# Haplotype reconstruction with Markov models

Lauri Eronen

Department of Computer Science
University of Helsinki

**Abstract**

Haplotypes extracted from DNA are a powerful form of genetic data for different tasks, whether an analysis of patterns across populations or gene mapping. A fundamental problem is that current practical laboratory methods do not give haplotype information. Recently, several statistical methods have been proposed for reconstructing haplotypes (pairs of sequences) from the available, unphased genotype data (sequences of unordered pairs). Often, these methods only work well if the number of markers, i.e., the length of the sequences, is relatively small, and if the markers are strongly correlated. In this paper, we present a new algorithm for statistical haplotype reconstruction. The algorithm exploits local regularities conserved in the haplotypes by modeling the haplotypes with a variable-order model. The algorithm is suitable for data sets with large numbers of markers with only moderately strong correlations.

## 1 Introduction

Most of our DNA sequence is common to all humans. However, due to mutations in the history, there are certain polymorphic sites in the genome, here called *markers*. The different variants that occur at markers are called *alleles*, and the string of alleles (ignoring all the DNA between the markers) in a given chromosome is called a *haplotype*. Markers and haplotypes can be used to study patterns of inheritance, and they are a powerful tool in gene mapping.

Each individual has two copies of each chromosome — one from each parent — and therefore also two alleles of each marker. Biological research based on individual chromosomes or patterns of inheritance has a fundamental problem here: current laboratory methods can only tell which two alleles a person has at a given marker, but they cannot tell which one is from each parent. In other words, haplotypes are not directly available, only *genotypes*, i.e., a list of unordered allele pairs.

The problem we consider is *haplotype reconstruction*: given the genotypes of a sample of individuals, the task is to predict the most likely haplotype pair for each individual. Haplotype reconstruction methods are based on statistical dependency between closely located markers, known as *linkage disequilibrium* (LD). LD is caused by the fact that markers close to each other are not likely to be separated by a recombination. Recombinations are relatively rare, so ancestral haplotype fragments of varying sizes will be shared by members of the same population. Our algorithm for haplotyping is based on finding and utilizing such shared structure across individuals.

Our motivation comes from gene mapping, where the goal is to locate a disease susceptibility gene by finding a haplotype fragment that is associated with the disease being studied. Typically, a gene mapping study uses a fixed set of markers called a *marker map*. Typical sizes of marker maps currently range from tens to thousands of markers, but maps with hundreds of thousands of markers are already available. Our goal is to develop a method that scales to large numbers of markers.

# 2 Problem setting and statistical background

We assume a set (map) $M$ of $\ell$ markers $1,\ldots,\ell$ and denote the set of all observed alleles of marker $i$ by $A_i$. A *haplotype* $H$ over $M$ is then a vector of alleles: $H \in \prod_{i=1,\ldots,\ell} A_i$. A *genotype* $G$ over $M$ is a vector of (unordered) allele pairs: $G \in \prod_{i=1,\ldots,\ell}\{\{a_1,a_2\} \mid a_1,a_2 \in A_i\}$.

The allele of haplotype $H$ at marker $i$ is denoted by $H(i)$. Similarly, the unordered allele pair of a genotype $G$ at marker $i$ is denoted by $G(i)$. Given a pair of haplotypes $\{H_1,H_2\}$ and a genotype $G$ such that $G(i) = \{H_1(i),H_2(i)\}$ for all $i$, we say that $\{H_1,H_2\}$ is *compatible* with $G$, or that $\{H_1,H_2\}$ is a (possible) *haplotype configuration* for genotype $G$. Two haplotypes determine a unique compatible genotype in the obvious way. A genotype, on the other hand, can have several compatible haplotype configurations. For a genotype $G$ with $k$ heterozygous markers (i.e., $k = |\{G(i) = \{a_1,a_2\} \mid a_1 \neq a_2\}|$), there are $2^{k-1}$ different haplotype configurations. The set of all haplotype configurations for a genotype $G$ is denoted by $C_G$, with $|C_G| = 2^{k-1}$.

The set of input genotypes is denoted by $\mathcal{G}$. The haplotype reconstruction problem is now defined as finding for each genotype $G \in \mathcal{G}$ the most probable haplotype configuration $\{H_1,H_2\} \in C_G$.

A common statistical approach to haplotype inference is to build a generative model of haplotype pairs, and use this model to resolve each genotype into the most probable consistent haplotype pair. The data is interpreted as a random sample of haplotype pairs drawn from the unknown distribution. We make the genetic assumption of *Hardy-Weinberg equilibrium*, which states that the haplotypes of an individual are independent of each other. Under this assumption, the probability of drawing a haplotype pair can be stated in terms of probabilities of the individual haplotypes, and the problem reduces to modeling the distribution of haplotypes:

$$P(\{H_1,H_2\}) = \begin{cases} 2 \cdot P(H_1)P(H_2) & \text{if } H_1 \neq H_2 \\ P(H_1)^2 & \text{if } H_1 = H_2. \end{cases} \tag{1}$$

As genotypes where $H_1 = H_2$ are unambiguous, in practice only the upper form of the probability in Equation 1 is used. Furthermore, the constant 2 can be ignored when comparing the probabilities, and will be left out to simplify the presentation.

Using Bayes' rule and Equation 1, the probability of a genotype $G$ resolving into any given haplotype configuration $\{H_1, H_2\} \in C_G$ can be stated as follows:

$$
\begin{aligned}
P(\{H_1, H_2\} \mid G) &\overset{Bayes'rule}{=} \frac{P(G \mid \{H_1, H_2\})P(\{H_1, H_2\})}{P(G)} \\
&\overset{1}{\propto} \frac{P(H_1)P(H_2)}{P(G)} \\
&\propto P(H_1)P(H_2). \qquad (2)
\end{aligned}
$$

The probability of a genotype, $P(G)$, can be treated as a normalizing constant. Most statistical haplotyping methods utilize the inference procedure captured by Equation 2. The main differences between the different methods lie in how the haplotype probability model, $P(H)$, is defined and how the parameters of the model are estimated.

The most straightforward way of modeling the haplotype distribution is to use a multinomial model [ES95], [NQXL02], with a frequency estimate, $\mathcal{F}(H)$, for each haplotype. In this setting, it is implicitly assumed that only few different haplotypes occur in the population, and most of the frequencies will be zero.

In practice, the frequencies usually have to be estimated from the genotype data itself. A common method of estimating parameters of a statistical model is the *maximum-likelihood approach:* find the parameters that maximize the likelihood of drawing the observed data. The probability of drawing a single genotype $G$ is just the sum of the probabilities of haplotype configurations compatible with the genotype:

$$
P(G \mid \mathcal{F}) = \sum_{\{H_1, H_2\} \in C_G} \mathcal{F}(H_1)\mathcal{F}(H_2),
$$

where $C_G$ is the set of haplotype configurations compatible with $G$. The likelihood of drawing a set $\mathcal{G}$ of genotypes is then just the product over all genotypes in $\mathcal{G}$:

$$
\begin{aligned}
L(\mathcal{G} \mid \mathcal{F}) &= \prod_{G \in \mathcal{G}} P(G \mid \mathcal{F}) \qquad (3) \\
&\propto \prod_{G \in \mathcal{G}} \sum_{\{H_1, H_2\} \in C_G} \mathcal{F}(H_1)\mathcal{F}(H_2).
\end{aligned}
$$

Unfortunately, analytically finding the haplotype frequencies that maximize $L(\mathcal{G} \mid \mathcal{F})$ is computationally infeasible already for a small number of markers. However, the EM algorithm [ES95], an iterative estimation procedure, can be used to find haplotype frequencies that (locally) maximize the likelihood of the observed genotype data. The algorithm consists of a sequence of two alternating steps. In the *E-step*, probabilities for each genotype consisting of all consistent haplotype pairs is calculated, conditioned on the haplotype frequencies obtained from the previous *M-step*. In the M-step, the haplotype frequencies are updated such that the likelihood for the observed data (conditioned on the resolution probabilities from the last E-step) is maximized. In practice this is done by computing the total estimated frequency of each haplotype in $\mathcal{G}$, based on the resolution probabilities from the previous M-step. These two steps are repeated, until the improvement in the likelihood (3) is smaller than some predefined threshold. For instance the uniform distribution can be used for initializing the algorithm. The EM algorithm converges to a local maximum on the likelihood surface; it is not guaranteed to find the global maximum. However, the algorithm has been found to work well in practice [FS00], in the case of a moderate number ($< 20$) of closely spaced markers.

## 3 Variable-order Markov model

Our focus is on marker maps with a large number of sparsely spaced markers. Under these conditions, recombinations between markers are common, and consequently haplotype diversity is large. This introduces two difficulties for the haplotype reconstruction procedure. First, the large number of markers means that the number of possible haplotype configurations for any single genotype becomes larger. Second, having a large number of different haplotypes makes it difficult to learn the haplotype distribution from data. Our central idea is to discover and utilize local LD structure, or *patterns* in the haplotypes, without requiring LD between distantly spaced markers. Here we restrict our attention to a simple class of patterns: a *haplotype fragment* is a haplotype restricted to a (continuous) sub-range of the original marker map.

We utilize the independence of non-neighboring markers (resulting from recombinations) by modeling the haplotypes by a (inhomogeneous) variable-order Markov model[EGT04]. The probability of a haplotype is broken into a product of conditional probabilities of individual alleles, with each marker conditional on a varying number of its immediate neighbors. When computing the probability of haplotype $H$, the distribution of alleles at marker $i$ is estimated using longest observed fragment that (1) matches haplotype $H$ and ends at marker $i-1$, and (2) has a frequency exceeding some given threshold, *minfr*:

$$P(H) = P(H(1)) \prod_{i=2,\ldots,\ell} P(H(i) \mid H(s_i, i-1)), \tag{4}$$

where $H(i)$ is the allele at marker $i$, $H(i,j)$ is the haplotype fragment covering markers $i,\ldots,j$ and $s_i = \min\{s \mid P(H(s,i-1)) \geq \mathit{minfr}\}$. Using a frequency threshold is motivated by the fact that it is not likely that a long fragment of haplotypes is shared by different individuals unless it is inherited from the same ancestral haplotype, thus giving a strong indication that also the next allele will be inherited from that same ancestor.

## 4 Algorithm

Our haplotype reconstruction algorithm is a modified version of the EM algorithm of [ES95]. The basic idea is the same: alternate steps for (1) estimating the model parameters and (2) reconstructing the haplotypes. These two steps are iterated, until the likelihood of the observed data does not improve any more. Note that the algorithm described in [EGT04] corresponds to performing only the first iteration of the algorithm.

**Haplotype reconstruction step (E)** In this step we calculate, for each $G \in \mathcal{G}$, the resolution probabilities of all compatible haplotype configurations. Let $\{H_1, H_2\} \in C_G$, then

$$P_t(\{H_1,H_2\} \mid G) = P(\{H_1,H_2\} \mid G, \mathcal{F}_{t-1}) = \frac{P(H_1 \mid \mathcal{F}_{t-1}) P(H_2 \mid \mathcal{F}_{t-1})}{\sum_{\{H,\overline{H}\} \in C_G} P(H \mid \mathcal{F}_{t-1}) P(\overline{H} \mid \mathcal{F}_{t-1})},$$

where $\mathcal{F}_{t-1}$ denotes the estimated model parameters at step $t$, and $P_t(\{H_1,H_2\} \mid G)$ denotes the probability of $G$ resolving into haplotypes $H_1$ and $H_2$. Note that the denominator is needed just for normalizing the the probabilities.

As the number of possible haplotype configurations for a genotype grows exponentially with the number of heterozygous markers, exhaustively going through all possible configurations is feasible only when the number of markers is small ($\approx 20$). When the number of markers is larger, we use a hierarchical pruning approach, motivated by a similar strategy used by Niu et al. [NQXL02]. The basic idea is to first break each genotype into short segments, and evaluate the probability of haplotype configurations separately for each segment. Then the configuration segments are combined into larger ones, until configurations for the complete marker map are reached. On each combination step, only a fixed number of most probable configurations are used to form the combined configurations.

**Parameter estimation step (M)** In this step, the model parameters are estimated conditioned on the previous resolution probabilities. The model parameters consist of the set of frequent fragments and the associated conditional allele probabilities. In practice, the conditional probabilities are derived from fragment frequencies as follows:

$$P(H(i) \mid H(i-d,i-1)) = \frac{\mathcal{F}(H(i-d,i))}{\mathcal{F}(H(i-d,i-1))},$$

where $\mathcal{F}\left(H(i,j)\right)$ denotes the frequency of haplotype fragment $H(i,j)$. It thus suffices to compute the set of frequent fragments, and extend each fragment with all possible alleles to obtain a set of fragments from which all the required probabilities can be computed. The resolution probabilities from the previous E-step give the expected number of occurrences of any haplotype in a given genotype. In a given genotype, the expected frequency of a haplotype fragment is obtained by a weighted sum over all haplotype configurations that match the fragment. Finally, the overall frequency of a fragment is obtained as an average over all genotypes:

$$\mathcal{F}_t(h) = \frac{1}{2|\mathcal{G}|} \sum_{G \in \mathcal{G}} \sum_{\{H_1,H_2\} \in C_G} P_t(\{H_1,H_2\} \mid G) \delta_{h,\{H_1,H_2\}},$$

where $\delta_{h,\{H_1,H_2\}}$ is the number of haplotypes in configuration $\{H_1,H_2\}$ that match $h$.

Depth-first search is used to find the set of frequent fragments from the haplotypes reconstructed in the previous E-step: start from frequent fragments of length one, and expand fragments to the right marker by marker until frequency drops below a given threshold.

**Initializing the algorithm** The initial fragment frequencies, $\mathcal{F}_0$, are computed analogously to the parameter estimation step. The difference is that there is not yet any information about the probability of different haplotype configurations; thus all configurations of a genotype must initially be considered equally likely. The number of haplotype configurations compatible with a genotype $G$ is exponential in the number of heterozygous markers in $G$, and computation by enumerating all the possible configurations in $C_G$ is thus infeasible. Fortunately, the initial frequencies can be counted directly from the genotype data as follows, without explicitly generating the elements of $C_G$:

$$\mathcal{F}(H(i,j)) = \frac{1}{2|\mathcal{G}|} \sum_{G \in \mathcal{G}, G \text{ matches } \mathrm{H}(i,j)} 2^{1-k_{G(i,j)}}, \tag{5}$$

where $k_{G(i,j)}$ is the number of heterozygous markers in genotype fragment $G(i,j)$ (note that a homozygous genotype has two identical haplotypes both matching the fragment, and thus weight 2).

# 5 Discussion

Modeling local regularities within a longer marker map is not a completely new idea. Recently, the hypothesis of *haplotype blocks* was introduced by Daly et al [DRS$^+$01]. In the block model, it is assumed that the human genome consists of consecutive regions of low haplotype diversity that are separated by so-called *recombination hotspots*, locations with a high recombination rate. Probabilistic models that describe the distribution of haplotypes in terms of haplotype blocks have been introduced recently by various authors [KPV$^+$03, AN03]. However, their models are targeted at describing the data in terms of haplotype blocks, and are not targeted at haplotype inference.

Although fixed-and variable order Markov models have been studied and used for many applications, they have not been applied to haplotype reconstruction earlier. The special structure of the haplotyping problem differentiates the work in this paper from usual applications of Markov chains. The Markov models employed here are *inhomogeneous*, i.e. each marker has its own states and transition probabilities. This in contrast with applications of Markov chains in e.g. sequence modeling problems, where the states and transition probabilities are not dependent on the location on the chromosome. The variable order Markov model used in this paper is also related to work on probabilistic suffix trees [RST94]; again, the difference is in the inhomogeneity of the Markov model.

Finally, there are connections to work on hidden Markov models [RJ86]. Similar to HMMs, haplotype reconstruction is an incomplete-data problem; the underlying haplotypes are not observed directly, but instead as a sequence of unordered allele pairs. In contrast to an ordinary HMM, there are 2 parallel hidden sequences (the haplotypes) which give rise to the observed sequence. Also unlike HMMs, the observations (allele pairs at each marker) are in this paper considered a deterministic function of the underlying haplotypes, and are thus assumed to be noise-free.

Standard HMM algorithms, such as Viterbi and Baum-Welch algorithms could be relatively straightforwardly adapted to the haplotyping problem using a fixed-order Markov model. It is, however, unclear whether they are applicable for the variable order case. One issue not handled by the current method is modeling of genotyping errors, which could possibly be incorporated into a Baum-Welch based method.

## References


[AN03] Eric C. Anderson and John Novembre. Finding haplotype block boundaries by using the minimum-description-length principle. *The American Journal of Human Genetics*, 73:336–354, 2003.

[DRS+01] Mark J. Daly, John D. Rioux, Stephan F. Schaffner, Thomas J. Hudson, and Eric S. Lander. High-resolution haplotype structure in the human genome. *Nature Genetics*, 29:229–232, 2001.

[EGT04] Lauri Eronen, Floris Geerts, and Hannu Toivonen. A Markov chain approach to reconstruction of long haplotypes. In *Pacific Symposium on Biocomputing 2004 (PSB 2004)*, pages 104–115, Hawaii, USA, 2004. World Scientific.

[ES95] Laurent Excoffier and Montgomery Slatkin. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Molecular Biological Evolution*, 12(5):921–927, 1995.

[FS00] Daniele Fallin and Nicholas J. Schork. Accuracy of haplotype frequency estimation for biallelic loci, via the expectation-maximization algorithm for unphased diploid genotype data. *The American Journal of Human Genetics*, 67:947–959, 2000.

[KPV+03] M. Koivisto, M. Perola, T. Varilo, W. Hennah, J. Ekelund, M. Lukk, L. Peltonen, E. Ukkonen, and H. Mannila. An MDL method for finding haplotype blocks and for estimating the strength of haplotype block boundaries. *Proceedings of the Pacific Symposium on Biocomputing 2003*, 2003.

[NQXL02] Tianhua Niu, Zhaohui S. Qin, Xiping Xu, and Jun S. Liu. Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. *The American Journal of Human Genetics*, 70:17–169, 2002.

[RJ86] L. R. Rabiner and B. H. Juang. An introduction to hidden Markov models. *IEEE ASSP Magazine*, 3(1):4 – 16, January 1986.

[RST94] Dana Ron, Yoram Singer, and Naftali Tishby. The power of amnesia. In Jack D. Cowan, Gerald Tesauro, and Joshua Alspector, editors, *Advances in Neural Information Processing Systems*, volume 6, pages 176–183. Morgan Kaufmann Publishers, Inc., 1994.