Detecting selective sweeps using hidden markov models (HMMs)

Ali Sheikhi\textsuperscript{1}, David Ramsey\textsuperscript{1} and Gilbert MacKenzie\textsuperscript{1,2}

\textsuperscript{1} Center of Biostatistics, Department of Mathematics and Statistics, University of Limerick, Limerick, Ireland
\textsuperscript{2} ENSAI, Rennes, France

Abstract: A selective sweep occurs when a positive mutation spreads through a population. It causes a reduction in variation among the nucleotides in the neighbourhood of the mutation. Whether or not a selective sweep has occurred can be investigated in various ways. To detect selective sweeps, we usually use the concept of the coalescent tree. It employs a sample of individuals from a population to trace all the alleles of a gene shared by all members of the population to a single ancestral copy, known as the most recent common ancestor (MRCA). Recent advances in theory and technology have created a background for genome-wide surveys for selective sweep events. These advances include the development of new statistical tests tailored to detect incomplete, or partial, selective sweeps associated with weaker selection, and large-scale acquisition of Deoxyribonucleic acid (DNA) sequence data which provide ample source for the detection of polymorphism patterns.

In this work, we aim to detect selective sweeps by using Hidden Markov Models (HMMs).

Keywords: Selective sweep; Mutation; HMMs.

1 Background

Strong selection for a favourable new allele can result in a \textit{selective sweep}. As the frequency of the new mutant increases, adjacent chromosomal regions linked to the mutation are also swept to fixation. This process is called \textit{hitchhiking}, which leads to a reduction in variation at linked sites.

A strong selective sweep occurs when a positive mutation that was previously not observed in a population spreads. This results in a region of the genome where the positively selected haplotype (the mutated allele and its neighbours) is essentially the only one that exists in the population, resulting in a large reduction of the total genetic variation in that chromosome region. A weak selective sweep occurs when a previously neutral variant that is already present in a population becomes positively selected for. The lower the initial frequency of such a variant the more such a sweep will resemble a strong selective sweep.
We can determine the frequency and locations of the sweeps by scanning chromosomes for valleys of reduced variation. However, this reduction in variation may be a result of demographic factors (such as a rapid increase in the population or a population bottleneck), which makes the process complicated. Demographic factors affect all the genes in a genome, whereas selection has only local effects. The key to distinguishing between these two possibilities is the pattern of variation at other loci on the same chromosome. Demographic factors should affect the entire chromosome in the same way, while selection should affect only particular regions of the chromosome. So, one can detect selection by comparing multiple chromosome. If there is strong statistical evidence against the neutral equilibrium model for a particular locus, but the model fits the data in other loci quite well, this will usually be interpreted as evidence for selection at that locus.

There are several statistical methods to detect selective sweeps. Genomic scans using single-nucleotide polymorphism (SNP) data for the detection of these sweeps has recently received considerable attention. Recent advances in DNA sequencing technology have lead to an enormous increase in the amount of DNA sequence available for testing evolutionary hypotheses. In this work, we aim to use Hidden Markov Models (HMMs) to detect selective sweeps.

1.1 Linkage Disequilibrium (LD)

Linkage disequilibrium is a measure of the association of alleles on gametes or chromosomes. A population is said to be at linkage equilibrium at a set of loci if the alleles are independently distributed on chromosomes. It should be noted that a locus is the specific location of a gene or DNA sequence on a chromosome. So, linkage disequilibrium is the non-random association of alleles at two or more loci, not necessarily on the same chromosome.

To illustrate the problem suppose there are 3 loci, A, B and C. We denote the alleles at locus A by $A_1, A_2, \ldots$, at locus B by $B_1, B_2, \ldots$ and at locus C by $C_1, C_2, \ldots$ So $A_2B_3C_1$ is an individual carrying the $A_2$ allele at locus A, the $B_3$ allele at locus B and the $C_3$ allele at locus C ($A_2B_3C_1$ is called a haplotype). In addition, consider $p_{ijk}$ is the frequency of $A_iB_jC_k$ haplotype. We denote the frequency of $A_i$, $B_j$ and $C_k$ by $p_{i..}$, $p_{.j.}$ and $p_{..k}$. Obviously $p_{i..} = \sum_{j,k} p_{ijk}$. The frequency of $B_j$ and $C_k$, $p_{.j.}$ and $p_{..k}$ can be obtained from the haplotype frequency analogously.

Now, if for all i, j and k, $p_{ijk} = p_{i..}p_{.j.}p_{..k}$, then we say the variation at these three loci is at linkage equilibrium. Otherwise, there is linkage disequilibrium.

A number of measures can be used to measure the departure from linkage equilibrium. Suppose we look at haplotypes for two loci, A and B with two alleles each. We denote the frequency of the haplotype $A_iB_j$ by $x_{ij}$. Now, consider the table below:
One can use the above frequencies to determine the frequency of each of the alleles:

\[
\begin{align*}
 f(A_1) &= x_{11} + x_{12} = p_1 \\
 f(A_2) &= x_{21} + x_{22} = p_2 \\
 f(B_1) &= x_{11} + x_{21} = q_1 \\
 f(B_2) &= x_{12} + x_{22} = q_2
\end{align*}
\]

If the two loci and the alleles are independent of each other, i.e. the variation at these two loci is at linkage equilibrium, then \(x_{ij} = p_ip_j\).

Now, if the variation at these loci is not at linkage equilibrium, there is a degree of linkage disequilibrium which is often measured by a \(D\), defined according to the table below:

\[
\begin{align*}
 B_1 & \quad x_{11} = p_1q_1 + D \quad x_{21} = p_2q_1 - D \quad q_1 \\
 B_2 & \quad x_{12} = p_1q_2 - D \quad x_{22} = p_2q_2 + D \quad q_2 \\
 \text{Total} & \quad p_1 \quad p_2 \quad 1
\end{align*}
\]

Therefore when we say two alleles are not in LD, it means that \(D \neq 0\).

### 1.2 The genetic data used

The genetic data used for this study are in fact the sequencing of the human genome and consist of strings of nucleotides. Nucleotides are molecules that when joined together, make up the structural units of DNA. They are the building blocks of the human genome. Nucleotides are composed of three units: base, sugar (monosaccharide) and phosphate. There are four nucleotides: adenine, thymine, cytosine and guanine, abbreviated by A, T, C, and G, respectively.

Recently, The International HapMap Consortium (2003) and Seattle SNP Project (2004) have provided extensive maps of the genome of samples of humans from different subpopulations. The goal of the International HapMap Project is to determine the common patterns of DNA sequence
variation in the human genome and to make this information freely available in the public domain.

At a large majority of the sites observed, all the individuals have the same nucleotide. Sites at which variation is observed are called segregating sites. At virtually all such sites just two nucleotides (variants) are observed. The variation observed at a segregating site is termed a single nucleotide polymorphism (SNP).

Suppose the genomes of $n$ individuals are scanned. The following are commonly used as summary statistics:

1. The non-polarised frequency spectrum. This is made up of the frequencies of the least common (minor) variant at each of the $k$ segregating sites, together with the position of each segregating site. The non-polarised frequency at a segregating site is between 1 and $\left\lfloor \frac{n}{2} \right\rfloor$, where $\lfloor x \rfloor$ is the integer part of $x$.

2. The polarised frequency spectrum. This is made up of the frequencies of the wild type variant (assumed to have been the prevalent variant in a recent ancestral population) at each of the $k$ segregating sites, together with the position of each segregating site. The polarised frequency at a segregating site is between 1 and $n - 1$.

These data are available from Seattle SNPs Variation Discovery Resource at http://pga.gs.washington.edu.

2 Markov chains (MC)

A sequence of discrete random variables $\{C_t, t \in N\}$, is said to be a (discrete time) Markov chain (MC) if for all $t \in N$ the Markov property,

$$Pr(C_{t+1}|C_t, ..., C_1) = Pr(C_{t+1}|C_t)$$

is satisfied.

It means that conditioning on the history of the process up to time $t$ is equivalent to conditioning only on the most recent value $C_t$.

Important quantities associated with a Markov chain are the conditional probabilities called transition probabilities,

$$Pr(C_{s+t} = j|C_s = i) = \gamma_{i,j}(t)$$

If these probabilities do not depend on $s$, the Markov chain is called homogeneous. The transition matrix $\Gamma(t)$ is defined as the matrix with $(i, j)$ element $\gamma_{i,j}(t)$.

3 Hidden markov models (HMMs)

A hidden Markov model (HMM) $\{C_t, t \in N\}$ is given by a particular kind of dependency structure. The model consists of two parts: first, an unobserved
parameter process \{C_t : t = 1, 2, \ldots\} satisfying the Markov property, and second the state dependent process \{X_t : t = 1, 2, \ldots\} such that, when \(C_t\) is known, the distribution of \(X_t\) depends only on the current state \(C_t\) and not on previous states or observations. With \(X^{(t)}\) and \(C^{(t)}\) representing the histories from time 1 to time \(t\), one can summarize the simplest model of this kind by:

\[
Pr(C_t|C^{(t-1)}) = Pr(C_t|C_{t-1}), t = 2, 3, \ldots
\]

\[
P(X_t|X^{(t-1)}, C^{(t)}) = P(X_t|C_t), t \in \mathbb{N}
\]

where \(C^{(t)} = \{C_1, C_2, \ldots, C_t\}\).

A hidden Markov model (HMM) \{\(C_t, t \in \mathbb{N}\)\} is given by a particular kind of dependency structure. The model consists of two parts: first, an unobserved parameter process \(\{C_t : t = 1, 2, \ldots\}\) satisfying the Markov property, and second the state dependent process \(\{X_t : t = 1, 2, \ldots\}\) such that, when \(C_t\) is known, the distribution of \(X_t\) depends only on the current state \(C_t\) and not on previous states or observations. With \(X^{(t)}\) and \(C^{(t)}\) representing the histories from time 1 to time \(t\), one can summarize the simplest model of this kind by:

\[
Pr(C_t|C^{(t-1)}) = Pr(C_t|C_{t-1}), t = 2, 3, \ldots
\]

\[
P(X_t|X^{(t-1)}, C^{(t)}) = P(X_t|C_t), t \in \mathbb{N}
\]

where \(C^{(t)} = \{C_1, C_2, \ldots, C_t\}\).

### 3.1 Elements of an HMM

1. \(N\), the number of states in the model. We denote the set of states as \(S = \{S_1, S_2, \ldots, S_N\}\) and the (unobserved) state at time \(t\) as \(Q_t\).
2. \(T\), the number of distinct observation symbols for each state. Let \(X_t\) be the observation at time \(t\). Therefore, \(X = \{X_1, X_2, \ldots, X_T\}\).
3. The state transition probability distribution \(A = \{a_{ij}\}\) where,

\[
a_{ij} = P[Q_{t+1} = s_j|Q_t = s_i], 1 \leq i, j \leq N
\]

4. The observation symbol probability distribution in state \(j\), \(B = \{b_j(k)\}\) where,

\[
b_j(k) = P[X_t = x_t|Q_t = s_j], 1 \leq j \leq N, 1 \leq k \leq T
\]

5. The initial state distribution,

\[
\pi_i = P[Q_1 = s_i], 1 \leq i \leq N
\]
4 The model

We aim to introduce a new approach based on HMMs to detect selective sweeps by using DNA sequence data. In our work, we consider a sample consisting of \( n \) aligned DNA sequences of length \( L \) taken from the same population. Using these data, we wish to determine whether a selective sweep has occurred in the corresponding chromosomal region. For \( i = 1, \ldots, L \), let \( Y_i = 1, \ldots, n - 1 \) be the number of derived alleles at site \( i \), assuming an infinite site model. In our model, \( Y_i = 0, \ldots, n - 1 \) is the observed state at site \( i \). The hidden state \( X_i \) indicates whether site \( i \) has been affected by selection. We consider that there are only three hidden states: neutral, intermediate and sweep. A site is in a sweep state when it is very close to the selected locus. Its site frequency spectrum is strongly influenced by the sweep. The intermediate state applies to those loci that are only slightly influenced by the sweep because of their larger distance to the selected locus. We consider the following transition probability matrix:

\[
T = \begin{pmatrix}
1 - p & p & 0 \\
0 & 1 - p & \frac{p}{2} \\
0 & p & 1 - p
\end{pmatrix}
\]

where \( T_{j,k} \) denotes the transition probability from state \( j \) to state \( k \). The index \( j = 1 \) refers to the neutral state, \( j = 2 \) refers to the intermediate state, and \( j = 3 \) refers to the sweep state.

Acknowledgments: The authors is grateful for the support of Science Foundation Ireland under the BIO-SI project (no. 07MI012)

References


