A Markov Chain Approach to Reconstruction of Long Haplotypes

L. Eronen, F. Geerts, and H. Toivonen

Pacific Symposium on Biocomputing 9:104-115(2004)

A MARKOV CHAIN APPROACH TO RECONSTRUCTION OF LONG HAPLOTYPES

L. ERONEN, F. GEERTS, H. TOIVONEN
HIIT-BRU and Department of Computer Science, University of Helsinki

Abstract

Haplotypes are important for association based gene mapping, but there are no practical laboratory methods for obtaining them directly from DNA samples. We propose simple Markov models for reconstruction of haplotypes for a given sample of multilocus genotypes. The models are aimed specifically for long marker maps, where linkage disequilibrium between markers may vary and be relatively weak. Such maps are ultimately used in chromosome or genome-wide association studies.

Haplotype reconstruction with standard Markov chains is based on linkage disequilibrium (LD) between neighboring markers. Markov chains of higher order can capture LD in a neighborhood of a given size. We introduce a more flexible and robust model, MC-VL, which is based on a Markov chain of variable order. Experimental validation of the Markov chain methods on both a wide range of simulated data and real data shows that they clearly outperform previous methods on genetically long marker maps and are highly competitive with short maps, too. MC-VL performs well across different data sets and settings while avoiding the problem of manually choosing an appropriate order for the Markov chain, and it has low computational complexity.

1 Introduction

Haplotypes capture information about regions descended from ancestral chromosomes. They are essential for many genetic studies, especially for association (or linkage disequilibrium, LD) based gene mapping: haplotypes can be much more informative than single markers, and they give higher power for assigning a phenotype to a genetic region in association studies [1]. Being able to use haplotypes is particularly important for SNP (single nucleotide polymorphism) markers, which are alone relatively uninformative.

Current practical laboratory techniques provide unphased genotype information for diploids, i.e., an unordered pair of alleles for each marker. Reconstruction of haplotypes from genotype data is then a crucial step in the analysis process. There are two approaches to the problem. One is based on trios: haplotypes are inferred from the genotypes of a subject's parents. This

involves significant additional genotyping costs and potential recruiting problems. Further, in the case of SNPs, on average up to one eighth of the alleles can still remain ambiguous. The second approach is to apply computational or statistical inference to find the most likely haplotype configuration consistent with the observed genotype data. This population-based alternative is fast and cheap and has been recently researched a lot: Clark's parsimony method [2,3], the expectation–maximization (EM) algorithm [4] and its Partition Ligation (PL) variant [5], Phase [6], Haplotyper [7], and the phylogenetic approach [8,9].

We propose and evaluate Markov chain models for population-based haplotype reconstruction and compare them with previous methods. While the existing methods typically assume that each haplotype is descended as a unit from generation to generation, we consider models that better accommodate recombinations. Our approach is motivated by gene mapping studies using genetically long, even genome wide maps [10]. In a typical study for gene mapping by LD, a map of markers is selected from the region of interest, which may span from millions to hundreds of millions of base pairs. For economical reasons, only a sparse subset of all known markers (polymorphisms) in the region is used. Chromosome and even genome-wide association studies are considered to have potential for efficient mapping of common disease genes [10,11].

Instead of estimating frequencies of full haplotypes, like previous models for population-based haplotyping, the Markov chain (MC) models we propose in Section 2 estimate and use frequencies of local haplotype fragments, i.e., shorter regions potentially conserved for several generations and thus more likely to be reliably identifiable in a population sample. The method does not assume haplotype blocks[12] in the population; in a sense our model allows each individual haplotype to have its own structure. Higher adaptivity to the data at hand is obtained by using haplotype fragments of different lengths at different regions, based on the strength of evidence for a fragment to be identical by descent in several haplotypes. We propose a Markov chain model of variable order, MC-VL, to obtain this adaptivity. Models related to the ones proposed in this article have been applied to various other sequence modeling and prediction problems[13,14] but, to our knowledge, not to haplotyping. We provide a hierarchical algorithm for constructing haplotypes (Section 3). Finally, we give an experimental evaluation of the proposed methods under varying linkage disequilibrium and compare the methods with previous techniques (Section 4). For the evaluation we use a wide range of simulated data as well as Daly's data [12]. We conclude in Section 5.

2 Models for Haplotype Reconstruction

Concepts and Notation We assume a set (map) M of ℓ markers $1, \ldots, \ell$ and denote the set of 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 *(multilocus) genotype* G over M is a vector of (unordered) allele pairs: $G \in \prod_{i=1,\ldots,\ell} \{\{a_1, a_2\} \mid a_1, a_2 \in A_i\}$. For SNPs, $|A_i| = 2$. Assuming alleles are labeled "1" and "2", SNP haplotypes are vectors in $(1,2)^\ell$ and SNP genotypes vectors in $(\{1,1\},\{1,2\},\{2,2\})^\ell$. In our terminology, a haplotype thus refers to the alleles in a chromosome over the whole marker map (and not e.g. to a segment descended as such from a founder). In a similar way, here the term genotype refers the data over the whole marker map (and not e.g. to just one marker).

Let $H(i,j)$ denote the sequence from the ith to the jth marker in haplotype H. We call $H(i,j)$ a *(haplotype) fragment*. We will denote $H(i,i)$ simply by $H(i)$. Also, let $G(i,j)$ denote the sequence of allele pairs from the ith to the jth marker in genotype G. Again, $G(i,i)$ is denoted by $G(i)$. Given two 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 and G are *consistent* or that $\{H_1, H_2\}$ is a *(possible) haplotype configuration* for genotype G. Two haplotypes determine a unique consistent genotype in the obvious way. A genotype, on the other hand, can have several haplotype configurations. For a genotype G with k heterozygous markers ($k = |\{i \mid |G(i)| = 2\}|$), there are 2^{k-1} different haplotype configurations. The set of all possible haplotype configurations for a genotype G will be denoted by C_G, with $|C_G| = 2^{k-1}$. Finally, we say that a fragment $H(i,j)$ and a genotype G *match* if there exists a string $\overline{H} \in \prod_{k=i,\ldots,j} A_i$ such that $\{H(i,j), \overline{H}\}$ is consistent with $G(i,j)$.

Breakdown of the Haplotype Reconstruction Problem In this paper we address the *haplotype reconstruction problem*: given a set $\mathcal{G}$ of genotypes the task is to output the most likely haplotype configuration for each genotype $G \in \mathcal{G}$. We assume Hardy-Weinberg equilibrium and use the equation

$$P(\{H_1, H_2\} \mid G) = \begin{cases} \frac{P(H_1)P(H_2)}{\sum_{\{H,\overline{H}\} \in C_G} P(H)P(\overline{H})} & \text{if } \{H_1, H_2\} \in C_G \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

to reduce the problem of estimating the probability of haplotype pairs to estimating the probability of single haplotypes. The genotypes are assumed to come from the same population and thus to share haplotype fragments, based on which the probability of different haplotype configurations can be estimated.

Estimation of Haplotype Fragment Probabilities We estimate the probabilities of haplotype fragments by their frequencies computed from the genotype data $\mathcal{G}$. Whenever a genotype fragment $G(i,j)$ has more than one heterozygous marker, it has several possible haplotype configurations. To compensate for this ambiguity, the matching genotypes are weighted according to their heterozygosity:

$$P(H(i,j)) \approx fr(H(i,j)) = \frac{1}{2\,|\mathcal{G}|} \sum_{\substack{G \in \mathcal{G}, \\ G \text{ matches } H(i,j)}} 2^{1-k_{G(i,j)}}, \tag{2}$$

where $k_{G(i,j)}$ is the number of heterozygous markers in $G(i,j)$ and $fr(\cdot)$ denotes frequency of the parameter. A homozygous genotype has two identical haplotypes both matching the fragment, and thus weight 2. This approach is very simple and in a strong contrast with the previous work on the topic, where the main emphasis is on haplotype frequency estimation.

Markov Chains *Markov chains* are simple models that capture statistical dependence between neighboring alleles:

$$P(H) \approx P(H(1)) \prod_{i=2,\ldots,\ell} P(H(i) \mid H(i-1)).$$

The motivation is that knowing a neighboring allele can tell a lot about the next allele, due to linkage disequilibrium between alleles of nearby markers. We estimate $P(H)$ from frequencies of haplotype fragments of length one and two:

$$P(H) \approx fr(H(1)) \prod_{i=2,\ldots,\ell} \frac{fr(H(i-1,i))}{fr(H(i-1))}. \tag{3}$$

The obvious shortcoming of this model is that although linkage is strongest between neighbors, a neighborhood of several markers is more informative and can show stronger LD.

Markov chains of order d (MC-d) are a more powerful alternative:

$$P(H) \approx P(H(1,d)) \prod_{i=d+1,\ldots,\ell} P(H(i) \mid H(i-d,i-1)). \tag{MC-d}$$

Here d can be used to tune the size of the neighborhood. With $d = 1$ we obviously have the standard Markov chain as a special case. To estimate $P(H)$ we compute the set $\mathcal{F}_d$ of all haplotype fragments of size d and $d+1$ and use their frequencies as in formula (3).

Variable Order Markov Chains Markov chains of variable order aim at adjusting the size of the neighborhood for each marker and haplotype individually. Informally, the goal is to use haplotype fragments that maximize the informativeness of LD. The exact model we propose is a *Markov chain of variable order determined by longest fragments (MC-VL)*. For this model, we compute the set $\mathcal{F}_{\text{vl}}$ of the N most frequent haplotype fragments:

$$\mathcal{F}_{\text{vl}} = \{H(i,j) \mid fr(H(i,j)) > fr(H') \text{ for all } H' \notin \mathcal{F}_{\text{vl}}\},\ |\mathcal{F}_{\text{vl}}| = N.$$

The idea is that we always use the longest fragments in $\mathcal{F}_{\text{vl}}$ from which we estimate the probabilities:

$$P(H) \approx P(H(1)) \prod_{i=2,\ldots,\ell} P(H(i) \mid H(s_i, i-1)), \qquad \text{(MC-VL)}$$

where $s_i = \min\{s | H(s,i) \in \mathcal{F}_{\text{vl}}\}$. In an area where there are long frequent fragments, the order of the Markov chain will be high. Since these fragments are frequent they are more likely to be identical by descent and thus are evidence for the haplotype to be reconstructed.

Handling Missing Data In real applications, marker data is often missing, due to changes in the marker map during the study, or due to genotyping problems. The MC methods can be extended to handle missing data with the following two modifications (we assume that either both alleles of a marker are known or both are missing).

First, the estimation of fragment frequencies needs to be adjusted so that information in genotypes with missing data is included. This is done by distributing the probability mass of a genotype over all the fragments obtained by imputing possible alleles at the missing markers, weighted by the frequencies of the alleles. Recall the frequency estimate $fr(H(i,j))$ in Equation 2, and let G match $H(i,j)$ if they match in all markers where G does have data. Then $fr'(H(i,j))$, the frequency estimate when G can have missing data, is defined as

$$fr'(H(i,j)) = fr(H(i,j)) \prod_{\substack{m \in [i,j], \\ G(m) \text{ is missing}}} fr(H(m)),$$

where $fr(H(m))$ is the frequency of allele $H(m)$.

Second, to reconstruct haplotypes for genotypes with missing data, probabilities (frequencies) need to be estimated for fragments $H(i,j)$ that potentially have missing data (no alleles are imputed, though). The estimate is the sum of the frequency of all fragments $H'(i,j)$ in $\mathcal{F}_{\text{vl}}$ or $\mathcal{F}_d$ that match $H(i,j)$ wherever it has data.

3 Haplotype Reconstruction Algorithm

The number of haplotype configurations for a genotype grows exponentially with the number of heterozygous markers, so exhaustive search is feasible only for small marker maps. If the marker map is long, we use use a hierarchical "partition ligation" (PL) search strategy, motivated by a similar strategy used by Niu et al. [7]. We use MC-VL as probability model in the description of our haplotype reconstruction algorithm (Figure 1). It is obvious how to adapt the algorithm for use with MC-d.

Given a set of genotypes $\mathcal{G}$ the algorithm HaploRec computes the fragment frequencies and then uses the PL strategy to search a subspace of all possible configurations for each genotype G individually. First, the Partition procedure recursively splits G until the genotype fragments consist of at most $\ell_{\max}$ markers. $\ell_{\max}$ is chosen such that the evaluation of all possible configurations of a fragment of length $\ell_{\max}$ is computationally feasible. In our experiments, we have used $\ell_{\max} = 8$. When $\ell_{\max} = \ell$, i.e., the total number of markers, then the algorithm performs the exhaustive search strategy.

HaploRec($\mathcal{G}$)
Compute the set $\mathcal{F}_{\text{vl}}$ of fragments and their frequencies using Equation (2);
for each $G \in \mathcal{G}$ **do**
  Output the most probable element of Partition(G);

Partition(G)
if $|G| \leq \ell_{\max}$ **then**
  Compute the set C_G of all haplotype configurations for G;
  Estimate their probabilities using Equations (1) and (MC-VL);
  Output the B most probable elements of C_G;
else
  $\mathcal{H}_1 =$ Partition($G(1, |G|/2)$);
  $\mathcal{H}_2 =$ Partition($G(|G|/2 + 1, |G|)$);
  $\mathcal{H} =$ Ligate($\mathcal{H}_1, \mathcal{H}_2$);
  Estimate the probabilities of elements of $\mathcal{H}$ using Equations (1) and (MC-VL);
  Output the B most probable elements of $\mathcal{H}$;
end if

Ligate($\mathcal{H}_1, \mathcal{H}_2$)
for each $\{H_1, \overline{H}_1\} \in \mathcal{H}_1$ and $\{H_2, \overline{H}_2\} \in \mathcal{H}_2$ **do**
  Output $\{H_1H_2, \overline{H}_1\overline{H}_2\}$ and $\{H_1\overline{H}_2, \overline{H}_1H_2\}$;

Figure 1: Haplotype reconstruction algorithm using probability model MC-VL and hierarchical partition ligation.

Once G is partitioned into small fragments, the B most probable haplotype configurations from all possible configurations according to MC-VL are obtained for each fragment. On all other levels of recursion, the LIGATION procedure produces $2B^2$ haplotype configurations by joining configurations for shorter fragments, obtained from the deeper recursion level and returns the B most probable ones. In the end, we obtain the B most probable haplotype configurations for the full genotype.

The method is greedy and not guaranteed to find the haplotype configuration with the largest probability. It is possible, although not likely, that a fragment of the most probable configuration is not among the B most likely fragments, and thus the global optimum is not found. However, in our experiments with $B = 10$ this was rarely the case.

Both MC-VL and MC-d have linear time complexity in $|\mathcal{G}|$; are exponential in $\ell_{\max}$, quadratic in B, and subquadratic in ℓ; MC-d is exponential in d. The space complexity of MC-VL is linear in N; MC-d is again exponential in d.

4 Experimental Results

Test setting We used simulated data sets in order to be able to perform controlled experiments. The setting corresponds to an association study in a population isolate. We simulated a population with effective founder population of size 20 (20 founders each with independent random haplotypes with uniformly distributed alleles). The population then expanded for 20 generations with random mating, leading to a final population of 100000 individuals. We used a sample of 500 genotypes, drawn randomly and independently from the last generation. We experimented separately with biallelic markers (SNPs) and 6-allele markers (microsatellites).

In our experiments we used a marker map of 32 evenly spaced markers. The major parameter varied in the experiments was the distance between adjacent markers: it ranged between 0.01 and 1 cM. The simulated chromosomal regions had, respectively, genetic lengths between 0.31 and 31 cM. We ran 10 independent population simulations for each of the different marker spacings and report results averaged over the 10 simulations.

In data sets and populations like the ones simulated, recombination is practically the only factor affecting haplotype (fragment) sharing between individuals in the final population. In 20 generations, 0.062–6.2 recombinations are expected per genotype for regions of length 0.31–31 cM, so reasonable mixing and fragmentation of founder haplotypes can be expected with the longer regions simulated. Marker mutations are unlikely in 20 generations and 32 markers, and were ignored in the simulation.

As a dense and real benchmark data we use the public Daly set [12] which consists of 129 genotyped trios from a European derived population. The map consists of 103 SNPs ranging over 500 kb located on chromosome 5q31 (Crohn's disease). We inferred the haplotypes of 129 children from pedigree data and used the nontransmitted chromosomes as an extra 129 (pseudo) haplotype pairs. Markers for which both alleles could not be inferred were marked as missing. From the set of 258 genotypes, the ones with more than 20% missing alleles were removed, leaving 147 genotypes in the final test set.

We measure the performance of the methods by the average number of switches ("recombinations") needed in the computer-generated haplotype configuration to recover the original haplotype configuration [15]. Switch distance is a natural error measure for this problem: many applications using inferred haplotypes will look at local haplotype segments and they are correct unless one of the needed switches is within the segment.

For benchmarking, we used available implementations of Phase [6], Snphap (see D. Clayton's website) and PL-EM [5]. We used default parameters where possible. For Phase, no step-wise mutation model was assumed, the number of iterations and burn-in iterations were both set to 10000, and the thinning interval was 100. For PL-EM, we set buffer size to 50, number of iterations to 20, and parsize to 1, as in our case a lot of haplotype diversity was assumed to be present.

We did not succeed in running our experiments with Haplotyper [7] (version 1.0, linux). Haplotyper worked fine for smaller test data sets, but terminated with an error in most of the data sets that were used in our experiments.

Evaluation of the models The performance of the methods is illustrated in Figure 2. Results with SNP data sets are on the left, with microsatellites on the right. The first row shows the performance of different Markov chain models, as a function of the marker map density. An immediate observation is that as markers are more sparsely spaced, the problem becomes more difficult and the error increases, as expected.

A useful and positive result is that the problem is solvable with quite a small error. Best models have switch distances between 0 and 3.5 (MC-10 and MC-VL, SNP data) or 0 and 2 (MC-4 and MC-VL, microsatellites), practically linear in the marker spacing. The results are excellent, less than 0.5 switches with SNP data for marker spacings 0.01–0.15 cM.

Markov chain models MC-d of a fixed order give mixed results. With $d = 1$, i.e., the standard Markov chain, the results are poor. With a growing d, results first improve but later deteriorate for sparse maps (see especially MC-12 for SNPs and MC-5 for microsatellites). This is due to overfitting as d

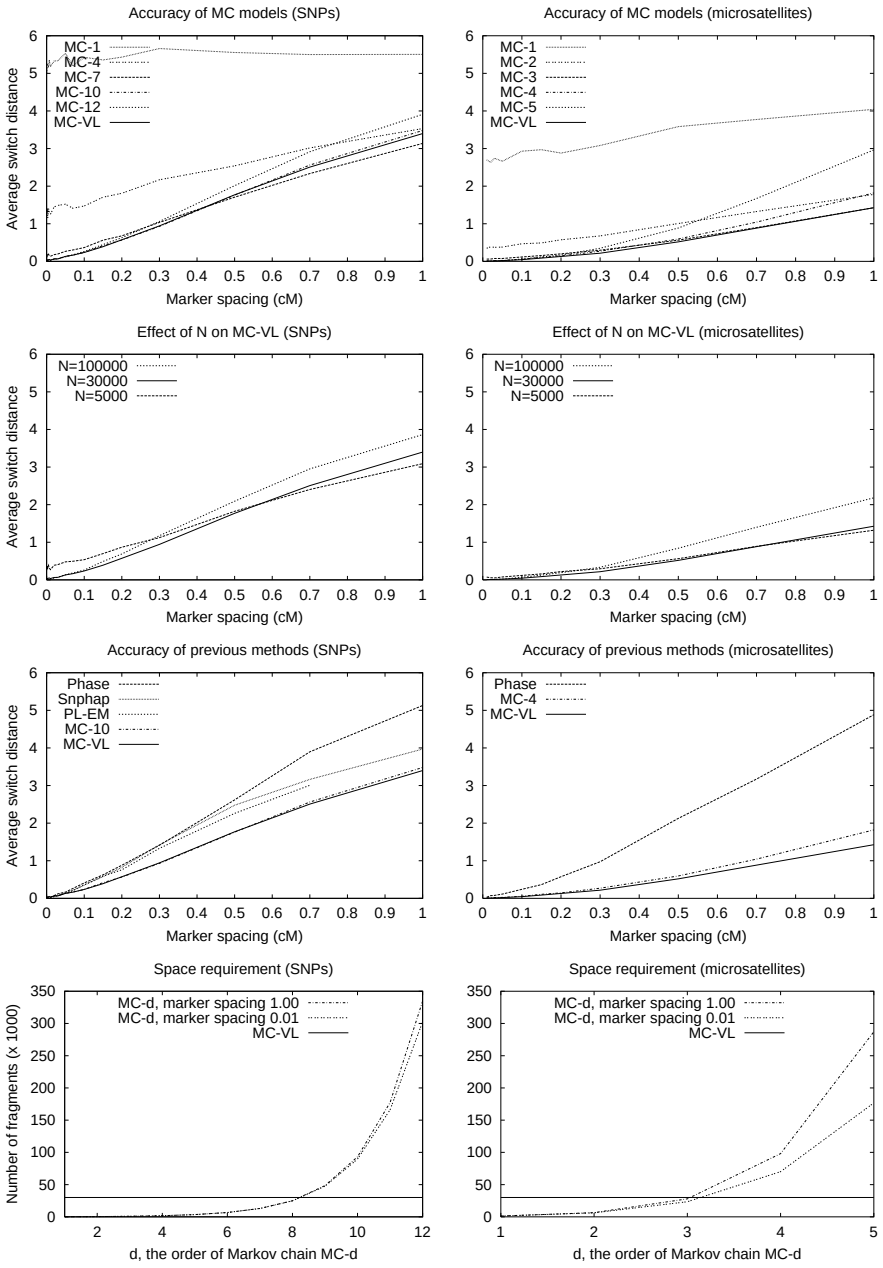

Figure 2: Experimental evaluation of proposed methods on simulated data.

is larger than the informative neighborhood of a marker. The suitable value of d can be quite different not only for different marker densities but also for different datasets: in our data, $d = 10$ is a good choice for SNPs, and $d = 4$ for microsatellites. Further, in a real data set with a less systematic marker map, no single value of d would necessarily be suitable across the whole map.

The second row of Figure 2 tests the robustness of MC-VL to N, the number of most frequent fragments used. The tested range is from 5000 to 100000, and MC-VL shows quite stable behavior, especially if contrasted to the large variance in results obtained with MC-d models. In all other figures we have used $N = 30000$ most frequent fragments.

A comparison to state-of-the-art methods is provided on the third row of Figure 2: Phase, Snphap, and PL-EM are applied to the same data sets, and results of MC-VL and MC-10 (SNPs) or MC-4 (microsatellites) are included for comparison. (The available implementations of Snphap and PL-EM assume SNP data and could not be run with microsatellites. The implementation of PL-EM failed for 1 cM marker spacing, resulting in a missing data point.) The performance of MC models is solid throughout the different settings and superior over previous models on marker densities larger than 0.05 cM.

A comparison on the Daly data set shows that the MC models are most competitive also with dense, real data sets with missing data. MC-VL and MC-9 outperform Snphap in terms of switch distance (0.90, 0.93, and 1.29, respectively). Switch distance could not be measured for Phase, as it often gave haplotype configurations not consistent with the observed genotype data. If the accuracy is measured in terms of haplotypes that are not completely correct, then Snphap, MC-VL, and MC-9 outperform Phase with a clear margin (0.41, 0.45, 0.48, and 0.97, respectively). (PL-EM did not complete in few days.)

The bottom row of Figure 2 illustrates the space requirements of the MC models, in number of haplotype fragments stored, for the simulated data sets. MC-VL has a constant space requirement, whereas the MC-d models have roughly exponential space requirement in d.

The running times of MC-VL ranged from 70 to 140 seconds for both SNPs and microsatellites, depending on N. The time requirement of MC-d is proportional to its space requirement, i.e., exponential in d. With the values of d reported in Figure 2, the running times varied between 40 and 120 seconds. Larger values took too long for repeated experimental testing. In comparison, Snphap takes around 2 to 20 seconds for the current data sets, PL-EM 3 to 100 seconds, and Phase between 5 hours (SNPs) and 30 hours (microsatellites). All experiments were run on a PC with an AMD 1400 Mhz processor. The MC models were implemented in Java, other implementations were provided by their authors.

5 Conclusion

We proposed Markov chain models for the haplotype reconstruction problem, motivated by association studies with wide marker maps. We experimentally tested the performance on simulated and real data. Normal Markov chains (of order $d = 1$) did not perform well. Higher order Markov chains did, but a suitable order d needs to be found for each data set. Variable order Markov chains (MC-VL) showed consistently good behaviour.

In experimental tests the MC models outperformed previous methods with sparse maps and were most competitive with dense maps, too. With SNPs the margin is clear and the switch distance of MC models is tens of percents smaller; with microsatellites the switch distance is less than half of Phase's. The wide applicability of the MC models was demonstrated on real data.

Why do the MC models perform well on sparse data? Previous haplotyping methods that are based on estimating haplotype frequencies are not well suited for situations where many haplotypes are unique. In the simulated setting, almost half of the haplotypes (480/1000) are unique with marker density 0.2 cM; with a density of 0.5 cM there are already 828 unique haplotypes. Estimating frequencies of haplotypes that occur only once is obviously difficult.

Among the MC models, MC-VL has some nice properties. It seems to adjust for a suitable neighborhood, and the user does not need to worry about setting the order d of a Markov chain; the model is not sensitive to the selection of its model parameter N. The computational complexity is low and predictable, compared to the exponential time and space of MC-d in d.

Our future work will include improved methods for estimating fragment probabilities. A promising idea is to use an iterative approach similar to EM. The performance of different components of the solutions could be evaluated: it is not fully clear which fraction of errors is due to fragment frequency estimation, which is due to models, and which to the heuristic search strategy. Probably each component has room for improvement. The effect of the haplotype reconstruction algorithm on the subsequent analysis, especially haplotype-based gene mapping, remains yet to be evaluated systematically.

An implementation of the methods introduced in this article is available at `http://www.cs.helsinki.fi/group/genetics/haplotyping.html`.

Acknowledgments The population simulator was provided by Vesa Ollikainen. We thank authors of Phase, Snphap and PL-EM implementations for kindly making their programs available.

References

1. Joshua Akey, Li Jin, and Momiao Xiong. Haplotypes *vs* single marker linkage disequilibrium tests: what do we gain? *European Journal of Human Genetics*, 9:291–300, 2001.
2. Andrew G. Clark. Inference of haplotypes from PCR-amplified samples of diploid populations. *Molecular Biological Evolution*, 7:111–122, 1990.
3. Dan Gusfield. Inference of haplotypes from samples of diploid populations: Complexity and algorithms. *Computational Biology*, 8:305–324, 2001.
4. Laurent Excoffier and Montgomery Slatkin. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Molecular Biological Evolution*, 12(5):921–927, 1995.
5. Tianhua Niu, Zhaohui S. Qin, and Jun S. Liu. Partition-ligation-expectation-maximization algorithm for haplotype inference with single-nucleotide polymorphisms. *The American Journal of Human Genetics*, 71:1242–1247, 2002.
6. Matthew Stephens, Nicholas J. Smith, and Peter Donnelly. A new statistical method for haplotype reconstruction from population data. *The American Journal of Human Genetics*, 68:978–989, 2001.
7. 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.
8. Dan Gusfield. Haplotyping as perfect phylogeny: conceptual framework and efficient solutions. In *Proceedings of the sixth annual international conference on Computational biology*, pages 166–175. ACM Press, 2002.
9. E. Eskin, E. Halperin, and R. M. Karp. Large scale reconstruction of haplotypes from genotype data. In *Proceedings of the seventh annual international conference on Computational biology*, pages 104–113. ACM Press, 2003.
10. Neil Risch and Kathleen Merikangas. The future of genetic studies of complex human diseases. *Science*, 273:1516–1517, 1996.
11. L Kruglyak. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genetics*, 22:139–144, 1999.
12. 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.
13. Mukund Deshpande and George Karypis. Evaluation of techniques for classifying biological sequences. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 417–431, 2002.
14. 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.
15. Shin Lin, David J. Cutler, Michael E. Zwick, and Aravinda Chakravarti. Haplotype inference in random population samples. *The American Journal of Human Genetics*, 71:1129–1137, 2002.