A Nonhomogeneous Hidden Markov Model for Gene Mapping Based on Next-Generation Sequencing Data

FATEMEH ZAMANZAD GHAVIDEL, JÜRGEN CLAESSEN, and TOMASZ BURZYKOWSKI

ABSTRACT

The analysis of polygenetic characteristics for mapping quantitative trait loci (QTL) remains an important challenge. QTL analysis requires two or more strains of organisms that differ substantially in the (poly-)genetic trait of interest, resulting in a heterozygous offspring. The offspring with the trait of interest is selected and subsequently screened for molecular markers such as single-nucleotide polymorphisms (SNPs) with next-generation sequencing. Gene mapping relies on the co-segregation between genes and/or markers. Genes and/or markers that are linked to a QTL influencing the trait will segregate more frequently with this locus. For each identified marker, observed mismatch frequencies between the reads of the offspring and the parental reference strains can be modeled by a multinomial distribution with the probabilities depending on the state of an underlying, unobserved Markov process. The states indicate whether the SNP is located in a (vicinity of a) QTL or not. Consequently, genomic loci associated with the QTL can be discovered by analyzing hidden states along the genome. The aforementioned hidden Markov model assumes that the identified SNPs are equally distributed along the chromosome and does not take the distance between neighboring SNPs into account. The distance between the neighboring SNPs could influence the chance of co-segregation between genes and markers. To address this issue, we propose a nonhomogeneous hidden Markov model with a transition matrix that depends on a set of distance-varying observed covariates. The application of the model is illustrated on the data from a study of ethanol tolerance in yeast.

Key words: next-generation sequencing, nonhomogeneous hidden Markov model, quantitative trait loci analysis, single-nucleotide polymorphisms.

1. INTRODUCTION

The analysis of polygenetic characteristics for mapping quantitative trait loci (QTL) has received a substantial amount of consideration in molecular genetics over the last decades. QTL analysis is a method that links two types of information—phenotypic data (trait measurements) and genotypic data (usually molecular markers)—in an attempt to interpret the genetic basis of variation in traits and to link certain phenotypes to specific regions of chromosomes. The goal of this process is to identify the precise
location of these linked chromosomal regions across the genome by looking for evidence of co-segregation, that is, the tendency for closely linked genes and genetic markers to be inherited together.

QTL analysis requires two or more strains of organisms that differ substantially in the quantitative trait of interest and reliable scoring of genetic markers. For this purpose, a cross is obtained from two parental strains, creating the first generation of heterozygous offspring (F1). The offspring with the phenotype of interest is identified and screened for genetic markers, such as single-nucleotide polymorphisms (SNPs). The advent of high-throughput technologies such as next-generation sequencing (NGS) (Shendure and Ji, 2005) provides a new way to score large numbers of genetic markers. Markers that are linked to a QTL influencing the trait of interest have a tendency to be inherited more frequently with this locus, whereas unlinked markers will not show any substantial association with the phenotype of interest.

Recently, a hidden Markov model (HMM) was proposed by Claesen and Burzykowski (2014) to map QTLs using NGS-assisted bulked segregant analysis (BSA) (Schneeberger et al., 2009). HMMs (MacDonald and Zucchini, 1997) consist of two stochastic processes: a (hidden) sequence of states characterized by the state transition probability distribution with a Markov chain property and a sequence of observations with state-dependent probability distribution. The HMM proposed by Claesen and Burzykowski (2014) provides a flexible approach to classify each identified SNP into one of several predefined states, which have their own specific biological interpretation. The identified states of the HMM allow to identify genomic regions that may be likely to contain trait-related genes.

The model proposed by Claesen and Burzykowski (2014) assumes that the identified SNPs are equally spaced across the whole genome. This assumption is not necessarily correct. On the other hand, the chance of co-segregation may depend on the distance between the SNPs. Hence, an extension of the HMM that accommodates the distance between SNPs is of interest. Toward this aim, the assumption that the latent Markov chain is homogeneous, that is, that the transition probabilities are constant, can be weakened. More specifically, the transition probabilities can be assumed to depend on the distance between SNPs. This results in a nonhomogeneous HMM (NH-HMM) model. Such models have been considered, for example, in environmental studies (Zucchini and Guttorp, 1991; Hughes and Guttorp, 1994; Charles et al., 1999; Hughes et al., 1999; Bellone et al., 2000; Charles et al., 2004; Robertson et al., 2004; Betró et al., 2008). In our article, we apply an NH-HMM to the analysis of an experiment aimed at finding QTLs for high ethanol tolerance in Saccharomyces cerevisiae.

2. DATA

In the case study, BSA (Schneeberger et al., 2009) was combined with NGS to allow simultaneous identification of many genetic markers.

A highly ethanol-tolerant yeast strain (S. cerevisiae) was crossed with a laboratory strain (without the trait) of a moderate ethanol tolerance, resulting in 5974 viable haploid yeast cells (Swinnen et al., 2012a,b). Haploid offspring was screened for high ethanol tolerance, first in a medium containing 16% of ethanol producing 136 ethanol-tolerant segregants and subsequently with 17% ethanol giving rise to 31 segregants. The genomic DNA from both pools, 16% ethanol (pool 1) and 17% ethanol (pool 2), and the parent strains with high ethanol tolerance was submitted to a pooled-segregant genome-wide sequencing analysis by means of high-throughput NGS using the Illumina/Solexa technique. The technique produced DNA sequences. These reads were subsequently aligned to a DNA sequence of the parental laboratory yeast strain (without the trait) and SNPs as genetic markers were identified. For each identified SNP, the chromosomal position, the number of sequencing events (reads), and the number of times that the nucleotides A, C, G, and T were present in the offspring were recorded. The larger the proportion of differences in terms of the nucleotides (mismatch/SNP frequency) between the offspring and the parental strain, the higher the chance of a presence of a trait-related gene in the vicinity of the chromosomal location.

3. METHODOLOGY

3.1. A homogeneous HMM

HMMs are a particular kind of mixture models that allow for serially dependent observations. An HMM consists of two components. The first component is a “parameter process” (Zucchini and MacDonald,
2009). It is a sequence \( \{C_1, C_2, \ldots \} \) of discrete random variables that can assume one of \( m \) possible values ("states") from a set \( \Sigma = \{s_1, s_2, \ldots, s_m\} \). We assume that the sequence of states forms a first-order Markov chain; that is, \( P(C_i | C_1, \ldots, C_{i-1}) = P(C_i | C_{i-1}) \). Moreover, we assume that, for all \( i \), \( P(C_i = s_j | C_{i-1} = s_k) = \gamma_{jk} \), where \( \gamma_{jk} \) is the probability of transition from state \( s_j \) to state \( s_k \). The transition probabilities form the state-transition probability matrix \( \Gamma \), with the element in \( j \)th row and \( k \)th column equal to \( \gamma_{jk} \).

Additionally, we specify the "initial state" probability distribution; that is, \( P(C_1 = s_j) \equiv \delta_j \), say.

The second part of an HMM is a "state-dependent" process \( \{X_1, X_2, \ldots \} \). The distribution of random variable \( X_i \) depends on state \( C_i \); that is, \( P(X_i | C_1, \ldots, X_{i-1}, C_i, \ldots, C_N) = P(X_i | C_i) \). In particular, we assume that \( P(X_i = x \mid C_i = s_j) \), the "emission probability" of \( x \) for state \( s_j \), depends on a vector of parameters \( \theta_j \).

We define \( \theta = (\theta_1, \ldots, \theta_m) \).

In practice, we observe only a sequence of values \( \{X_1 = x_1, X_2 = x_2, \ldots, X_N = x_N\} \), without the corresponding sequence of the generating states \( \{C_1 = c_1, C_2 = c_2, \ldots, C_N = c_N\} \). In other words, the Markovian "parameter process" is hidden from the observer and hence the name of the model.

Figure 1 schematically presents the HMM described above.

If the state sequence \( \{C_1, C_2, \ldots, C_N\} \) is known exactly, that is, that \( C_1 = c_1, C_2 = c_2, \ldots, C_N = c_N \), then the likelihood function based on the observed sequence \( \{X_1 = x_1, X_2 = x_2, \ldots, X_N = x_N\} \) and the state sequence can be represented as follows:

\[
P(X_1, x_2, \ldots, x_N, c_1, c_2, \ldots, c_N) = \delta_{c_1} \gamma_{c_1, c_2} \cdots \gamma_{c_{N-1}, c_N} P_{c_1}(x_1) \cdots P_{c_N}(x_N).
\]

(1)

The likelihood function, given in Equation (1), is referred to as the complete-data likelihood, because it makes an assumption that both the observed sequence and the states are known.

Define \( u_j(i) \) as an indicator variable taking the value 1 if \( C_i = s_j \) and 0 otherwise. Moreover, define \( v_{jk}(i) \) taking the value 1 if \( C_{i-1} = s_j \) and \( C_i = s_k \) and 0 otherwise. Using this notation, we can express the logarithm of Equation (1) as follows:

\[
\log [P(x_1, \ldots, x_N, c_1, \ldots, c_N)] = \sum_{j=1}^{m} u_j(1) \log (\delta_j) + \sum_{j=1}^{m} \sum_{k=1}^{m} \sum_{i=2}^{N} v_{jk}(i) \log (\gamma_{jk}) + \sum_{j=1}^{m} \sum_{i=1}^{N} u_j(i) \log p_j(x_i).
\]

(2)

Note that Equation (2) is composed of three components, each depending on a different set of parameters. However, likelihood (1) cannot be used to estimate the parameters of the model, because the state sequence is not observed. Instead, the observed-data likelihood has to be used, given by

\[
L_N = P(x_1, x_2, \ldots, x_N) = \sum_{j_1=1}^{m} \cdots \sum_{j_N=1}^{m} \{ \delta_{j_1} \gamma_{j_1, j_2} \cdots \gamma_{j_{N-1}, j_N} P_{j_1}(x_1) \cdots P_{j_N}(x_N) \}.
\]

(3)

**FIG. 1.** A HMM underlying the sequence of data values \( \{X_1, \ldots, X_{i-1}, X_i, X_{i+1}, \ldots, X_N\} \). \( C_i \) is the hidden state for the observation \( X_i \), \( P(X_i | C_i) \) is the probability of emission of \( X_i \) in state \( C_i \), and \( P(C_i | C_{i-1}) \) is the probability of transition from state \( C_{i-1} \) to state \( C_i \). HMM, hidden Markov model.
In essence, the observed-data likelihood results from treating the sequence of states as missing data. The direct use of the likelihood in Equation (3) to estimate the parameters of an HMM would not be feasible, as it involves an enormous number of calculations (even for a moderate $N$). However, the maximization of Equation (3) can be obtained by using the forward–backward algorithm (also known as the Baum–Welch algorithm) (Baum et al., 1970), which is a form of the expectation–maximization algorithm (Dempster et al., 1977). In particular, in the expectation step (E-step), $u_{ji}(i)$ and $v_{jk}(i)$ are estimated given the observed data and current estimates of the model parameters. In the maximization step (M-step), the log-likelihood function (2) is maximized with respect to $\delta$, $\Gamma$, and $\theta$, given the estimates of $u_{ji}(i)$ and $v_{jk}(i)$. Since the parameters appear in separate components of Equation (2), the maximum of each of the components can be sought for separately.

The E- and M-steps are repeated until a convergence criterion is met. For instance, the difference between the values of parameters estimated in two consecutive iterations may be required to become smaller than a predefined threshold. Standard errors of the parameter estimates can be obtained by the method proposed by Louis (1982).

After estimating the parameters of the HMM, the sequence of the hidden states, which could have generated the observed sequence of symbols, can be predicted. For instance, the most likely sequence (“global decoding”) can be found with the help of the Viterbi algorithm (Viterbi, 1967). Alternately, for each $X_i$, the most likely state can be assigned based on the conditional state distribution given $X_i$ (“local decoding”). Note that the so-determined most likely state may differ from the one assigned by the Viterbi algorithm (though in particular applications, the differences may be minor).

### 3.2. A nonhomogeneous HMM

In a homogeneous HMM (H-HMM), the transition probabilities are constant. In a NH-HMM, the transition probabilities may depend on the position of $C_i$ in the state sequence $\{C_1, C_2, \ldots, C_N\}$. In other words, the model is characterized by position-specific transition-probability matrices $\Gamma_i$. For instance, assume that, for each random variable $X_i$, from the observable sequence $\{X_1, X_2, \ldots, X_N\}$, information is available about covariate $Y_i$. The transition-probability matrix can now depend on the covariate; that is, $\Gamma_i = \Gamma(Y_i)$. The form of the dependence of the transition probabilities on the covariates has to be specified in a way that is proper to the particular application. Hughes (1993) discussed several different parameterizations for transition probabilities.

A possible form of dependence of transition probabilities on covariates can be specified as follows:

$$
\pi_j(k, i) = P(C_{i+1} = s_k | C_i = s_j, Y_i) = \frac{\exp(x_j + \beta_{jk}Y_i)}{\sum_{k=1}^{m} \exp(x_j + \beta_{jk}Y_i)},
$$

in which the transition probabilities depend on covariate $Y_i$ through multinomial logit link functions. The unknown parameters $\pi_j(k, i)$ and $\beta_{jk}$ are the coefficients of the link function that have to be estimated. The resulting nonhomogeneous transition matrices $\Gamma_i$ are used in Equation (3). The maximization of Equation (3) can be obtained by using the forward–backward algorithm, as for the HMM. After estimating the parameters of the NH-HMM, the sequence of the hidden states can be predicted using global or local decoding (see Section 3.1).

### 3.3. A nonhomogeneous HMM for BSA experiments

Claesen and Burzykowski (2014) proposed to use three states ($m = 3$) to map QTLs. These states correspond to linkage or no linkage with genes responsible for the phenotype of interest. For a given SNP, at location $i$, no linkage implies that the number of nucleotides concordant with the parental strain without the trait is equal to the number of discordant counts. As a consequence, the resulting SNP frequency should be equal to 50%. In case of linkage, the number of discordant counts and that of discordant counts are no longer equal to each other. If the number of discordant counts is larger than the number of concordant counts, linkage with a locus of the parent with the trait could be inferred. If the number of concordant counts is larger than the number of discordant counts, one could postulate that there is a linkage with genes of the parent without the trait.

For each identified SNP, we observe the set of counts $(n_{A,i}, n_{C,i}, n_{G,i}, n_{T,i})$ which provides the number of times nucleotides $A$, $C$, $G$, and $T$ were present in the offspring. Conditioning on the reference nucleotide
observed for the particular SNP, the (“emission”) probability for the observed set of counts for the \( j \)th state can be expressed as follows:

\[
P_j(n_{rA,i}, n_{rC,i}, n_{rG,i}, n_{rT,i}) = \left( \prod_{l \in \{A,C,G,T\}} \left( \frac{\mu_{rl,j}}{\sum_{r \in \{A,C,G,T\}} \mu_{rj}} \right)^{n_{rl,i}} \right),
\]

where \( \mu_{rl,j} \) is the probability of observing the parental nucleotide \( r \) and offspring nucleotide \( l \) for the \( j \)th state.

By assuming that “nucleotide mismatch” probabilities \( \mu_{rl,j} = \mu_j' \) for \( r \neq l \) and that “nucleotide match” probabilities \( \mu_{rr,j} = \mu_j \), with \( \mu_j + 3\mu_j' = 0.25 \), and using the transformation

\[
\theta_j = \log \left( \frac{\mu_j}{3\mu'_j} \right),
\]

the probability, given in Equation (5), can be expressed in the following form:

\[
P_j(n_{rA,i}, n_{rC,i}, n_{rG,i}, n_{rT,i}) = \left( \prod_{l \neq r} \left( \frac{e^{\theta_j}}{1 + e^{\theta_j}} \right)^{n_{rl,i}} \right) \prod_{l \neq r} \left( \frac{1}{3(1 + e^{\theta_j})} \frac{1}{3(1 + e^{\theta_j})} \right)^{n_{rr,i}}. 
\]

Because of higher error rates in NGS sequencing (Dohm et al., 2008), the possibility of including sequencing errors in the model has been proposed by Claesen and Burzykowski (2014). In particular, assuming that there are no sequencing errors in the reference sequence, that the probability of a sequencing
error ($e$) is identical for all nucleotides, and that the probability of a sequencing error in the offspring sequence does not depend on the sequenced nucleotide, Equation (7) takes the following form:

$$P_j(n_{rA,i}, n_{rC,i}, n_{rG,i}, n_{rT,i}) = \left(\frac{n_i}{n_{rA,i}, n_{rC,i}, n_{rG,i}, n_{rT,i}}\right)^{n_{r,i}} \prod_{i \neq j} \left(\frac{1 + e(3^{e_i} - 1)}{3(1 + e_i)}\right)^{n_{r,i}}. \tag{8}$$

where $e = e_{tot}/3$.

In the proposed NH-HMM, the transition probabilities depend upon the distance between adjacent SNPs (in 10,000 bp). Denote the distance between the $i$th and $(i + 1)$th SNP by $y_i$. The transition probabilities can then be modeled as a function of $y_i$ by the multinomial logit link function (4). To ensure estimability of the parameters $\alpha_{jk}$ and $\beta_{jk}$, we constrain $\alpha_{j1}$ and $\beta_{j1}$ to be equal to zero.

4. RESULTS

Figure 2 presents the distribution of the distance between neighboring SNPs (in 10 kbp) for pool 1 of chromosome XIV and chromosome IX of the high ethanol-tolerance dataset. As it can be seen from the histograms, the SNPs are not equally spaced. Therefore, we applied the proposed NH-HMM to the data (the results of chromosome IX can be found in the Supplementary Materials available online at www.liebertonline.com/cmb). We compare the results of the proposed NH-HMM with the outcome of the HMM with the homogeneous transition matrix in terms of the parameter estimates and the allocation of the SNPs to particular states.

4.1. Chromosome XIV, pool 1

Table 1 shows the estimated concordance probabilities, initial probabilities, Akaike’s information criterion (AIC), and Bayesian information criterion (BIC) for the H-HMM and NH-HMM.

The comparison of the first and the second row of Table 1 reveals that on the basis of the information criteria, the NH-HMM is worth considering. Note that the estimates of the concordance probabilities $\mu_j$ are almost similar for both models. In particular, based on the NH-HMM, the total probabilities of discordance between the reference parent and the offspring with high ethanol tolerance can be estimated to be equal to $1 - 4 \times 0.2268 = 0.0928$, $1 - 4 \times 0.128 = 0.488$, and $1 - 4 \times 0.0559 = 0.776$ for the first, second, and third

<table>
<thead>
<tr>
<th>Parameters</th>
<th>$j = 1$</th>
<th>$j = 2$</th>
<th>$j = 3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha_{1j}$</td>
<td>0</td>
<td>-3.591</td>
<td>-3.126</td>
</tr>
<tr>
<td>$\alpha_{2j}$</td>
<td>-2.705</td>
<td>-4.088</td>
<td>-4.055</td>
</tr>
<tr>
<td>$\alpha_{3j}$</td>
<td>-3.926</td>
<td>-5.483</td>
<td>-5.483</td>
</tr>
<tr>
<td>$\beta_{1j}$</td>
<td>-3.167</td>
<td>-2.939</td>
<td>-2.939</td>
</tr>
<tr>
<td>$\beta_{2j}$</td>
<td>3.402</td>
<td>5.619</td>
<td>5.619</td>
</tr>
<tr>
<td>$\beta_{3j}$</td>
<td>3.402</td>
<td>5.619</td>
<td>5.619</td>
</tr>
</tbody>
</table>
state, respectively. Thus, following the reasoning outlined in Section 3.3, the first state can be seen as corresponding to SNPs linked with a gene locus of the parent without the trait, the second state can be considered as corresponding to the SNPs not linked with any genes, whereas the third state can be seen as SNPs that are potentially linked to one or more genes responsible to high ethanol tolerance.

Table 2 shows the estimated coefficients of the logit link function (4) describing the dependence of the transition probabilities on the distance between the adjacent SNPs (in 10 kbp). The estimated transition probabilities of the Markov chain, as a function of the distance between SNPs, are shown in Figure 3.

![Figure 3](image1.png)

**FIG. 3.** Probabilities of transition from one state to another, as estimated by the three-state nonhomogeneous HMM for chromosome XIV, pool 1.

![Figure 4](image2.png)

**FIG. 4.** Chromosome XIV, pool 1, predicted SNP-specific states for the model with homogeneous HMM (left panel) and with nonhomogeneous HMM (right panel). Red stands for state 1, green for state 2, and blue for state 3.
estimated coefficients indicate that, conditional on being in state 1, increasing the distance to the neighboring SNP increases the probability of transition to the second state. Similarly, for state 2, the larger the distance between the adjacent SNPs, the higher the chance of transition to state 3 as compared to state 1. For state 3, results similar to those for state 1 are obtained: increasing the distance between SNPs increases the probability of transition to the second state. In other words, for all three states, increasing the distance between the neighboring SNPs increases the chance of transition to another state. As a consequence, the transition probabilities of staying in the same state decrease as the distance between the neighboring SNPs increases. For state 1 (Fig. 3) this decline is especially large as compared to SNPs in states 2 and 3. The small number of SNPs, less densely spread out along the chromosome with SNPs frequency below 20% (Fig. 4), is most likely responsible for this effect. On the other hand, the probability of co-segregation decreases as the distance between the adjacent SNPs increases. The presence of QTLs for high ethanol tolerance in state 3 could be a possible reason for the high chance of transition from state 2 to 3 and from state 3 to 2 over large distances.

The left-hand-side panel of Figure 4 presents the states predicted for each SNP based on the most likely state sequence (Viterbi algorithm) obtained from the fitted NH-HMM (Table 2), while the right-hand-side panel presents the states predicted for each SNP obtained from the NH-HMM. Both plots clearly illustrate that state 3 (blue) is associated with a high frequency of mismatched nucleotides, state 2 (green) with an intermediate frequency, and state 1 (red) with a low frequency. Except in few cases (10 SNPs), NH-HMM and H-HMM assign the SNPs to the same states.

### 4.2. Chromosome XIV, pool 2

Table 3 shows the results for the H-HMM and NH-HMM for pool 2 of chromosome XIV. The corresponding total discordance probabilities are equal to 0.321, 0.609, and 0.876 for state 1, state 2, and state 3, respectively. Table 4 shows the estimated coefficients of the logit link function describing the dependence of the transition probabilities on the distance between the adjacent SNPs (in 10 kbp) for the NH-HMM. It can be observed that (see Fig. 5), similarly to pool 1, the probabilities of staying in the same state decrease as the distance between the neighboring SNPs increases. However, for state 1 of pool 2, this decline is not as large as for pool 1 (see Fig. 3). The larger number of SNPs around 200,000 bp, due to the presence of a minor QTL, could be a possible reason for this behavior (Fig. 6). If the distance between SNPs increases, the possibility of moving from state 2 to state 3 and for state 3 to 2 also increases. Except for 10 SNPs, NH-HMM and H-HMM assign the SNPs to the same states (see Fig. 6).

### Table 3. Chromosome XIV, Pool 2: Comparison of Two Forms of the Hidden Markov Model with Three States and Their Corresponding Parameter Estimates (95% Confidence Intervals in Brackets)

<table>
<thead>
<tr>
<th>Model</th>
<th>$\mu_1$</th>
<th>$\mu_2$</th>
<th>$\mu_3$</th>
<th>$\delta_1$</th>
<th>$\delta_2$</th>
<th>$\delta_3$</th>
<th>$-\log L_{20}$</th>
<th>AIC</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>H-HMM</td>
<td>0.1696</td>
<td>0.0979</td>
<td>0.0309</td>
<td>1a</td>
<td>0a</td>
<td>0a</td>
<td>74.749.81</td>
<td>149,521.6</td>
<td>149,583.6</td>
</tr>
<tr>
<td></td>
<td>[0.168,0.171]</td>
<td>[0.096,0.099]</td>
<td>[0.029,0.032]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NH-HMM</td>
<td>0.1695</td>
<td>0.0979</td>
<td>0.0308</td>
<td>1a</td>
<td>0a</td>
<td>0a</td>
<td>74.739.59</td>
<td>149,501.2</td>
<td>149,563.1</td>
</tr>
<tr>
<td></td>
<td>[0.168,0.171]</td>
<td>[0.096,0.099]</td>
<td>[0.029,0.032]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*aConfidence interval could not be calculated as these parameters are at the boundary of the parameter space.

### Table 4. Parameter Estimates (95% Confidence Intervals in Brackets) for the Nonhomogeneous Hidden Markov Model with Three States of Pool 2 of Chromosome XIV

<table>
<thead>
<tr>
<th>Parameters</th>
<th>$j = 1$</th>
<th>$j = 2$</th>
<th>$j = 3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha_j$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_j$</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

ZAMANZAD GHAVIDEL ET AL.
5. CONCLUSIONS

In this article, we have presented an NH-HMM model for QTL mapping. The approach adopted by the NH-HMM has a number of advantages over a homogeneous HMM. Most importantly, by taking into account the distance between adjacent SNPs, an NH-HMM better models chromosomes where some regions are densely covered and others are covered at lower density. In addition, an NH-HMM can be extended to include other covariates except the distance between the adjacent SNPs. However, further investigation should be devoted to the choice of the relevant covariates for the NH-HMM. One of the main concerns in an NH-HMM model is choosing a suitable link function together with an appropriate distance.
scale. For instance, in our case, we choose the standard (multinomial) logit link function between covariates and the transition probabilities. Other link functions, such as probit, are also possible. In our analysis, the distance scale of 10 kbp between the adjacent SNPs was used.

Another important issue for using the NH-HMM is the number of parameters in the model that have to be estimated. Applying the NH-HMM can increase the number of estimated parameters related to the transition probabilities. For larger number of states, the number of these parameters increases the computation time of the forward–backward algorithm.

To decide which stochastic model fits the data best, AIC and BIC can be computed. In our case, there was a slight difference between the values of these information criteria for the basic HMM and the NH-HMM. Moreover, there was little difference in the estimated values of the parameters of interest, that is, the concordance probabilities. Thus, in our application, the results did not depend on the assumption of (non)homogeneity of the transition matrix.

ACKNOWLEDGMENTS

The authors are grateful to Steve Swinnen, Thiago Pais, Maria R. Foulquie-Moreno, and Johan M. Thevelein of the Laboratory of Molecular Cell Biology, Institute of Botany and Microbiology, KU Leuven, and Department of Molecular Microbiology, VIB (Flemish Institute for Biotechnology), for providing the data. The authors gratefully acknowledge the support from the IAP Research Network P7/06 of the Belgian State (Belgian Science Policy).

AUTHOR DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES


Address correspondence to:
Fatemeh Zamanzad Ghavidel
Interuniversity Institute for Biostatistics and statistical Bioinformatics
Hasselt University
Agoralaan—Building D
3590 Diepenbeek
Limburg
Belgium

E-mail: fatemeh.zamanzadghavidel@uhasselt.be