The Second International Symposium on Optimization and Systems Biology (OSB'08)
Lijiang, China, October 31– November 3, 2008
Copyright © 2008 ORSC & APORC, pp. 212–223

Improved Approach for Haplotype Inference Based on Markov Chain*

Ling-Yun Wu[1,†] Ji-Hong Zhang[2] Raymond Chan[3]

[1] Institute of Applied Mathematics, Academy of Mathematics and Systems Science,
Chinese Academy of Sciences, Beijing 100190, China
[2] School of International Business, Beijing Foreign Studies University, Beijing 100089, China
[3] Department of Mathematics, Chinese University of Hong Kong, Hong Kong

Abstract Variable-order Markov model (VMM) is an important statistical method for haplotype inference problem. It is well-suited for sparse marker maps and large-scale data. The existing algorithm, HaploRec, solves VMM by a greedy algorithm with pruning strategy. We present an improved Expectation-Maximization (EM) algorithm for VMM, which is based on dynamic programming (DP). The computational experimental results with simulated and real data show that the proposed algorithm can greatly improve the accuracy of VMM with an acceptable running time. The methods described in this paper are implemented in a software package, HMC, which is available from the internet.

Keywords Haplotype Inference; SNP; Markov Chain; Dynamic Programming.

1 Introduction

In many genetic studies such as association studies of common complex disease, haplotypes provide higher power for assigning a phenotype to a genetic region than individual SNP (single nucleotide polymorphism) markers since they can capture more information about regions descended from ancestral chromosomes [1, 14]. However current laboratory methods for obtaining haplotypes directly from DNA samples are not practical and time consuming and expensive [17, 19]. Many computational methods have been developed to reconstruct haplotypes from unphased genotype data, which are based on statistical dependency between neighboring markers. Some of them are combinatorial models, e.g. Clark's parsimony method [2, 8] and the phylogenetic approach [8, 6], while some are the statistical models, e.g. the expectation-maximization (EM) algorithm [7] and its Partition Ligation (PL) variant [12], PHASE [16], Haplotyper [11]. We refer the interested readers to the review paper [9] and the references therein for more details about these methods.

In this paper, we focus on the variable-order Markov models which is first proposed by Eronen *et al.* in [4]. In these models, the frequency of each haplotype is estimated

[*] This work is supported by National Natural Science Foundation under Grant No. 60503004 and 10631070.

[†] Corresponding author. Email: lywu@amt.ac.cn. The work of first author is partly supported by Department of Mathematics, Chinese University of Hong Kong, Hong Kong.

from short fragments, which implicitly considers varied recombination rates throughout long full haplotype. Since the fragments are shorter regions potentially conserved for several generations, they are more likely to be reliably identifiable in a population sample. However, in [4], the algorithm used to reconstruct the haplotype resolutions with maximum likelihood is a greedy algorithm. The algorithm applies a partition-ligation (PL) pruning strategy on the possible haplotype resolutions, hence produces only near-optimal solutions and without any guarantee of solution quality. In [18], we proposed a dynamic programming (DP) method which can solve the problem optimally. The computational experiments in [18] show that the solution of PL method maybe far away from true optimum in some cases such as large marker spacing. Therefore, DP method can greatly improve the results in [4], while the time and space complexity remains in same low magnitude of PL method. Recently, Eronen *et al.* [5] introduced an expectation-maximization (EM) algorithm to iteratively improve the accuracy of fragment frequencies estimation. This algorithm re-estimates the fragment frequencies from the combined set of the most probable haplotype configurations for all genotypes, which is sampled in previous step. In this paper, we will show that, by using DP method proposed in [18], the fragment frequencies can be re-estimated more accurately based on all possible haplotype configurations for all genotypes, which will further improve the accuracy of haplotype inference.

2 Notations

Each diploid individual has two nearly identical copies of each chromosome and hence of each region of interest. A description of the alleles of markers on a single copy is called a *haplotype*, while a description of the conflated alleles (i.e. the unordered pair of alleles for each marker) on the two homogenous copies of chromosomes is called a *genotype*. We denote a set (map) of the n markers by $\mathscr{M} = \{1, 2, \cdots, n\}$ and the set of alleles of marker i by $\mathscr{A}_i$. Then the haplotypes are vectors in $\prod_{i=1}^{n} \mathscr{A}_i$ and the genotypes are vectors in $\prod_{i=1}^{n}\{\{a_1, a_2\}|a_1, a_2 \in \mathscr{A}_i\}$. For SNPs, $|\mathscr{A}_i| = 2$ and alleles are often labeled “0” (wild type) and “1” (mutant type).

For a haplotype H, a *haplotype fragment* $H(i, j)$ denotes the allele sequence from the ith to the jth marker in H. The $H(i,i)$ will be denoted simply by $H(i)$. Similarly, a genotype fragment $G(i, j)$ denotes the sequence of allele pairs from the ith to the jth marker in genotype G and $G(i)$ represents $G(i,i)$. For a haplotype fragment $H(i, j)$ denoted by h, $h(k)$ is defined as $H(k)$ for $i \leq k \leq j$, and undefined elsewhere. The genotype fragment $g = G(i, j)$ is defined similarly. In this paper, we always use capital letters to denote full haplotypes (genotypes) while small letters denote haplotype (genotype) fragments. Denote the first and last marker at which the fragment h is defined by $start(h)$ and $end(h)$ respectively. We only consider consecutive fragment, i.e., $h(i)$ is defined for any i such that $start(h) \leq i \leq end(h)$.

For a haplotype fragment h and a genotype fragment g, h is said to be *consistent* with g (denoted by $h \in g$), if $h(k) \in g(k)$ for all k such that $h(k)$ and $g(k)$ are both defined. For a haplotype H and a genotype G such that $H(1,n) \in G(1,n)$, H is called a consistent haplotype of G and also denoted by $H \in G$. The set of all consistent haplotypes of a genotype G is denoted by $\mathscr{H}(G)$. For a set of genotypes $\mathscr{G}$, let $\mathscr{H}(\mathscr{G}) = \bigcup_{G \in \mathscr{G}} \mathscr{H}(G)$.

Given two haplotypes H_1, H_2 and a genotype G such that $\{H_1(i), H_2(i)\} = G(i)$ for all $i \in \mathscr{M}$, the haplotype pair $\{H_1, H_2\}$ is called a *haplotype configuration* for genotype G.

A genotype may have many haplotype configurations. Namely, there are 2^{k-1} different haplotype configurations for a genotype G with k heterozygous markers, i.e. the markers that have different alleles in two homogenous chromosomes. Conversely, two haplotypes uniquely determine a genotype. The set of all possible haplotype configurations for a genotype G is denoted by $\mathscr{C}(G)$.

The *haplotype inference problem* addressed in this paper is as follows: **given a set of genotypes $\mathscr{G}$, find the most likely haplotype configuration for each genotype $G \in \mathscr{G}$.**

Given a set of genotype $\mathscr{G}$, the haplotype inference problem is to find a pair of haplotypes H_1 and H_2 for each $G \in \mathscr{G}$ such that $\{H_1, H_2\} \in \mathscr{C}(G)$ and the probability of haplotype configuration $P(\{H_1, H_2\}|G)$ is maximized over $\mathscr{C}(G)$. Under the assumption of Hardy-Weinberg equilibrium, the evaluation of the probability of haplotype configuration is reduced to estimating the probabilities of single haplotypes as follows.

$$P\big(\{H_1,H_2\}|G\big) = \begin{cases} \dfrac{P(H_1)P(H_2)}{\sum_{\{H,H'\}\in\mathscr{C}(G)} P(H)P(H')} & \text{if } \{H_1,H_2\} \in \mathscr{C}(G), \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

3 Results

The improved algorithm proposed in this paper is implemented as a part of software package HMC [1]. In order to report comparative performance of the proposed method, several sets of simulated and real data are used in the computational experiments. All experiments are run on a Pentium D processor 3.4GHz without using any parallel technique. Both simulated and real datasets are used in the experiments.

The simulated datasets are the datasets used in Marchini *et al.* [10], which are developed as benchmark datasets to evaluate haplotype inference methods. While most of the datasets are distributed without the answers, a set of trial datasets with answer files is available on their website[2]. The trial datasets used in this paper are ST1, ST2 and ST3. There are 20 datasets in each datasets with same simulation model. ST1 datasets are simulated with constant recombination rate across the region, constant population size, and random mating. ST2 datasets are same as ST1, but with the addition of a variable recombination rate across the region. ST3 datasets are same as ST2, except a model of demography consistent with white Americans was used. The datasets are trios data, but only the data of parents are used in this paper as unrelated individuals.

The real dataset is the public Daly set [3], which is a real genotype set with missing data from a European derived population. The map consists of 103 SNPs ranging over 500kb on chromosome 5q31 (Crohn's disease). After removing the genotypes with more than 20% missing alleles, 147 genotypes are obtained in the final test set. The Daly dataset used here is same as that in Eronen *et al.* [4] and Zhang *et al.* [18].

Three different criteria are used to assess the performance of methods: Switch error, IHP and IGP [10]. Switch error is the percentage of possible switches in inferred haplotype configuration to recover the original one (i.e. the correct haplotype configuration). IHP (incorrect haplotype percentage) is the percentage of ambiguous individuals that haplotype configuration is not correctly inferred. IGP (incorrect genotype percentage) is the

[1] The version of HMC used in this paper is v0.8.

[2] http://www.stats.ox.ac.uk/ marchini/phaseoff/trial.data.tgz

percentage of ambiguous genotypes that are not correctly phased. We refer users to [10] for more details.

The detail of computational experiments results are shown in Tab. 1. Haplorec is the software developed by Eronen *et al.* [4, 5]. The version of haplorec used in the experiments is 2.0, in which the algorithm in [5] is implemented. In order to compare with haplorec, the parameters of HMC are selected in such way that the number of generated fragments (patterns) is no more than that by haplorec in all instances. The experimental results in Tab. 1 show that the HMC archives better results than haplorec in most cases, whenever EM procedure is used or not.

The comparison results of two methods are illustrated in Tab. 2. When EM procedure is not used, the results of HMC is averagely 7.97%, 2.38%, 6.99% better than that of haplorec in terms of switch error, IHP, IGP respectively. If EM procedure is used, the improvement of HMC over haplorec is more significant: 15.38%, 6.73%, 12.19% in terms of switch error, IHP, IGP respectively. While considering the effect of EM procedure, haplorec with EM is 28.42%, 12.49%, 19.32% better than haplorec without EM in terms of switch error, IHP, IGP respectively. HMC with EM is 32.64%, 16.44%, 23.08% better than haplorec without EM in terms of switch error, IHP, IGP respectively. In other words, HMC can exert more power of EM procedure than haplorec in improving the accuracy of variable-order Markov model, which is due to the precise estimation of fragment frequencies in maximization step. For the same reason, it is also observed in the experiments that the EM iterations of HMC are less than that of haplorec.

The average computation time of different methods are shown in Tab. 3. When EM procedure is not used, the average running time of HMC is same magnitude order or even less than that of haplorec. If using EM, the average running time of HMC is larger than that of haplorec, but still in the same magnitude order.

The experimental results of variable-order Markov model are also compare with that of PHASE [15], the best algorithm for haplotype inference, in Tab. 1. The result of HMC with EM is worse than that of PHASE, but very close. Considering the huge running time of PHASE (see Tab. 3), the performance of HMC is acceptable. Therefore, the variable-order Markov model based methods (HMC, haplorec) can be used as a fast haplotype inference tool, complemented to PHASE and other methods.

4 Conclusion

In this paper, we focus on the variable-order Markov models for haplotype inference problem studied in [4, 5, 18]. Due to the exponential number of possible haplotype configurations, Eronen *et al.* [5] solve VMM by a greedy algorithm with pruning strategy. They also note that, for the FMM, the reconstruction of haplotype configurations with maximal probability can be solved by an adaptation of well-known Viterbi algorithm, and the EM procedure can be improved by an adaptation of Baum-Welch algorithm. However, they are unclear whether the two approaches can be extendable to the VMM (see Additional file 1 of [5]). We answer this question positively in this paper. Namely, for most well-defined haplotype fragment set, e.g. $\mathscr{F}_{minfr}$ used in [4, 5, 18], the VMM can be solve by an modified version of Baum-Welch algorithm, which will improve the accuracy. The computational experiments on simulated and real data show that the presented method outperforms Haplorec [5] in terms of solution quality for reconstructing haplotypes. Al-

Table 1: Experimental results of methods

Error measure	Error rate (%)				
and datasets	haplorec (no EM)	HMC (no EM)	haplorec	HMC	PHASE
Switch error:					
ST1	0.067489	0.065044	0.040952	0.034140	0.019873
ST2	0.127536	0.128040	0.074804	0.057529	0.034067
ST3	0.095310	0.090634	0.073912	0.067101	0.056260
Daly	0.047186	0.032532	0.039859	0.028722	0.024055
IHP:					
ST1	0.819066	0.817434	0.656109	0.619928	0.380943
ST2	0.979096	0.983263	0.833828	0.766271	0.532740
ST3	0.899562	0.908856	0.840777	0.816638	0.682641
Daly	0.554688	0.484375	0.546875	0.453125	0.441860
IGP:					
ST1	0.069697	0.070191	0.051322	0.046936	0.022741
ST2	0.095242	0.094798	0.069643	0.059467	0.034046
ST3	0.081373	0.080935	0.075167	0.071054	0.058249
Daly	0.029883	0.019502	0.024322	0.016462	0.014926

though the new approach is more complicated and time-consuming than the method in [5], it is deserved for the improved reconstruction accuracy, and the computation times for moderate scale problems are acceptable.

5 Methods

5.1 Markov models

Since the genotypes are assumed to come from the same population, there exists linkage disequilibrium between neighboring markers and the real haplotype configurations are believed to share some common haplotype fragments in some meaning. The motivation of Markov models is to model the local disequilibrium by regard the haplotype as a Markov chain [4, 18, 5]. By estimating the transition probabilities from fragment frequencies, the probability of single haplotype can be computed from the haplotype fragment probabilities. In this paper, two Markov chain models are studied: the fixed-order Markov model (FMM) and the variable-order Markov model (VMM). Herein, we briefly introduce two models and refer readers to [4, 18, 5] for more details.

In FMM, the conditional probability for an allele in a maker i depends only on the preceding d marker(s):

$$P(H) = P\big(H(1,d)\big) \prod_{i=d+1,\cdots,n} P\big(H(i)|H(i-d,i-1)\big).$$

For $d = 1$, this is a standard Markov chain. The VMM is different from FMM in that the order of Markov chain, d, varies for each haplotype and each position:

$$P(H) = P\big(H(1)\big) \prod_{i=2,\cdots,n} P\big(H(i)|H(s_i,i-1)\big),$$

Table 2: Comparison results of methods

Error measure	Improvement of HMC over haplorec (%)		Improvement of EM over no EM (%)	
and datasets	no EM	EM	haplorec	HMC
Switch error:				
ST1	3.62	16.63	39.32	47.51
ST2	-0.40	23.09	41.35	55.07
ST3	4.91	9.22	22.45	25.96
Daly	31.06	27.94	15.53	11.71
Average	7.97	15.38	28.42	32.64
IHP:				
ST1	0.20	5.51	19.90	24.16
ST2	-0.43	8.10	14.84	22.07
ST3	-1.03	2.87	6.53	10.15
Daly	12.68	17.14	1.41	6.45
Average	2.38	6.73	12.49	16.44
IGP:				
ST1	-0.71	8.55	26.36	33.13
ST2	0.47	14.61	26.88	37.27
ST3	0.54	5.47	7.63	12.21
Daly	34.74	32.32	18.61	15.59
Average	6.99	12.19	19.32	23.08

where s_i is the smallest value such that haplotype fragment $H(s_i, i-1)$ is in a pre-determined fragment set $\mathscr{F}$. Obviously, FMM can be viewed as a special case of VMM. When $\mathscr{F}$ is the set of all haplotype fragments of length $\leq d$, VMM degenerate to FMM of order d. Hence, we only consider VMM in the rest of this paper and the method and conclusions are also applicable to FMM. The often used fragment set of VMM is the set of frequent haplotype fragments, namely, the set of fragments which have a frequency exceeds the given threshold [4, 5, 18], denoted by $\mathscr{F}_{minfr}$.

The parameters of Markov models are the transition probabilities, which are derived from haplotype fragment frequencies as follows:

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

where $F(h)$ is the estimated frequency of fragment h. Therefore, we denote the model parameters λ as the fragment frequencies. The model parameters are learned from data by EM method.

5.2 EM method

The EM algorithm is applied in many haplotype inference methods [7, 12, 5]. The EM algorithm proposed in this paper is similar as that in [5]. In the EM framework, the haplotype configurations of each genotype are considered as latent variables, and the goal

Table 3: Running time of methods

	Average running time (seconds)				
Datasets	haplorec (no EM)	HMC (no EM)	haplorec	HMC	PHASE
ST1	2.772750	0.830100	15.874550	85.302050	18868.164700
ST2	2.704750	0.769200	19.063200	139.308550	32933.549100
ST3	2.753900	0.830050	16.267700	97.438950	53908.853100
Daly	5.861000	6.629000	35.127	760.702000	22523.972000

is to find a maximum likelihood estimate for the model parameters. The likelihood of a genotype dataset is the product of likelihood over all genotypes, while the likelihood of each genotype is the sum of probabilities of all its possible haplotype configurations. Namely, the total likelihood of whole data is:

$$L(\mathscr{G}|\lambda) = \prod_{G\in\mathscr{G}} \left(\sum_{\{H_1,H_2\}\in\mathscr{C}(G)} P\big(\{H_1,H_2\}|\lambda\big) \right), \tag{2}$$

where the probability of the haplotype configuration $\{H_1,H_2\}$, given the model parameters λ, is obtained as follows:

$$P\big(\{H_1,H_2\}|\lambda\big) = P(H_1|\lambda)P(H_2|\lambda)$$

according to the assumption of Hardy-Weinberg equilibrium.

The EM algorithm find a local maximum likelihood by iteratively computing the values of latent variables (Expectation step) and estimating the model parameters (Maximization step). The brief procedure is described as follows:

Step 1. Start with initial guesses of the parameters λ_0;

Step 2. Expectation step: compute the likelihood and the values of latent variables based on current estimate of parameters λ_k;

Step 3. Maximization step: determine the new estimate of parameters λ_{k+1};

Step 4. Iterate steps 2 and 3 until convergence.

Step 5. Output the reconstructed haplotypes, for each genotype, by selecting the configurations with maximum probability.

5.3 Reformed model

Exhaustively computing all possible configurations for a genotype is time consuming since the numbers of possible configurations are growth exponentially with the numbers of heterozygous makers. Eronen *et al.* [5] use a greedy algorithm with pruning strategy to find a set of most probable configurations, for each genotype, maker by marker. There are two disadvantages of the method in [5]. First, the algorithm is a greedy algorithm. That is, the set of configurations returned by the algorithm are not guaranteed exactly the most probable ones. Hence, there is chance that some haplotype configurations with higher probabilities are missing. Second, only a small fraction of haplotype configurations for each genotype are explored, while all rest configurations are assumed having zero probability. Therefore, the consequent re-estimate of parameters λ will become inaccurate, which may affect the final results and convergence of the algorithm.

Based on the preliminary work in [18], we can solve the problem more accurate by a DP algorithm. In [18], we show that, for most well-defined haplotype fragment set $\mathscr{F}$ (e.g. $\mathscr{F}_{minfr}$ and the fragment set for FMM), each haplotype $H \in \mathscr{H}(\mathscr{G})$ can be exactly represented by a unique Markov chain of n fragments in $\mathscr{F}$, $h_1, h_2, \cdots, h_n$, which is denoted by $H = (h_1, h_2, \cdots, h_n)$. And the transition probability is defined as follows:

$$P(h_i|h_{i-1}) = \begin{cases} \frac{F(h_i)}{F(o(h_i,h_{i-1}))}, & \text{if } h_i \text{ is a successor of } h_{i-1}, \\ 0, & \text{otherwise,} \end{cases}$$

where $o(h_i, h_{i-1})$ is the overlapping fragment of h_i and h_{i-1}. For a fragment h and an allele a at marker $end(h)+1$, the successor of h is the longest fragment $h' \in \mathscr{F}$ such that $start(h') \geq start(h)$, $end(h') = end(h)$ and $h'(end(h)+1) = a$.

Consider a non-homogenous Markov model $\mathscr{R}$, in which each state is a ordered pair of haplotype fragments from $\mathscr{F}$, $\{h_1, h_2\}$, such that $end(h_1) = end(h_2)$. Following the above discussion, each ordered haplotype pair $\{H, H'\}$, H and $H' \in \mathscr{H}(\mathscr{G})$, corresponds to a unique Markov chain of n states, $(s_1, s_2, \cdots, s_n)$, where $s_i = \{h_i, h'_i\}$, $H = (h_1, h_2, \cdots, h_n)$ and $H' = (h'_1, h'_2, \cdots, h'_n)$. Define the transition probability in Markov model $\mathscr{R}$ as follows:

$$P(s_i|s_{i-1}) = P(h_i|h_{i-1})P(h'_i|h'_{i-1}).$$

then the probability of haplotype configurations $\{H, H'\}$ can be calculated from Markov model $\mathscr{R}$:

$$P(\{H, H'\}) = P(s_1) \prod_{i=2,\cdots,n} P(s_i|s_{i-1}).$$

The implementation of EM method in this paper is based on the reformed Markov model $\mathscr{R}$.

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 Biology and Evolution*, 7(2):111–122, 1990.

[3] 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.

[4] Lauri Eronen, Floris Geerts, and Hannu Toivonen. A markov chain approach to reconstruction of long haplotypes. In *Proceedings of 9th Pacific Symposium on Biocomputing (PSB'04)*, pages 104–115. World Scientific, January 2004.

[5] Lauri Eronen, Floris Geerts, and Hannu Toivonen. Haplorec: Efficient and accurate large-scale reconstruction of haplotypes. *BMC Bioinformatics*, 7(542), 2006.

[6] E. Eskin, E. Halperin, and R. M. Karp. Large scale reconstruction of haplotypes from genotype data. In *Proceedings of 7th Annual International Conference on Research in Computational Molecular Biology (RECOMB)*, pages 104–113. ACM Press, 2003.

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

[8] Dan Gusfield. Haplotyping as perfect phylogeny: Conceptual framework and efficient solutions. In *Proceedings of 6th Annual International Conference on Research in Computational Molecular Biology (RECOMB)*, pages 166–175. ACM Press, 2002.

[9] Bjarni V. Halldórsson, Vineet Bafna, Nathan Edwards, Ross Lippert, Shibu Yooseph, and Sorin Istrail. A survey of computational methods for determining haplotypes. In Sorin Istrail et al., editor, *SNPs and Haplotype Inference*, number 2983 in Lecture Notes in Bioinformatics, pages 26–47. Springer-Verlag, Berlin Heidelberg, 2004.

[10] Jonathan Marchini, David Cutler, Nick Patterson, Matthew Stephens, Eleazar Eskin, Eran Halperin, Shin Lin, Zhaohui S. Qin, Heather M. Munro, Gonçalo R. Abecasis, and Peter Donnelly. A comparison of phasing algorithms for trios and unrelated individuals. *American Journal of Human Genetic*, 78:437–450, 2006.

[11] Tianhua Niu, Zhaohui S. Qin, Xiping Xu, and Jun S. Liu. Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. *American Journal of Human Genetics*, 70:157–169, 2002.

[12] Zhaohui S. Qin, Tianhua Niu, and Jun S. Liu. Partition-ligation-expectation-maximization algorithm for haplotype inference with single-nucleotide polymorphisms. *American Journal of Human Genetics*, 71:1242–1247, 2002.

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

[14] J. C. Stephens, J. A. Schneider, D. A. Tanduay, and et al. Haplotype variation and linkage disequilibrium in 313 human genes. *Science*, 293:489–493, 2001.

[15] Matthew Stephens and P. Scheet. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *American Journal of Human Genetics*, 76:449–462, 2005.

[16] Matthew Stephens, Nicholas J. Smith, and Peter Donnelly. A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics*, 68:978–989, 2001.

[17] Hai Yan, Nickolas Papadopoulos, Giancarlo Marra, Claudia Perrera, Josef Jiricny, C. Richard Boland, Henry T. Lynch, Robert B. Chadwick, Albert de la Chapelle, Karin Berg, James R. Eshleman, Weishi Yuan, Sanford Markowitz, Steven J. Laken, Christoph Lengauer, Kenneth W. Kinzler, and Bert Vogelstein. Conversion of diploidy to haploidy. *Nature*, 403(6771):723–724, 2000.

[18] Ji-Hong Zhang, Ling-Yun Wu, Jian Chen, and Xiang-Sun Zhang. A new statistical method for haplotype inference from genotype data. In *Proceedings of IASTED International Conference on Computational and Systems Biology (CASB 2006)*, pages 7–12. ACTA Press, 2006.

[19] Kun Zhang, Jun Zhu, Jay Shendure, Gregory J. Porreca, John D. Aach, Robi D. Mitra, and George M. Church. Long-range polony haplotyping of individual human chromosome molecules. *Nature Genetics*, 38(3):382–387, 2006.

A Algorithm

A.1 Estimation step

The posterior probability of haplotype configuration $\{H_1, H_2\}$ for a given genotype G, in iteration k, can be calculated as follows:

$$P_t\big(\{H_1,H_2\}|G\big) = P\big(\{H_1,H_2\}|G,\lambda_{k-1}\big)$$
$$= \begin{cases} \dfrac{P(H_1,H_2|\lambda_{k-1})}{\sum_{\{H,H'\}\in\mathscr{C}(G)} P(H,H'|\lambda_{k-1})} & \text{if } \{H_1,H_2\}\in\mathscr{C}(G), \\ 0 & \text{otherwise.} \end{cases}$$

Note that the reformed Markov model $\mathscr{R}$ is one order Markov model. Given the model parameter λ, the problem to find the best haplotype configuration that maximizes the posterior probability for a given genotype G can be solved optimally by the DP method in [18], which is an adaptation of well-known Viterbi algorithm [13]. Besides that, with minor modifications, the DP method in [18] can also compute the sum of probabilities of all possible configurations for the given genotype G, which is needed in estimation of the posterior probability of configurations.

A.2 Maximization step

Based on the estimated probabilities of latent variables (haplotype configurations) in previous estimation step, the model parameters λ are re-estimated to improve the total likelihood (see Eq. 2). As discussed above, the model parameters λ are directly derived from haplotype fragment frequencies. Therefore, it is sufficient to re-estimate haplotype fragment frequencies in maximization step. The expected frequency of a haplotype fragment h in a single genotype is the sum over all the haplotypes that match h in its possible configurations. The overall expected frequency of the fragment h is an average over all genotypes.

$$F_t(h) = \frac{1}{|\mathscr{G}|}\sum_{G\in\mathscr{G}}\left(\sum_{\{H_1,H_2\}\in\mathscr{C}(G)} \frac{1}{2}P_t(\{H_1,H_2\}|G)\delta_{h,\{H_1,H_2\}}\right),$$

where $\delta_{h,\{H_1,H_2\}} \in \{0,1,2\}$ is the number of haplotypes in $\{H_1,H_2\}$ that match h.

In order to describe the procedure for re-estimation of model parameters, we first define the forward variables and backward variables. Denote the states in $\mathscr{R}$ as $\mathscr{S} = \{S_1, S_2, \cdots, S_N\}$, and the state at time (marker) t as q_t. The forward variable $\alpha_t(i)$ is the probability of the partial observation genotype $G(1,t)$ and the state S_i at time t, given the model parameters λ, i.e.:

$$\alpha_t(i) = P(G(1,t), q_t = S_i|\lambda).$$

The forward variables can be calculated inductively as follows:

$$\alpha_1(i) = \begin{cases} P(S_i|\lambda), & \text{if } end(S_i) = 1 \text{ and } S_i \text{ matches } G, \\ 0, & \text{otherwise,} \end{cases}$$

$$\alpha_t(i) = \begin{cases} \sum_{j=1}^{N} \alpha_{t-1}(j)P(S_i|S_j,\lambda), & \text{if } end(S_i) = t \text{ and } S_i \text{ matches } G, \\ 0, & \text{otherwise,} \end{cases}$$

for $t = 2, 3, \cdots, n$. Similarly, we can define the backward variables $\beta_t(i)$ as the probability of the partial observation genotype $G(t+1,n)$, given the state S_i at time t and the model parameters λ, i.e.:

$$\beta_t(i) = P(G(t+1,n)|q_t = S_i, \lambda).$$

The backward variables can be calculated inductively as follows:

$$\beta_n(i) = \begin{cases} 1, & \text{if } end(S_i) = n \text{ and } S_i \text{ matches } G, \\ 0, & \text{otherwise,} \end{cases}$$

$$\beta_t(i) = \begin{cases} \sum_{j=1}^{N} \beta_{t+1}(j) P(S_j|S_i,\lambda), & \text{if } end(S_i) = t \text{ and } S_i \text{ matches } G, \\ 0, & \text{otherwise,} \end{cases}$$

for $t = 1,2,\cdots,n-1$.

Denote S_i^1, S_i^2 as the first and second fragment of S_i respectively, and denote G^1 and G^2 as the first and second haplotype of genotype G. Define $\xi_t^0(i,h)$, $\xi_t^1(i,h)$, $\xi_t^2(i,h)$ as the probabilities of the partial observation genotype $G(1,t)$, the state S_i at time t and haplotype fragment $h(start(h),t)$ at both or one of two haplotypes in the configuration of G, given the model parameter λ, i.e.:

$$\begin{aligned} \xi_t^0(i,h) &= P\big(G(1,t), q_t = S_i, h(s,t) = G^1(s,t) = G^2(s,t), s = start(h)|\lambda\big), \\ \xi_t^1(i,h) &= P\big(G(1,t), q_t = S_i, h(s,t) = G^1(s,t) \neq G^2(s,t), s = start(h)|\lambda\big), \\ \xi_t^2(i,h) &= P\big(G(1,t), q_t = S_i, h(s,t) = G^2(s,t) \neq G^1(s,t), s = start(h)|\lambda\big). \end{aligned}$$

For $i = 1,2,\cdots,N$, $1 \leq t < start(h)$,

$$\xi_t^0(i,h) = \alpha_t(i), \qquad \xi_t^1(i,h) = \xi_t^2(i,h) = 0.$$

For $start(h) \leq t \leq end(h)$,

$$\xi_t^0(i,h) = \begin{cases} \sum_{j=1}^{N} \xi_{t-1}^0(j) P(S_i|S_j,\lambda), & \text{if } t > 1 \text{ and } h(t) = S_i^1(t) = S_i^2(t), \\ \alpha_t(i), & \text{if } t = 1 \text{ and } h(t) = S_i^1(t) = S_i^2(t), \\ 0, & \text{otherwise.} \end{cases}$$

$$\xi_t^1(i,h) = \begin{cases} \sum_{j=1}^{N} (\frac{1}{2}\xi_{t-1}^0(j) + \xi_{t-1}^1(j)) P(S_i|S_j,\lambda), & \text{if } t > 1 \text{ and } h(t) = S_i^1(t) \neq S_i^2(t), \\ \frac{1}{2}\alpha_t(i), & \text{if } t = 1 \text{ and } h(t) = S_i^1(t) \neq S_i^2(t), \\ 0, & \text{otherwise.} \end{cases}$$

$$\xi_t^2(i,h) = \begin{cases} \sum_{j=1}^{N} (\frac{1}{2}\xi_{t-1}^0(j) + \xi_{t-1}^2(j)) P(S_i|S_j,\lambda), & \text{if } t > 1 \text{ and } h(t) = S_i^2(t) \neq S_i^1(t), \\ \frac{1}{2}\alpha_t(i), & \text{if } t = 1 \text{ and } h(t) = S_i^2(t) \neq S_i^1(t), \\ 0, & \text{otherwise.} \end{cases}$$

Then the frequency of haplotype fragment h can be re-estimated as follows:

$$F_t(h) = \frac{1}{|\mathscr{G}|} \sum_{G \in \mathscr{G}} \left(\sum_{i=1}^{N} \left(\sum_{j=1}^{3} \xi_t^j(i,h) \right) \beta_t(i) \right).$$

where $t = end(h)$.

A.3 Initialization

The initial setting of model parameters λ are computed under the simple assumption that all possible haplotype configurations for a genotype have equally probabilities, since there is no prior knowledge about the distribution of different configurations. The calculation method is same as in [18].

For each haplotype fragment h where $start(h) = s$, $end(h) = e$,

$$F_0(h) = \frac{1}{|\mathscr{G}|} \sum_{\substack{G \in \mathscr{G} \\ h \in G(s,e)}} \prod_{s \leq i \leq e} fr(h,i,G)$$

where $fr(h,i,G)$ is the frequency of allele $h(i)$ at marker i given genotype G, and defined as follows:

$$fr(h,i,G) = \begin{cases} 1, & h(i) \in G(i) \text{ and G(i) is homozygous,} \\ 0.5, & h(i) \in G(i) \text{ and G(i) is heterozygous,} \\ 0.5(1+fr(h,i)), & h(i) \in G(i) \text{ and G(i) is partial missing,} \\ 0.5fr(h,i), & h(i) \notin G(i) \text{ and G(i) is partial missing,} \\ fr(h,i), & \text{G(i) is missing,} \\ 0, & \text{otherwise,} \end{cases}$$

and $fr(h,i)$ is the frequency of allele $h(i)$ at marker i in whole data $\mathscr{G}$.

A.4 Complexity

Let $|\mathscr{F}| = c$, $|\mathscr{G}| = m$, $|\mathscr{M}| = n$. Obviously $|\mathscr{S}| = O(c^2)$. The complexity of estimation step is $O(mc^2)$ [18]. Note that for each i, there are only a constant number of j such that $P(S_i|S_j,\lambda) > 0$, and $\alpha_t(i) = 0$, $\beta_t(i) = 0$ for any $t \neq end(S_i)$. The complexity of calculation of $\alpha_t(i)$, $\beta_t(i)$ for a genotype G, is $O(c^2)$. Similarly, noting that $\xi_t^j(i,h) = 0$ for any $t \neq end(S_i)$, the complexity of calculation of $\xi_t^j(i)$ for a fragment h is also $O(c^2)$. Therefore, the re-estimation of each fragment will take $O(mc^2)$ and the complexity for re-estimation of all fragments in $\mathscr{F}$, i.e. the complexity of maximization step, is $O(mc^3)$.

Note that $\xi_t^j(i,h) = \xi_t^j(i,h')$ if $s = start(h) = start(h')$, $e = end(h) \leq end(h')$ and $h(s,e) = h'(s,e)$, i.e. h is a prefix of h', where $t <= e$ and $j = 0,1,2$. Therefore the calculation of maximization step can be accelerated by organizing the haplotype fragments in a prefix tree and using a depth-first-search method. The computational experiments show that the improved implementation can be one order of magnitude faster than the simple one. Similarly, the calculation of initial model parameters can also be implemented by same manner to improve the speed.