IADIS International Journal on Computer Science and Information Systems
Vol. 5, No.1, pp. 101-112
ISSN: 1646-3692

CONTINUOUS-TIME HIDDEN MARKOV MODELS FOR THE COPY NUMBER ANALYSIS OF GENOTYPING ARRAYS

Matthew Kowgier and Rafal Kustra. *Dalla Lana School of Public Health, University of Toronto, Toronto Canada.*

ABSTRACT

We present a novel Hidden Markov Model for detecting copy number variations (CNV) from genotyping arrays. Our model is a novel application of HMM to inferring CNVs from genotyping arrays: it assumes a continuous-time framework and is informed by prior findings from previously analyzed real data. This framework is also more realistic than discrete-time models which are currently used since it does not assume that CNV breakpoints occur at the genotyped loci. We show how to estimate the model parameters using a training data of normal samples whose CNV regions have been confirmed, and present results from applying the model to a set of HapMap samples containing aberrant SNPs.

KEYWORDS

Hidden Markov Models; Copy Number Variation; Markov Chain Monte Carlo; SNP genotyping arrays.

1. INTRODUCTION

In this paper we propose a novel application of a continuous-time hidden Markov model (CHMM) for interrogating genetic copy number (CN) information from genome-wide Single Nucleotide Polymorphism (SNP) arrays. DNA Copy number changes, either deletions or amplifications of a region of DNA, result in having less or more than the usual 2 versions of DNA sequence. Such DNA alterations constitute an important class of genetic mutations which are proving extremely useful in understanding the genetic underpinnings of many diseases and other phenotypic information (McCarroll and Altshuler, 2007). The Affymetrix Mapping 500K chip set (Affymetrix, 2006) is a pair of arrays that interrogate over 500,000 human SNPs. In the rest of the paper, we refer to this pair of arrays as a SNP array. While

SNP arrays were originally developed for genome-wide genotyping, the technology has also proven to be capable of producing copy number calls. The copy number analysis of SNP arrays consists of the following sequence of steps: (1) the preprocessing of the low-level (i.e. probe-level) data, which involves removing various biases that exist in the data (such as PCR fragment length); (2) single locus copy number estimation at each SNP location, which summarizes the DNA concentration; and (3), chromosome-wide modelling to infer *regions* of copy number changes, called copy number variations (CNVs). This paper focuses on the improvement of methodology for the third step.

Copy number data from normal cell lines is often characterized by long stretches of no variation, interspersed with typically small regions of CNVs. Hidden Markov models (HMMs) are well-suited to modelling such sudden changes in the data, enabling them to make good predictions of copy number along the genome. For this reason, HMMs are a commonly used technique for the genome-wide detection of CNVs. Indeed, numerous HMMs have already been proposed for the analysis of SNP arrays, including dChip (Lin et al., 2004) and VanillaICE (Scharpf, 2008). Both of these implementations of HMMs model the copy-number process as discrete with respect to the genomic location and are therefore called discrete-time HMMs (DHMMs). Furthermore, since DHMMs do not model the process between the observed locations they necessarily force copy number changes to occur at the observed SNP locations, which is an unrealistic assumption as pointed out by Stjernqvist et al. (2007). In reality, CNV breakpoints are likely to occur between the locations interrogated by the SNP arrays; even the most dense SNP arrays interrogate only a small fraction (less than 0.1%) of the genome. In this paper we adopt a continuous-time HMM in which the Markov process governing copy number changes along the genome is viewed as continuous in time, so that copy number changes can occur at any point along the genome. Another advantage of the continuous-time framework is that the uneven distances between the SNPs on the array are naturally taken into account. Such models have been previously used by Stjernqvist et al. (2007) in the context of array CGH data (Snijders et al., 2001). The copy number analysis of array CGH data differs from that of SNPs arrays in two important ways. First, since the endpoints from array CGH data are ratios of the signal from the test DNA sample to that of a reference DNA sample, it is difficult to determine whether a detected CNV occurred in the test or reference sample. Thus, such an approach is not effective when interested, as we are in this study, in finding germline CNVs – CNVs in normal cell lines, rather than tumor cell lines. Second, the probes are designed differently. For example, SNP arrays use short probes that produce allele-specific measurements, while array CGH use probes that are longer, non-allele-specific, and tend to overlap. These structural differences require different model specifications and, therefore, model fitting procedures.

In this paper we propose a fully Bayesian continuous-time HMM (CHMM) for the analysis of SNP arrays. We show that Bayesian copy number estimation addresses some shortcomings of the standard approach – the combination of the Baum-Welch (EM) algorithm for parameter estimation and the Viterbi algorithm for copy number estimation. We assess the methodology on a set of a previously verified aberrant loci and also with a simulation study consisting of 100 samples from chromosome 8 data. We end the paper with some concluding remarks and comments about future work.

2. BODY OF PAPER

2.1 Overview of our Procedure for Copy Number Determination

The SNP arrays produce a number of intensity values for each interrogated SNP. The description of the underlying technology and meaning of these values is beyond the scope of this paper; for further details, readers are asked to consult Kennedy et al. (2003) and the references therein. For the purpose of CN determination, a summary of the total intensity at each SNP, regardless of the underlying genotype present at the site, is needed. We used a popular procedure called Copy-number estimation using Robust Multichip Analysis (CRMA) (Bengtsson et al., 2008) that summarizes the total raw intensity data for each SNP. We refer to these single-locus, non-polymorphic, continuous, copy number estimates as raw CNs. Each raw CN is computed independently of other loci, in the sense that information from nearby loci is not utilized at this stage. One of the advantages of CRMA over other methods is that it utilizes information, if available, from multiple arrays (observations) to improve estimates at a given SNP location. In sum, sufficiently large raw CN estimates indicate evidence of a copy number gain, whereas sufficiently small raw CN estimates indicate evidence of a copy number loss.

2.2 Emission Distribution for the Raw CNs

We focus on the analysis of a single individual sample, so the data we analyze consists of a sequence of continuous raw CNs Y_i for $i = 1, \ldots, M$, where M is the number of SNPs. Additionally, we also know the physical location in bps of the observed SNPs, which we denote by d_i for $i = 1, \ldots, M$. Let C_i be the underlying copy-number value where $C_i = \{1,2,3\}$. The raw CNs are assumed to be generated from a conditional Gaussian model, whose parameters depend on the underlying and hidden, CN state. These Gaussians are usually called *emission distributions*. So, independently for all i,

$$Y_i|\mu_c, \sigma_c^2, C_i = c \sim N(\mu_c, \sigma_c^2). \quad (1)$$

Since regions with altered CN states are assumed to be of genetic length that usually encompasses more than one SNP site, a hidden Markov model is used to estimate Gaussian model parameters and hence the underlying CN states across each chromosome.

2.3 Titration and Human Population Data to set Hyper-Parameters of our Model

We used two previously published datasets to help specify (hyperparameters) prior distributions of our model parameters; see Section 2.8 for more details on how this is done.

The X chromosome titration data set (3X, 4X, and 5X) contains three artificially constructed DNA samples containing abnormal amplification of the whole X chromosome (aneuploidies). There are four replicates of each DNA sample. The aneuploidies are a X trisomy (presence of three copies of chromosome X); a X chromosome tetrasomy (presence of

four copies of chromosome X); and a X chromosome pentasomy (presence of five copies of chromosome X). These data were downloaded from the Affymetrix data resource center (Affymetrix) The Coriell Cell Repository numbers for these three cell lines are NA04626 (3X), NA01416 (4X), and NA06061 (5X). We used this data to specify the hyperparameters of the emission distribution.

McCarroll et al. (2008) report genomic coordinates for 1,320 copy number polymorphisms from the 270 HapMap samples. We used these data to inform about mean lengths of copy number deletions and amplifications, which are both hyperparameters of the transition intensity matrix.

2.5 The Copy Number Model

The copy number process records the number of copies of DNA at specific locations along the genome. We let $\{C(t)\}_{0<=t<=T}$ denote the unobserved copy number process of one sample which we wish to infer, where T is the length of the chromosome in bps. We allow the process to take three possible values: 1 (haploid), 2 (diploid) or 3 (triploid). This could easily be extended to include more states, such as 0- and 4-copy states. For convenience, we will denote the copy number at the observed SNP locations $(C(d_1), \ldots, C(d_M))$ by $(C_1, \ldots, C_M) = C$. Our goal is to infer C based on the observed data $Y = (Y_1, \ldots, Y_M)$.

We model C as a continuous-time Markov process. The continuous-time Markov process is parameterized in terms of a 3×3 transition intensity matrix of copy number changes $Q = \{q_{ij}\}_{i=1,2,3;j=1,2,3}$, where, for $i \neq j$, $q_{ij} > 0$, and $q_{kk} = -\sum_{l \neq k} q_{kl}$, so that $\sum_j q_{ij} = 0$ and $q_{kk} <= 0$. Unlike a discrete-time Markov process whose state transitions are defined in terms of transition probabilities, the continuous-time Markov process is defined in terms of its instantaneous transition intensities q_{ij}. The intensity q_{ij} represents the instantaneous risk of moving from state i to state j:

$$q_{ij} = \lim_{h \to 0} P(C(t+h) = j | C(t) = i)/h. \tag{2}$$

The complete specification of the transition intensity matrix is given by

$$Q = \begin{pmatrix} -(q_{12} + q_{13}) & q_{12} & q_{13} \\ q_{21} & -(q_{21} + q_{23}) & q_{23} \\ q_{31} & q_{32} & -(q_{31} + q_{32}) \end{pmatrix}.$$

This model assumes that, for example, the rate of deletion in a normal (2-copy) region is different than the rate of deletion in a amplified (3-copy) region. More specifically, this model specification is based on the reasonable belief that the rate of deletions will be larger from the normal state than from the amplified state (i.e. $q_{21} > q_{31}$), since deletions rarely follow regions of amplification.

The parameters of the Q matrix govern the occurrence of copy number changes along the chromosome. The model is shown graphically in Figure 1. Perhaps a better way of understanding the evolution of the Markov chain is through the time it spends in a state or, in the context of genomics, the number of bases pairs before a copy number change occurs. (Note that the number of bps before a copy number change occurs simply corresponds to the

length of the region; for example, the length of a copy number deletion.) Assuming that the chain begins in the diploid (2-copy) state, the chain stays in the diploid state for a length of time (or distance) that is exponentially distributed with rate parameter $\nu_2 = q_{21} + q_{23}$. Thus, under this model, the expected length of a diploid region is $1/\nu_2$. Once the stay in the diploid state is complete, the chain then moves to either the deleted (1-copy) state with probability $q_{21}/(q_{21} + q_{23})$ or the amplified (3-copy) state with probability $q_{23}/(q_{21} + q_{23})$. This process repeats itself, over and over. We denote the transition probabilities of the chain by $p_{ij} = q_{ij}/(\sum_{k \neq j} q_{ik})$ and the transition rates by $\nu_j = \sum_{k \neq j} q_{jk}$ for $j = 1, 2, 3$. Note that $p_{ii} = 0$ for $i = 1, 2, 3$; once the process leaves state i, it must proceed to a new state. Also $p_{13} = 1 - p_{12}$, $p_{23} = 1 - p_{21}$ and $p_{32} = 1 - p_{12}$. As we will see later, it is convenient to reparameterize the model in terms of the p- and ν-parameters. Under this parameterization for the CHMM, the collection of all parameters is $\theta = (\mu_1, \mu_2, \mu_3, \sigma_1^2, \sigma_2^2, \sigma_3^2, p_{12}, p_{21}, p_{31}, \nu_1, \nu_2, \nu_3)$ and we sometimes also distinguish between parameters governing the emission distribution, $\theta_E = (\mu_1, \mu_2, \mu_3, \sigma_1^2, \sigma_2^2, \sigma_3^2)$, and the parameters governing the Q-matrix, $\theta_Q = (p_{12}, p_{21}, p_{31}, \nu_1, \nu_2, \nu_3)$.

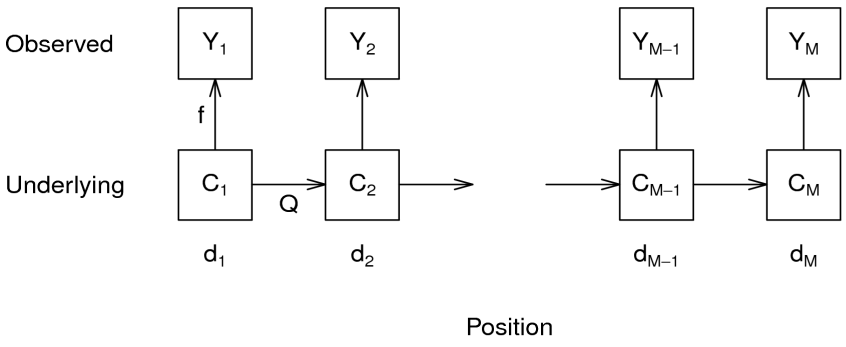

Figure 1. Graphical representation of the continuous-time HMM for SNP arrays.

2.6 Computing the Distance-d Transition Probability Matrix

In order to specify the likelihood of the model, $p(Y|\theta)$, we need to compute the corresponding distance-d transition probability matrix, $T(d)$. The matrix $T(d)$ is defined by the elements $T_{ij}(d) = P(C(t+d) = j | C(t) = i)$ for $i = 1, 2, 3$ and $j = 1, 2, 3$. This matrix can be derived from the $Q-$matrix by computing the matrix exponential of Q: $T(d) = \exp\{dQ\}$; see Ross (2003) for details. For our model, though, there are no closed-form solutions for the matrix exponential. So, instead, we assume that there is at most one transition between adjacent SNPs. This is a reasonable assumption as we do not expect more than one transition between adjacent SNPs due to the close proximity of SNPs on the arrays. With this assumption, if we let L_i be a random variable representing the length of the stay in

state i, then, assuming that we are in state i at SNP $j-1$, the probability that we are still in state i at SNP j is

$$\begin{aligned} T_{ii}(d_j - d_{j-1}) &= P(C_j = i | C_{j-1} = i, d_j, d_{j-1}) \\ &= P(L_i > d_j - d_{j-1}) \\ &= e^{-v_i(d_j - d_{j-1})} \\ &= e^{-v_i \Delta_j} \end{aligned}$$

where $\Delta_j = d_j - d_{j-1}$, the distance between SNP j and $j-1$. This holds because we know that L_i is exponentially distributed with parameter ν_i. Similarly, the probability that we transition to state k from state i between SNPs j and $j-1$ is

$$\begin{aligned} T_{ik}(\Delta_j) &= P(C_j = k | C_{j-1} = i, \Delta_j) \\ &= P(C_j = k | C_{j-1} = i) P(L_i \leq \Delta_j) \\ &= (q_{ik}/v_i)(1 - e^{-v_i \Delta_j}) \\ &= p_{ik}(1 - e^{-v_i \Delta_j}). \end{aligned}$$

Therefore, the corresponding distance-d transition probability matrix is given by

$$T(d) = \begin{pmatrix} e^{-\nu_1 d} & p_{12}(1 - e^{-\nu_1 d}) & (1 - p_{12})(1 - e^{-\nu_1 d}) \\ p_{21}(1 - e^{-\nu_2 d}) & e^{-\nu_2 d} & (1 - p_{21})(1 - e^{-\nu_2 d}) \\ p_{31}(1 - e^{-\nu_3 d}) & (1 - p_{31})(1 - e^{-\nu_3 d}) & e^{-\nu_3 d} \end{pmatrix}.$$

2.7 Estimation

The standard approach to parameter estimation for HMMs consists of two stages. First, we find the marginal posterior mode of the model parameters,

$$\hat{\theta} = \arg\max_{\theta} p(\theta | Y, \hat{\phi}). \tag{3}$$

This can be accomplished by using the Baum-Welch algorithm (Welch, 2003).

Second, given the parameter estimates $\hat{\theta}$, we then use the Viterbi algorithm (Rabiner, 1989) to calculate the most probable sequence of CN states,

$$\hat{C} = \arg\max_{C} p(Y | C, \hat{\theta}). \tag{4}$$

The primary problem with this approach is that it does not fully account for the uncertainty in the parameters. That is, its solution is conditional on a single point estimate $\hat{\theta}$ and it fails to take into account other reasonable values of θ. Other authors have also recognized this shortcoming and have proposed using a fully Bayesian approach; see e.g. Churchill and Lazareva (1999). Furthermore, the EM algorithm does poorly at estimating the parameters of the transition intensity matrix, especially the transition rate parameters, ν_i, which we expect to be very small. For example, this was observed by Rydén (2008) who recommend using priors

for these parameters. To address these issues, we propose an MCMC algorithm in Section 2.9.

2.8 Prior Distributions

We used a fully Bayesian approach to estimate the model parameters and copy number. We place prior distributions on the unknown parameters of the emission distribution which depend on the underlying copy number as follows.

$$\mu_c \sim N(m_c, v_c^2) \tag{5}$$

and

$$\frac{1}{\sigma_c^2} \sim \frac{1}{d_{0,c} s_{0,c}^2} \chi^2_{d_{0,c}}, \tag{6}$$

where $d_{0,c}$ are the degrees of freedom for the $\chi^2 - \text{distribution}$ and $s^2_{0,c}$ is the variance of a typical locus. This is the scaled $\text{Inv-}\chi^2$ specification for the variance, σ^2_{ik}; see Gelman et al. for a definition.

We also place priors on the parameters of the distance-d transition probability matrix:

$$\begin{aligned} p_{12} &\sim \text{Beta}(a_1, b_1), \\ p_{21} &\sim \text{Beta}(a_2, b_2), \\ p_{31} &\sim \text{Beta}(a_3, b_3), \\ \nu_i &\sim \text{Gamma}(2, 1/l_i) \quad i = 1, 2, 3. \end{aligned}$$

These priors are conjugate for all parameters except for the ν-parameters for which no conjugate prior exists. This leaves us with a set of hyperparameters, $\phi = (m_c, v_c^2, a_1, a_2, a_3, b_1, b_2, b_3, l_1, l_2, l_3)$, to specify. Our Beta prior parameters (a_1, a_2, a_3) and (b_1, b_2, b_3) can be interpreted as pseudo-counts of the number of transitions between states. For example, a_1 is the prior number of observations for transitions between the 1-copy state and the 2-copy state, and b_1 is the prior number of observations for transitions between the 1-copy state and the 3-copy state. Since it is rare for a duplication to follow a deletion, we expect the probability p_{12} to be close to 1. Thus, the specification of a_1 should be large relative to b_1. Similarly, the reciprocal of ν_1 represents the expected length of a deletion and, under the above prior specification, $E(1/\nu_1) = l_1$. Therefore, l_1 is the prior expected length of a deletion. In sum, informative priors are appealing for copy number data, because we can use prior biological knowledge to help specify the parameters while, at the same time, allow the data to adjust these *a priori* values accordingly.

2.9 A Markov Chain Monte Carlo Algorithm

The posterior distribution of (C, θ) can be written as

$$\begin{aligned} P(C,\theta|Y) &\propto P(\theta)P(Y|C,\theta)P(C|\theta) \\ &= P(\theta)\pi_{c_1} f(Y_1|c_1)\prod_{i=2}^{M} f(y_j|c_j)T_{c_{i-1},c_i}(\Delta_i), \end{aligned}$$

where $T_{ij}(d)$ is the $ij\text{th}$ element of the distance-d transition probability matrix described in Section 2.6. To generate samples from this posterior distribution, an MCMC algorithm was used. All model parameters have conjugate priors with the exception of the rate parameters for which we used the Metropolis-Hastings algorithm (Hastings, 1970) to sample from their distributions. We used the backward sampling algorithm proposed in Churchill and Lazareva (1999) to sample from $p(C|Y,\theta)$. Starting from initial values $\theta^{(0)}$ and $C^{(0)}$, iterate the following steps:

1. Given the current drawn values of θ_Q, update C using a backward sampling algorithm. The backward sampling algorithm is based on the equation

$$\begin{aligned} P(C_{l-1}=j|C_l,\ldots,C_M,Y,\theta_Q) &\propto P(C_{l-1}=j,C_l,\ldots,C_M,Y,\theta_Q) \\ &= P(Y_1,\ldots,Y_{t-1},C_{t-1}=j|C_t)P(Y_t,\ldots,Y_M,C_{t+1},\ldots,C_M|C_t)P(C_t) \\ &\propto P(Y_1,\ldots,Y_{t-1},C_{t-1}=j,C_t) \\ &= P(Y_1,\ldots,Y_{t-1},C_{t-1}=j)P(C_t|C_{t-1}=j) \\ &= \alpha_{t-1}(j)T_{jC_t}(\Delta_t). \end{aligned}$$

Here, $\alpha_{t-1}(j)$ is the joint probability of observing data up to SNP $t-1$ and j copies at SNP $t-1$, often called the forward probability. Note that $T_{ij}(\Delta_t)$ depends on θ_Q.

(a) For $j=1,2,3$, initialize the forward probabilities: $\alpha_1(j)=\pi_j f_j(Y_1)$.

(b) For $j=1,2,3$, and given $\alpha_1(j)$, compute the remaining forward probabilities using the recursive equation
$\alpha_t(j)=f_j(Y_t)\sum_{i=1}^{3}\alpha_{t-1}(i)T_{ij}(\Delta_t)$.

(c) Sample C_M from a $\text{Multinomial}_1(\Lambda_M)$, where
$\Lambda_M=(\frac{\alpha_M(1)}{\sum_k \alpha_M(k)},\frac{\alpha_M(2)}{\sum_k \alpha_M(k)},\frac{\alpha_M(3)}{\sum_k \alpha_M(k)})$.

(d) Sample C_{t-1} recursively for $t=M,M-1,\ldots,2$ from
$C_{t-1}\sim \text{Multinomial}(\Lambda_{t-1}(C_t))$, where
$\Lambda_{t-1}(C_t)=(\frac{\alpha_{t-1}(1)T_{1C_t}(\Delta_t)}{\sum_k \alpha_{t-1}(k)T_{kC_t}(\Delta_t)},\frac{\alpha_{t-1}(2)T_{2C_t}(\Delta_t)}{\sum_k \alpha_{t-1}(k)T_{kC_t}(\Delta_t)},\frac{\alpha_{t-1}(3)T_{3C_t}(\Delta_t)}{\sum_k \alpha_{t-1}(k)T_{kC_t}(\Delta_t)})$.

2. For $c\in\{1,2,3\}$ and given the current drawn C, sample μ_c and σ_c^2 from their full conditional distributions. Since these distributions are standard conjugate distributions, we omit the details here.
3. Given the current drawn C, sample p_{12}, p_{21} and p_{31} from their full conditional distributions. Since these distributions are standard conjugate distributions, we omit the details here.
4. For $c\in\{1,2,3\}$, given the data, other model parameters and the most recently drawn value of ν_c, denoted by $\nu_c^{(t-1)}$, sample ν_c using a Metropolis-Hastings algorithm:
   (a) Sample a proposal $\nu_c^{(t)}$ from a $\text{Gamma}(1,\nu_c^{(t-1)})$ distribution.

(b) Accept $\nu_c^{(t)}$ with probability $\max\left\{\frac{p(\nu_c^{(t)}|\sim)\Gamma(\nu_c^{(t-1)})}{p(\nu_c^{(t-1)}|\sim)\Gamma(\nu_c^{(t)})}, 1\right\}$, where $p(\nu_c|\sim)$ represents the complete conditional distribution of ν_c, and $\Gamma(\nu_c)$ is the density function for the $\mathrm{Gamma}(1, \nu_c)$ distribution.

After running the MCMC algorithm, we have a set of sampled copy number sequences $\{C^{(t)}\}_{1<=t<=T}$, where $C^{(t)} = (C_1^{(t)}, \ldots, C_M^{(t)})$ and T is the number of MCMC scans, as well as a set of sampled model parameters $\{\theta^{(t)}\}_{1<=t<=T}$, where $\theta^{(t)} = (\mu_1^{(t)}, \mu_2^{(t)}, \mu_3^{(t)}, \sigma_1^{2,(t)}, \sigma_2^{2,(t)}, \sigma_3^{2,(t)}, p_{12}^{(t)} p_{21}^{(t)}, p_{31}^{(t)}, \nu_1^{(t)}, \nu_2^{(t)}, \nu_3^{(t)})$.

2.10 Data Analyses

We analyzed data from a set of HapMap samples containing aberrant SNPs that have been experimentally verified by quantitative real-time PCR (qPCR) in a separate study (MacConaill et al., 2007). We used the titration data, which have known biological structure, to specify the hyperparameters of the emission distribution. Previous studies indicate that copy number deletions and amplifications may vary widely in size, between 5 bps and over 200 kb (McCarroll et al., 2008). Since we are interested in intermediate-size (10-100 kb) CNVs, we used $l_1 = l_3 = 50000$ and $l_2 = 1e + 06$. We collected 15,000 posterior samples using the MCMC algorithm and discarded the first 5000 as a burn-in phase. For each SNP, we estimated the copy number by taking the average of the sampled copy number values across the MCMC scans and rounded this number to the nearest integer.

For fitting the discrete-time HMM, we used the VanillaICE package with the default settings (see Scharpf (2008), for details), except that we change the emission distribution to be equivalent to the one that was used for the continuous-time model.

The results are presented in Table 1. Among the models the CHMM performed the best with 13 out of 14 SNPs called correctly. The discrete-time HMM was next with 11 out of 14 SNPs called correctly.

Table 1. Predictions by various HMMs on a set of aberrant SNPs that have been experimentally verified by qPCR. DHMM is the discrete-time HMM. CHMM is the continuous-time HMM.

SNP	Chr	Sample	qPCR	DHMM	dChip	CHMM
SNP_A-1941019	13	NA10851	0.86	1.00	1.00	1.00
SNP_A-4220257	8	NA10851	1.40	2.00	2.00	2.00
SNP_A-2114552	22	NA10863	2.74	2.00	2.00	3.00
SNP_A-1842651	17	NA10863	4.27	3.00	2.00	3.00
SNP_A-4209889	3	NA12801	1.24	2.00	2.00	1.00
SNP_A-2102849	8	NA10863	0.88	1.00	2.00	1.00
SNP_A-2122068	8	NA10863	0.85	1.00	1.00	1.00
SNP_A-1932704	7	NA10863	0.00	1.00	2.00	1.00
SNP_A-1889457	8	NA10863	1.05	1.00	1.00	1.00
SNP_A-4204549	8	NA10863	0.82	1.00	1.00	1.00
SNP_A-2125892	22	NA12707	0.00	1.00	2.00	1.00
SNP_A-2217320	22	NA12707	1.40	1.00	2.00	1.00
SNP_A-2126506	17	NA12707	4.51	3.00	2.00	3.00
SNP_A-1851359	17	NA12707	2.53	3.00	2.00	3.00

2.11 Results from a Simulation Study

CNV breakpoints were simulated from a model with $q_{21} = q_{23} = 3.33e-07$, $q_{12} = q_{32} = 1.998e-05$, and $q_{13} = q_{31} = 2e-08$, except in one region of CN polymorphism of length 100 kb which had a 6-fold increase in the rate of deletions. These breakpoints were simulated over a 140 Mb stretch, the length of chromosome 8, independently for 100 samples, and then were mapped onto the genomic locations corresponding to the observed SNP markers for the Affymetrix 500K Nsp chip. For each sample, this resulted in underlying copy number calls for 14,839 SNPs. With these simulated copy number calls, observed data were then simulated from the following hierarchical model.

1. For each copy number class $c \in \{1, 2, 3\}$ and SNP $i \in \{1, \ldots, 14839\}$;
   (a) sample $\mu_{i,c} \sim N(m_c, v_c^2)$;
   (b) sample $\frac{1}{\sigma_{i,c}^2} \sim \frac{1}{d_{0,c}s_{0,c}^2}\chi_{d_{0,c}}^2$.
2. For $i \in \{1, \ldots, 14839\}$ and $j \in \{1, \ldots, 100\}$, sample $Y_{ij} \sim N(\mu_{i,c_{ij}}, \sigma_{i,c_{ij}}^2)$.

This was done for $m = (-0.729, 0, 0.5)$, $v^2 = (0.01, 0.005, 0.01)$, $d_0 = (4, 4, 4)$ and $s_0^2 = (0.043, 0.032, 0.043)$. These values were chosen to mimic estimates from the titration data which have known biological structure.

For the purpose of saving time, we analyzed the first 5,000 SNPs for each of the 100 samples. For estimation of each simulated data set, we used the same values of the hyperparameters as were used for the data analyses described in the previous section, and collected 8,000 samples from the posterior using the MCMC algorithm.

We compared the results of the continuous-time HMM to two other methods: GLAD (Hupe et al., 2004) and CBS (Olshen et al., 2004). For GLAD, the default settings were used. GLAD provides output labels which correspond to loss/gain/diploid status for each SNP. For CBS, we post-processed the results by merging classes with predicted means within 0.25 of one another. Furthermore, the class with mean closest to zero was assigned the diploid class (normal class of two copies). The remaining classes were assigned to either gain or loss depending on whether their predicted class mean was larger or smaller than the diploid class.

The results are presented in Table 1. The continuous-time HMM performed the best in terms of detecting aberrant loci. However, it also detected more false-positives than both CBS and the DHMM.

Table 2. Prediction results for the simulation study. The second column is the misclassification error rate, the third column is the true positive rate of detection, and the fourth column is the true negative rate. These error rates are based on averages across the 100 samples.

Method	Misclassification rate	TPR	TNR
CHMM	2.80%	84.54%	97.29%
GLAD	3.26%	55.39%	98.11%
CBS	0.76%	79.80%	99.88%
DHMM	0.74%	82.15%	99.85%

3. CONCLUSION

In this paper we develop and apply a continuous-time Hidden Markov Model for the analysis of SNP array data, to infer regions of altered copy number. We use a number of previously published results to help specify priors for the Bayesian models underlying the HMM. The copy number analysis and databases are a novel development in the area of genomics, hence it is important for models to be flexible enough to enable novel discoveries. In particular, the data analysis in this paper underlines the importance of developing a reliable estimation procedure for the parameters of the transition intensity matrix, as the results produced by the Viterbi algorithm are quite sensitive to the specification of these parameters. While the MCMC framework addresses the need to account for the uncertainty in the estimation of model parameters θ, it does also bring forth a new challenge which is how to summarize the sampled copy number sequences from the MCMC algorithm. Unfortunately, there is no Viterbi-style algorithm for maximizing $p(C|Y)$. We are currently assessing various approaches for summarizing the sampled copy number sequences from the MCMC results.

Related to this is CNV inference. The Bayesian continuous-time HMM framework we use is a more natural setting, compared to discrete-time HMMs, to develop new prior and parameter specification models. The model readily provides estimates of posterior quantities, such as the probability that a region of interest contains a CNV. Such probabilities may be used to rank detected CNVs. Our results indicate that our CHMM is already competitive with the specialized DHMM implementation for such data (a VanillaICE package) while allowing for a more consistent modeling framework.

Future work also includes extending the model to the analysis of multiple samples, with the ultimate goal of detecting copy number polymorphisms.

ACKNOWLEDGEMENT

This work was supported by a Natural Sciences and Engineering Research Council (NSERC) scholarship.

REFERENCES

Genechip Human Mapping 500K Array Set. Affymetrix, 2006. Data Sheet.

Affymetrix. Data resource center. http://www.affymetrix.com/support/mas/datasets.affx.

Henrik Bengtsson et al. Estimation and assessment of raw copy numbers at the single locus level. *Bioinformatics*, 24(6):759–767, 2008.

Gary A. Churchill and Betty Lazareva. Bayesian restoration of a hidden Markov chain with applications to DNA sequencing. *Journal of Computational Biology*, 6(2): 261–277, 1999.

Andrew Gelman et al. *Bayesian Data Analysis*. Chapman and Hall, second edition, 2003.

W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

Philippe Hupe et al. Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. *Bioinformatics* 20(18):3413–3422, 2004.

Giulia C. Kennedy et al. Large-scale genotyping of complex DNA. *Nature Biotechnology*, 21:1233–1237, 2003.

Ming Lin et al. dchipSNP: significance curve and clustering of SNP-array-based loss-of-heterozygosity data. *Bioinformatics*, 20(8):1233–1240, 2004.

Laura E MacConaill et al. Toward accurate high-throughput SNP genotyping in the presence of inherited copy number variation. *BMC Genomics*, 8(211), 2007.

Steven A. McCarroll and David M. Altshuler. Copy-number variation and association studies of human disease. *Nature Genetics*, 39(7 Suppl):S37–42, 2007.

Steven A. McCarroll et al. Integrated detection and population-genetic analysis of snps and copy number variation. *Nature Genetics*, 40(10):1166–1174, 2008.

Adam B. Olshen et al. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, 5(4):557–572, 2004.

Lawrence R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

Sheldon M. Ross. *Introduction to probability models*. John Wiley, second edition, 2003.

Tobias Rydén. EM versus Markov chain Monte Carlo for estimation of hidden Markov models: A computational perspective. *Bayesian Analysis*, 3(4):659–688, 2008.

Robert Scharpf. VanillaICE: Hidden markov models for the assessment of chromosomal alterations using high-throughput SNP arrays. R vignette, 2008.

Antoine M. Snijders et al. Assembly of microarrays for genome-wide measurement of DNA copy number. *Nature Genetics*, 29:263–264, 2001.

Susann Stjernqvist et al. Continuous-index hidden Markov modelling of array CGH copy number data. *Bioinformatics*, 23(8):1006–1014, 2007.

Lloyd R. Welch. Hidden Markov models and the baum-welch algorithm. *IEEE Information Theory Society Newsletter*, 53(4), 2003.

George Zogopoulos et al. Germ-line DNA copy number variation frequencies in a large North American population. *Human Genetics*, 122(3-4):345–353, 2007.