome of all individuals, because the physical-chemical properties (e.g. the base composition) of one particular chromosome of different individuals (the same tumor cell line of same species) are very similar. This similarity leads to similar hybridization signal intensity over a chromosome.

The maximum number of dosage states M one individual can have remains to be determined. we employ Gaussian mixture model using penalized likelihood criteria such as AIC to select the number of states for each individual.

3 Experiments and Results

We tested the performance of GIMscan on simulated aCGH data and real data with complex aCGH patterns to demonstrate the working principle and general trends of our method in gene dosage prediction, and to evaluate our prediction quality under nontrivial genome imbalance and hybridization scenarios. The benefit of applying a sophisticated hybrid stochastic model to capture both discrete (e.g., changing DNA copy number) and continuous (e.g., varying hybridization efficiency) latent spatial trajectory underlying noisy aCGH measurements is evidenced in each level of genetic scales we have analyzed.

3.1 Simulated aCGH Data

We first validate GIMscan on simulated aCGH datasets, which mimic typical spatial patterns of LR sequences in real aCGH assays, and allow a quantitative assessment of model performance based on known underlying gene dosage states in the simulation.

In our simulation experiments, three methods—threshold, HMM as in [4], and GIMscan—were tested on 12 datasets simulated with different settings of two parameters, the Gaussian emission variance r and the KF transitional variance b (see Section 2). These two parameters represent the two sources of the overall noise in the data: r reflects the quality the LR measurements in an aCGH experiment, whereas b reflects the variability of the hybridization signal intensity along the chromosome. Our datasets correspond to three different values of r, ranging from low, to medium, high; and four values of b also spanning a significant range (see Fig. 3). For each combination of r and b, a total of 100 LR sequences each containing 100 clones were generated. For each sequence, we simulated a random 5×5 stochastic matrix, T, for modeling transitions between gene dosage states, and T was set to allow both short and long stretch of gene dosage alterations, but not high-frequency oscillations between different states. All three methods were applied to each dataset to infer the gene dosage states underlying the simulated LRs, and the experiments were repeated 100 times. Fig. 3 summarizes the medians, quantiles and ranges of the prediction

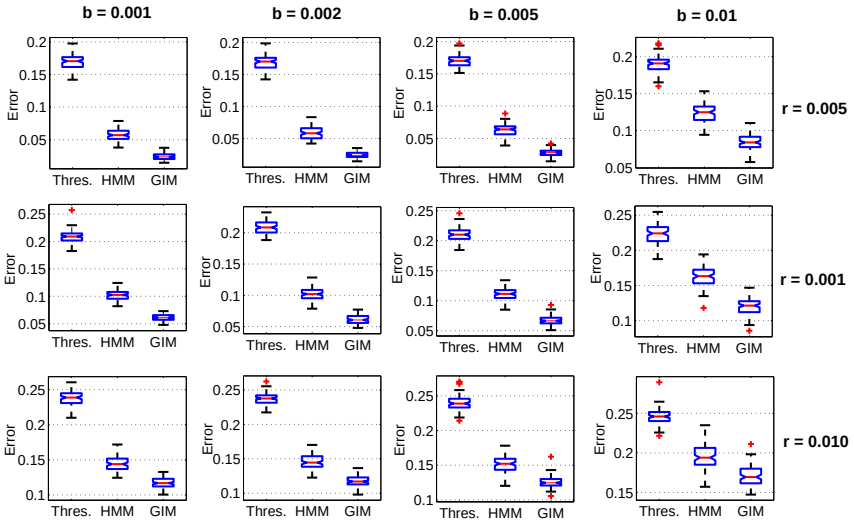

Fig. 3. Performance of gene dosage state prediction on simulated aCGH datasets. Each row corresponds to an emission noise, and each column corresponds to a Kalman filter transitional variance. In each case, we plot the results of threshold method (Thres.), HMM and GIMscan (GIM). The red line represents the median, and the blue box indicates upper and lower quantiles. The black bars are the range of the error rate. Outliers are plotted by "+".

error rates by different methods under various parameter settings. Consistently, GIMscan outperformed the other two methods by a significant margin.

As an illustration of the advantage offered by the SKF model adopted by GIMscan, and the effectiveness of our inference algorithm, Fig. 4 and Fig. 5 show two examples of GIMscan's performance in the simulated datasets. The first example concerns "high-quality" aCGH records simulated with low measurement noise ($r = 0.001$) over 100 clones switching between two gene dosage states both with low spatial drift in their corresponding true hybridization intensities ($b = 0.001$) (Fig. 4(a)). Figure 4(b) presents the inferred gene dosage state and the inferred dosage-state-specific "trajectories" (i.e., the latent dynamical trend captured by each KF) of the latent true hybridization intensities underlying the observed LR sequence shown in Fig. 4(a). As shown in this illustration, each inferred latent trajectory indeed represents a smoothed and spatially changing baseline of the LR signals corresponding to a particular dosage state; and all inferred trajectories agree well with the true trajectories of hybridization intensities used for simulating the observed LR signals. As a result, the inferred switching process over these trajectories gives a highly accurate prediction of the gene dosage states underlying the LR sequence. GIMscan can also estimate the confidence intervals (i.e., standard deviation) of the inferred hybridization trajectories, as shown in Fig. 4(c).

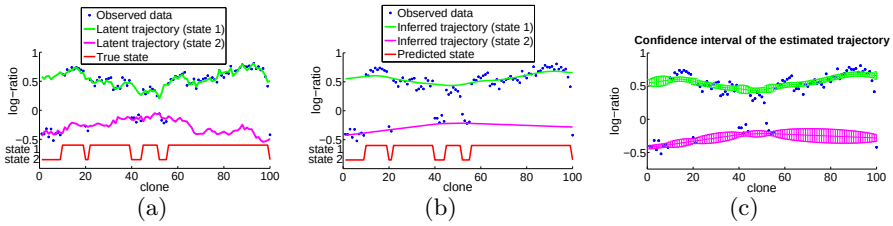

Fig. 4. (a) Simulated data (blue dots), two latent trajectory (green and pink), and switching process (red). The length of the simulated data is 100 clones. (b) Simulated data (blue dots), inferred trajectory (green and pink) and inferred switching process (red). (c) The confidence intervals of the inferred trajectories.

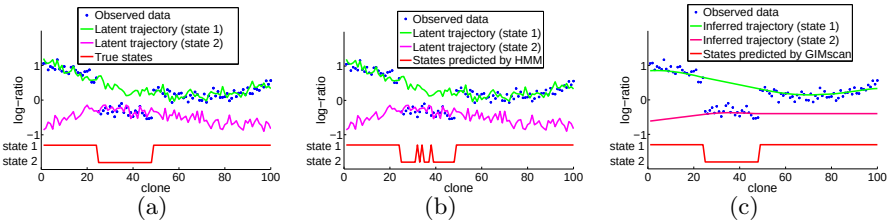

Fig. 5. (a) Simulated data (blue dots), two latent trajectories (green and pink), and switching process (red). The length of the simulated data is 100 clones. (b) Given the observed data in (a), the red line is the state predicted by HMM. (c) Given the observed data in (a), the green and pink curves are trajectories inferred by SKF, and the red line is the switching process inferred by SKF.

Another example shown in Fig. 5 concerns low-quality, arguably more realistic aCGH records simulated with high measurement noise ($r = 0.01$) and severer spatial change ($b = 0.01$) in the true hybridization trajectories. The combined effects of high measurement noise and high spatial variance of the hybridization trajectories are excepted to lead to misassignment of gene dosage state due to inaccurate estimation of the dosage-state-specific hybridization intensities when spatial trajectory of the hybridization intensities is ignored. Note that the trajectories in Fig. 5(a) of both dosage states are not flat, which reflect severe spatial drift of hybridization signal intensity within each state. When assuming spatial invariance of dosage-state-specific signal distribution, the unflatness of both trajectories can cause the estimated mean of LR signals to be highly biased (e.g., higher for state 1, and lower for state 2), and their variances to be significantly greater than the actual fluctuation. Consequently, the estimated dosage-specific signal distributions can be seriously overlapping, causing the LR signals from two states hard to distinguish. Fig. 5(b) shows exactly this effect, on the quality of state estimation by an HMM model. Whereas the SKF model underlying GIMscan readily mitigates this effect, and produces the correct estimation.

3.2 Real aCGH Data with Diverse Spatial Patterns

Now we present case studies of selected real aCGH data with a diverse spectrum of spatial patterns. Our dataset was obtained from an online repository of whole-genome aCGH profiles of 125 colorectal tumors originally studied in [5]. This dataset was found to contain highly stochastic LR measurements with severe spatial variance and drifts along the chromosomes, and bear rich cohorts of genome imbalance patterns. Such complications present a great challenge to naive algorithms for gene dosage inference, and are thus particularly suitable for evaluating our proposed method.

Given an aCGH profile, GIMscan first employs a k-nearest neighbor regression procedure (e.g., $k = 3$) to impute the missing values in the LR records. Then it fits the processed data with a Gaussian mixture based on maximum likelihood estimation, and performs model selection based on AIC to determine the total number of gene dosage states, M (which is constrained between 1 to 5), for each individual. Afterwards, the number of component KFs (i.e., dosage-state-specific hybridization trajectories) in GIMscan is set to be M, and the mean of the starting clone of each KF takes on the mean of a component in the estimated Gaussian mixture as initial value. Note that with this setup, we still need to establish the exact mapping between the KFs inferred by GIMscan and the possible gene dosage states, namely deletion, loss, normal, gain, and amplification. Since GIMscan provides estimations of the hybridization trajectories of each KF, we follow a straightforward statistical and biological argument and determine the corresponding dosage-state of each trajectory based on the relative mean-values of the estimated true hybridization intensities of all clones.

For comparison, we re-implemented the HMM methods according to Fridlyand *et al.* [4], with modest extension (i.e., parameter sharing) so that it can be applied to whole genome CGH profiles covering multiple chromosomes. Following [4] AIC is also used for model selection for the HMM.

The dataset we studied contains a total of $\sim 2.75 \times 10^5$ LR measurements from 23×125 chromosomes (i.e., 125 human genomes). Here we first present a small-scale case study of three representative chromosomes, each containing a typical spatial pattern for the LR sequence that was found to be difficult to analyze by conventional methods. For convenience, we refer to these patterns as, flat-arch, step, and spike, respectively, according to their shapes in the LR intensity plots (Fig. 6).

Pattern I: Flat-Arch Figure 6(a)(b) displays the LR measurements from chromosome 4 of individual X77, this pattern is marked by lower magnitudes of LRs at the two telomere regions of the chromosome and elevated magnitudes in the central region. Locally (i.e., along the plotted chromosomal region), there is a continuous trend of spatially evolving hybridization intensity along the chromosome, and there are few abrupt breakage points that would signal a dosage-state alteration. But due to the high dispersion of LR values as a result of such a spatial drift, methods based on invariant state-specific hybridization intensity, such as the HMM, would either fit the observed LR values with one biased and

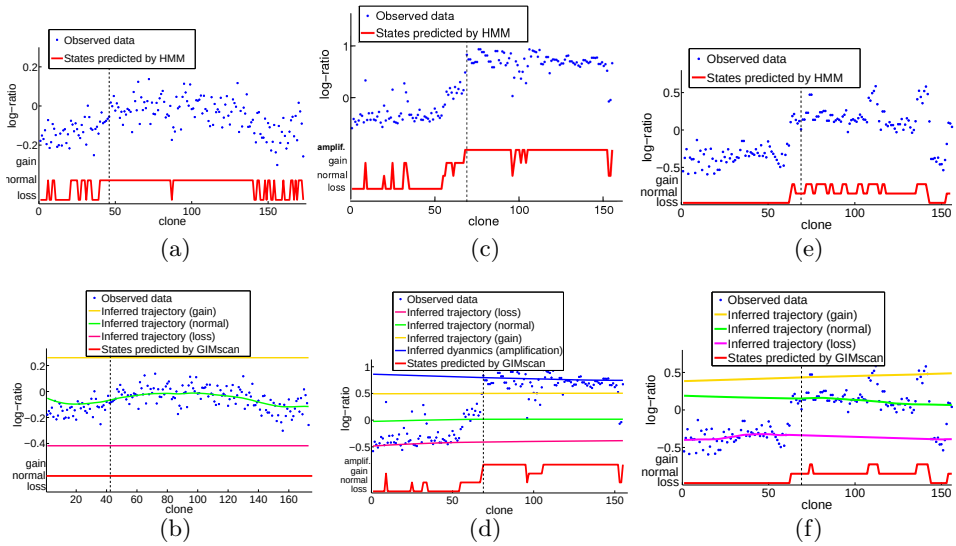

Fig. 6. Three typical spatial patterns for the LR sequence which were found to be difficult to analyze by convenctional methods: (a)(b)Flat-Arch pattern; (c)(d)Step pattern; (e)(f) Spikes pattern. (a)(c)(e) shows states predicted by HMM (red). In (b)(d)(f) the pink, green and yellow curves are inferred trajectories for loss, normal and gain, respectively. Red solid line indicates states predicted by GIMscan. Centromere position is indicated by dashed vertical line in these two plots.

high-variance Gaussian distribution, or split the LRs with two highly overlapping Gaussians. These caveats could seriously compromise the quality of gene dosage state estimation. Figure 6(a) shows the dosage estimation by an HMM fitted on this chromosome. The outcome suggests heavy oscillations between two dosage states throughout the chromosome, which is biologically implausible. Figure 6(b) shows the dosage state sequence and dosage-state-specific trajectories underlying chromosome 4 of individual X77 inferred by GIMscan. A whole-genome fitting resulted in three estimated dosage-states. On this particular chromosome, the trajectories of the loss and gain states (the pink and yellow curves, respectively) were not matched to any observations, and the entire region is determined to be corresponding to a normal state whose hybridization intensity varies along the chromosome (the green curve). Indeed, a more global visual inspection of these Flat-Arch patterns in the context of whole aCGH profile often reveals that the flat-arch shape in the LR-plots often merely reflects modest (but spatially correlated) change of the LR magnitude most likely within a single dosage state.

Pattern II: Step This pattern is typical when there appears to be a quantum change of LR magnitudes from one to the other end of the chromosome, but the boundary of the change is not sharp and the overall sequence is moderately noisy, such as shown in Fig. 6(c)(d) which is taken from chromosome 8 from individual

X265. In addition to the step, this sample also harbors a number of local spikes and short regions potentially implying dosage-state alterations. Via AIC model-selection, the HMM adopted four dosage state when processing this data. The states predicted by HMM are shown in Fig. 6(c). As can be seen, the results are reasonable, except that several positions near clone 100 contains highly frequent switching between states. The dosage state sequence and dosage-state-specific trajectories inferred by GIMscan are shown in Fig. 6(d). Note that there is a slightly decreasing trend in the trajectory corresponding to the amplification state. While the gain and normal trajectories correspond to only a few clones on this chromosome, a genome-level parameter sharing scheme adopted by GIMscan enables them to be reliably estimated, and thereby leads to plausible prediction of point changes on isolated clones (e.g., clone 97 and 152).

Pattern III: Spikes Spikes are a typical pattern often accompany other patterns, such as steps. It is marked by short sequences, sometimes singletons, of elevated or attenuated LR measurements along the chromosomes. Figure 6(e)(f) shows such an example from chromosome 8 of individual X318. In this chromosome, the copy-number loss was apparent on 8p arm, while three spikes (around clone 75, 110 and 140) were visible on 8q arm. These spikes correspond to the gain state with a large measurement variance. Figure 6(e) shows the states predicted by HMM. Although HMM correctly predicted the states on 8p, it predicted more clones on 8q to be gain state. However, by our visual check, some clones (e.g. around clone 79, 96) should have been classified to be normal state. The possibly faulty predictions of gain states resulted from the large variance of the spikes estimated by the HMM. GIMscan correctly detected and annotated the spikes, as well as giving convincing predictions on other clones (Fig. 6(f)). Compared to the case for the same chromosome (i.e., no. 8) from individual X265 shown in Fig. 6(c)(d), where four dosage-state-specific trajectories were determined, here we uncovered only three states for chromosome 8. This is because model selection for SKF in this individual based on the whole-genome aCGH only identifies three states—normal, loss and gain. Parameter-sharing was adopted by GIMscan for all chromosomes in this individual, and leads to three common trajectories. Comparing Fig. 6(d) with Fig. 6(f), one can notice that the elevates of the trajectories corresponding to the same dosage state (e.g., normal) can be quite different across individual, which is likely due to some unidentified systematic error or hybridization-efficiency difference across individuals. The parameter-sharing scheme adopted by GIMscan (i.e., sharing dosage-state-specific trajectories across chromosomes within individual, but not across individual) provides a reasonable strategy to tackle such variations.

We finally used GIMscan for populational analysis of Nakao *et al.*'s [5] dataset. Overall, over the 125 genomes each examined at ∼2200 clones uniformly distributed in the genome, on average each genome have 19.18% (or 407) of the clones suffered either gain or loss (9.25% and 9.94%, respectively), and another 1.33% of the clones were hit by amplification or deletion (0.93% and 0.4%, respectively). The whole-genome spatial spectrum of GIM rates over the entire

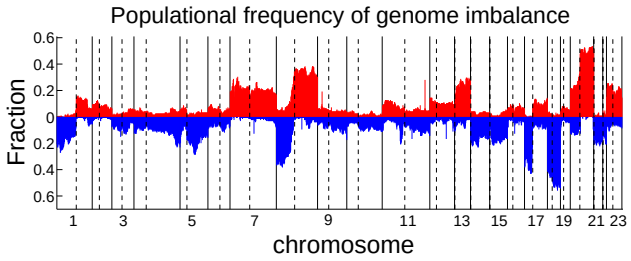

Fig. 7. Overall frequency of DNA dosage state alteration over entire genomes of 125 individuals. Blue bars represent clones with DNA copy number loss or deletion, whereas red bars for DNA copy number gain or amplification. Solid vertical lines show boundaries between chromosomes; dashed vertical lines show centromeres of chromosomes.

study population is displayed in Fig. 7. As can be seen, the population rates of gain and amplification of clones in chromosome 7, 8q, 13q, 20q and 23 were significantly higher than those of the other regions, suggesting possible presence of proto-oncogenes in these regions. Likewise, the population rates of loss and deletion in chromosome 1p, the distal-end of 4q, 5q, 8p, 14, 15, 17p, 18, and 21, were significantly higher than those of the other regions, suggesting possible presence of tumor suppressor genes in these regions.

4 Discussion

An important issue for the success of GIMscan is the parameters initialization. Our experience with GIMscan shows that the initial values for π and ϕ may be fairly arbitrary, while the initial values for μ and r are more essential. We can employ the Gaussian mixture to cluster the data points into M clusters. The mean LR value of one cluster is used as the initial value for the starting mean of the corresponding trajectory. The initial value of r can be determined similarly. We initialized a and b with some constants: a was fixed to 1 and b was fixed to 10^{-2}. σ was given the same initial value as r.

Acknowledgments. This material is based upon work supported by the National Science Foundation under Grant No. 0523757, and and by the Pennsylvania Department of Health's Health Research Program under Grant No. 2001NF-Cancer Health Research Grant ME-01-739. E.P.X. is also supported by a NSF CAREER Award under Grant No. DBI-0546594.

References

1. Diep, C.B. *et al.* : Genome characteristics of primary carcinomas, local recurrences, carcinomatoses, and liver metastases from colorectal cancer patients. Mol. Cancer. **3(1)** (2004) 6

2. Pinkel, D., Albertson, D.G.: Array comparative genomic hybridization and its applications in cancer. Nat. Genet. **37** (2005) S11–S17
3. Pinkel, D. *et al.* : High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. Nat. Genet. **20** (1998) 207–211
4. Fridlyand, J., Snijders, A.M., Pinkel, D., Albertson, D.G., Jain, A.N.: Hidden Markov models approach to the analysis of array CGH data. J. Multivariate Anal. **90(1)** (2004) 132–153
5. Nakao, K. *et al.* : High-resolution analysis of DNA copy number alterations in colorectal cancer by array-based comparative genomic hybridization. Carcinogenesis. **25(8)** (2004) 1345–1357
6. Hodgson, G. *et al.* : Genome scanning with array CGH delineates regional alterations in mouse islet carcinomas. Nat. Genet. **29** (2001) 459–464
7. Eilers, P., De Menezes, R.: Quantile smoothing of array CGH data. Bioinformatics. **21(7)** (2005) 1146–1153
8. Hsu, L., Self, S., Grove, D., Randolph, T., Wang, K., Delrow, J., Loo, L., Porter, P.: Denoising array-based comparative genomic hybridization data using wavelets. Biostatistics. **6(2)** (2005) 211–226
9. Myers, C.L., Dunham, M.J., Kung, S.Y., Troyanskaya, O.G.: Accurate detection of aneuploidies in array CGH and gene expression microarray data. Bioinformatics. **20(18)** (2004) 3533–3543
10. Olshen, A.B., Venkatraman, E.S., Lucito, R., Wigler, M.: Circular binary segmentation for the analysis of array based DNA copy number data. Biostatistics. **5(4)** (2004) 557–572
11. Jong, K., Marchiori, E., Meijer, G., Vaart, A.V., Ylstra, B.: Breakpoint identification and smoothing of array comparative genomic hybridization data. Bioinformatics. **20(18)** (2004) 3636–3637
12. Picard, F., Robin, S., Lavielle, M., Vaisse, C., Daudin, J.J.: A statistical approach for array CGH data analysis. BMC Bioinformatics. **6(1)** (2005) 27.
13. Hupe, P., Stransky, N., Thiery, J.P., Radvanyi, F., Barillot, E.: Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. Bioinformatics. **20(18)** (2004) 3413–3422
14. Wang, P., Kim, Y., Pollack, J., Narasimhan, B., Tibshirani, R.: A method for calling gains and losses in array CGH data. Biostatistics. **6(1)** (2005) 45–58
15. Daruwala, R.S., Rudra, A., Ostrer, H., Lucito, R., Wigler, M., Mishra, B.: A versatile statistical analysis algorithm to detect genome copy number variation. PNAS. **101** (2004) 16292–16297
16. Marioni, J.C., Thorne, N.P., Tavare, S.: BioHMM: a heterogeneous hidden Markov model for segmenting array CGH data. Bioinformatics. **22(9)** (2006) 1144–1146
17. Broet, P., Richardson, S.: Detection of gene copy number changes in CGH microarrays using a spatially correlated mixture model. Bioinformatics. **22(8)** (2006) 911–918
18. Shah, S.P., Xuan, X., De Leeuw, R., Khojasteh, M., Lam, W., Ng, R., Murphy, K.P.: Integrating copy number polymorphisms into array CGH analysis using a robust HMM. Bioinformatics. **22(14)** (2006) e431–e439
19. Engler, D.A., Mohapatra, G., Louis, D.N., Betensky, R.A.: A pseudolikelihood approach for simultaneous analysis of array comparative genomic hybridizations. Biostatistics. **7(3)** (2006) 399–421
20. Murphy, K.P.: Learning switching Kalman filter models. Compaq Cambridge Research Lab Tech Report 98-10. (1998)
21. Ghahramani, Z., Hinton, G.E.: Variational learning for switching state-space models. Neural Comput. **12(4)** (1998) 963–996