

A Markov Chain Model for Haplotype Assembly from SNP Fragments

Rui-Sheng Wang^{1,2}
wangrsh@amss.ac.cn

Ling-Yun Wu³
wlyun@amt.ac.cn

Xiang-Sun Zhang^{3*}
zxs@amt.ac.cn

Luonan Chen^{1,4,5*}
chen@eic.osaka-sandai.ac.jp

¹ Faculty of Engineering, Osaka Sangyo University, Osaka 574-8530, Japan

² School of Information, Renmin University of China, Beijing 100872, China

³ AMSS, Chinese Academy of Sciences, Beijing 100080, China

⁴ Institute of Systems Biology, Shanghai University, Shanghai 200444, China

⁵ ERATO Aihara Complexity Modelling Project, JST, and Institute of Industrial Science, University of Tokyo, Tokyo 153-8505, Japan

* Corresponding author

Abstract

Single nucleotide polymorphism (SNP) is the most frequent form of human genetic variations and of importance for medical diagnosis and tracking disease genes. A haplotype is a sequence of SNPs from a single copy of a chromosome, and haplotype assembly from SNP fragments is based on DNA fragments with SNPs and the methodology of shotgun sequence assembly. In contrast to conventional combinatorial models which aim at different error types in SNP fragments, in this paper we propose a new statistical model — a Markov chain model for haplotype assembly based on information of SNP fragments. The main advantage of this model over combinatorial ones is that it requires no prior information on error types in data. In addition, unlike exact algorithms with the exponential-time computation complexity for most combinatorial models, the proposed model can be solved in polynomial time and thus is efficient for large-scale problems. Experiment results on several data sets illustrate the effectiveness of the new method.

Keywords: Markov chain, haplotype assembly, SNP fragments, computational complexity

1 Introduction

Genome sequencing efforts [15] show that all human share about 99% identity at DNA level and it is small regions of genetic differences that account for the phenotype variations and disease susceptibilities [8] in a population. The most frequent form to address genetic differences is single nucleotide polymorphism (SNP) — a single DNA base with more than one nucleotide occurs in the population [2]. SNPs can be used as genetic makers and have important roles in medical diagnosis and disease association.

The nucleotides {A, G, C, T} in a SNP position are called *alleles*. Almost all SNPs have two instead of four different alleles which we denote the wild type as 0 and the mutant type as 1. In diploid organisms, the genomes are organized as pairs of chromosomes, a maternal copy and a paternal copy. A *haplotype* is the allele sequence information on each copy of a pair of chromosomes which is a string over {0, 1} or {A, G, C, T}. Most current techniques for SNPs discovery do not provide the haplotype information for each copy. Instead, these methods just give the *genotype* information, i.e. the unordered allele pair sequence on two copies of a pair of chromosomes. For a genotype, if a pair of alleles at a SNP site is made of two identical values, this SNP site is called *homozygous*, otherwise it is called *heterozygous*.

Haplotypes contain much more information than individual genetic markers or SNPs and have critical importance in studying disease association and genome evolution [14]. Haplotypes can be obtained by separating DNA sequences and then sequencing through constructing somatic cell hybrids [6, 17]. Such experimental method is very time-consuming, labor-intensive as well as expensive. Recently, much attention has focused on computational methods for obtaining haplotypes from genotype data [4, 7, 9, 11, 13] or SNP fragments [1, 3, 10, 12, 16, 18]. Haplotype assembly, also referred as individual haplotyping, is based on DNP fragments with SNPs and methodology of shotgun sequence assembly [10, 12]. The input data can be the aligned short DNA fragments with SNPs obtained by DNA shotgun sequencing or a resequencing effort for the purpose of large scale haplotyping. When we focus on SNP positions, these short DNA fragments are actually aligned SNP fragments. DNA sequencing errors and the diploidy of human genome make the problem complex.

Emphasizing different types of errors, Lancia *et al.* [10] have given two models for the haplotype assembly problem — the Minimum Fragment Removal (MFR) model and the Minimum SNP Removal (MSR) model. MFR aims at errors caused by fragment contamination and assumes that “bad” fragments are due to contamination, while MSR aims at the sequencing errors and tries to remove sequencing errors by removing some whole SNP sites. Another important computational model for haplotype assembly — the Minimum Error Correction (MEC) model was proposed and proved to be NP-hard in [12]. This model is also suitable for the case that all the fragments come from one organism but there are sequencing errors to be corrected. A common feature of these combinatorial models for haplotype assembly is that they are all NP-hard and even APX-hard [1, 3].

Although much work has been carried on haplotype assembly, there are still some problems unsolved whose solution will be significant. Firstly, as mentioned above, existing combinatorial models for haplotype assembly all aim at some type of data errors. However, when facing a data set without any prior information, we do not know the type of data errors and thus do not know which model we should choose. Hence, a unified and consistent model that can deal with various types of data errors is needed. In addition, existing combinatorial models for haplotype assembly are all NP-hard in terms of computational complexity, i.e. there is no polynomial time exact algorithm for these problems, therefore it is impossible to deal with large-scale problems, which are common in practice [16]. In this paper, we reformulate the haplotype assembly problem into a statistical framework and particularly model it by a Markov chain (MC) model which can accommodate various data errors. Furthermore, it can be solved exactly in polynomial time, thereby suitable for large-scale problems. In addition, haplotypes have very strong linkage disequilibrium property [5, 6] and MC model is an effective tool for characterizing this kind of linkage property. For example, a MC model [7] and a hidden Markov model [13] have been proposed as a tool of haplotype inference from genotypes. To demonstrate the proposed method, multiple numerical experiments including comparison with combinatorial models are conducted to confirm the effectiveness of this model.

2 Problem

We consider n consecutive SNP sites, where A_i is the set of possible alleles which the i th SNP site takes value on. A haplotype H can be viewed as a vector on alleles i.e. $H \in \prod_{i=1}^n A_i$. A genotype G is a vector consisting of unordered allele pairs, i.e. $G \in \prod_{i=1}^n (A_i \times A_i)$. The allele that haplotype H takes at the i th SNP site is denoted by $H(i)$. Similarly, the allele pair that genotype G takes at the i th SNP site is denoted by $G(i)$. $H(i, j)$ ($G(i, j)$) represents a fragment of haplotype H (genotype G) from the i th SNP site to the j th SNP site. X is a SNP matrix on $\{A, G, C, T, -\}$ or $\{0, 1, -\}$, where each row represents a SNP fragment. Generally, X can be obtained by using some DNA fragment alignment algorithms. The ‘-’ in X is called *hole* which is missing reads in DNA sequencing or uncovered site by some fragment since the fragments may have different lengths.

Given a pair of haplotypes (H_1, H_2) and a genotype G , if $G(i) = \{H_1(i), H_2(i)\}$ holds for all i , $i = 1, \dots, n$, then (H_1, H_2) and G are said to be compatible, denoted by $H_1 \oplus H_2 = G$, or $H_1 = G \ominus H_2$,

$H_2 = G \ominus H_1$. For a haplotype H , if $H(i) \in G(i)$ holds for all $i, i = 1, \dots, n$, then H and G are said to be compatible, denoted by $H \in G$, $\bar{H} = G \ominus H$. The compatibility of a haplotype fragment $H(i, j)$ and a genotype $G(i, j)$ can be defined similarly. Obviously, a pair of haplotypes uniquely determines a genotype, but a genotype can have many compatible haplotype pairs. Specifically, a genotype G with k heterozygous SNP sites has 2^{k-1} pairs of haplotypes that are compatible with G . These haplotype pairs are denoted as $C(G)$, $|C(G)| = 2^{k-1}$.

From a statistical view, haplotype assembly is such a problem: given a set of SNP fragments (a SNP matrix) X , find a most possible haplotype pair (i.e. with largest probability: $\Pr\{H_1^*, H_2^*\}$) over all possible haplotype pairs. That is, to find a pair of haplotypes $\{H_1^*, H_2^*\}$ such that

$$\Pr(\{H_1^*, H_2^*\} | X) = \max_{\{H_1, H_2\}} \Pr(\{H_1, H_2\} | X). \quad (1)$$

3 Method

Since genotype data can be easily obtained from biological experiments, we assume the genotype of a pair of haplotypes to be reconstructed is known. In addition, according to Hardy-Weinberg equilibrium, these two haplotypes are independent. From the Bayes rule, given a SNP matrix X and a genotype G , the probability estimation of a pair of haplotypes can be transformed into the probability estimation of a single haplotype:

$$\Pr(\{H_1, H_2\} | X, G) = \begin{cases} \frac{\Pr(H_1 | X) \Pr(H_2 | X)}{\sum_{\{H, \bar{H}\} \in C(G)} \Pr(H | X) \Pr(\bar{H} | X)} & \text{if } \{H_1, H_2\} \in C(G), \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

Thus, the problem (1) is transformed as finding a pair of haplotypes $\{H_1^*, H_2^*\}$ such that

$$\Pr(H_1^* | X) \Pr(H_2^* | X) = \max_{\{H_1, H_2\} \in C(G)} \Pr(H_1 | X) \Pr(H_2 | X). \quad (3)$$

There is a kind of statistical dependence (linkage disequilibrium) in the neighboring SNPs [5, 6]. SNPs that belong to one linkage disequilibrium block are not easily separated by recombinant event. Non-neighboring SNPs are relatively independent. Therefore, the probability of a haplotype can be represented by a set of conditional probabilities and the probabilities of alleles on each SNP site depend on that of preceding alleles. MC model can describe such dependence between SNPs:

$$\Pr(H | X) = \Pr(H(1) | X) \prod_{i=2}^n \Pr(H(i) | H(i-1), X). \quad (4)$$

This model indicates that owing to the correlation property between SNP sites, if we have some information for alleles on one SNP site we can also derive the related information for next alleles. Sometimes, not only neighboring SNPs have strong linkage disequilibrium property, but also even several contiguous SNP sites have strong association. For such a case, the following d -order MC model is required:

$$\Pr(H | X) = \Pr(H(1, d) | X) \prod_{i=d+1}^n \Pr(H(i) | H(i-d, i-1), X). \quad (5)$$

where d denotes that the probability of the alleles on current SNP site is dependent on that of d preceding SNP sites. When $d = 1$, the above model is standard MC model. The probability estimation in (4) and (5) is described in details in Appendix. The optimal solution to the MC model can be obtained by a dynamic programming algorithm (see Appendix).

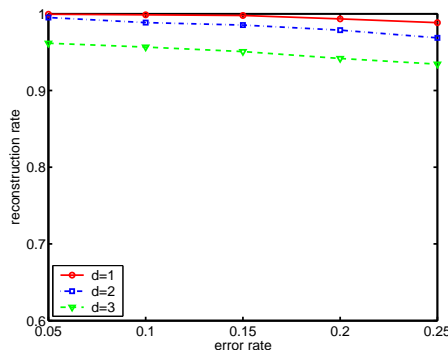


Figure 1: The reconstruction rate of the MC model on Daly set. d is the order of MC model.

4 Results

In this section, we will use both real data and simulation data to test the MC model for haplotype reconstruction. The algorithms are implemented on a 2.26G Hz Pentium 4 processor PC using Microsoft Visual C++ compiler 6. In our experiments, we use *reconstruction rate* [16], i.e. the similarity degree (the number of matching bases) between the original haplotypes and the assembled haplotypes, to measure performance of the model. A reconstruction rate 1 means that two original haplotypes are completely reconstructed.

4.1 Experiment on Daly Data Set and Simulated Data Set

We first use Daly's 129 genotypes from chromosome 5q31 as one test set [5]. This commonly used benchmark dataset is a sample from a European-derived population consists of 129 trios and spans a 500-kb region on human chromosome 5q31 with 103 SNPs. The haplotypes of 129 children from the trios can be inferred from the genotypes of their parents through pedigree information and the non-transmitted chromosomes as an extra 129 (pseudo) genotypes. Markers for which both alleles could not be inferred are marked as missing. From the resulting set of 258 haplotype pairs (genotypes), the ones with more than 20% missing alleles are removed, leaving 147 haplotype pairs. Among these pairs, 18 genotypes with no more than one heterozygous site are omitted, leaving 129 pairs of haplotypes as the test set. 5 instances are generated from each haplotype pair respectively according to the error rate of SNP fragments $e = 0.05, 0.1, 0.15, 0.2, 0.25$. Specifically, the SNP fragments in each instance are randomly copied from two haplotypes with certain error rate. The length of SNP fragments is between 5 and 8 and the starting sites of these fragments are uniformly distributed. The total number of fragments is relevant to sequence coverage. Coverage c is the number of SNP fragments that cover a SNP site. Generally, c is between 5 and 10 in shotgun sequencing. Here we set $c = 7$. The number of SNP fragments in each instance is about $m = 2nc/L$, where L is the average length of fragments and n is the length of the haplotypes. Here the length of fragments and the coverage of SNP sites are set as those in practice and the number of SNP fragments is determined by these parameters. Certainly, the larger the coverage, the more accurate the probability estimation.

Since the computational complexity for dynamic programming (DP) that computes the frequencies of haplotype pairs is on the order of polynomial-time, assembling haplotype pairs by MC model is very fast. For instances with about 100 SNP site (for this case, the total number of SNP fragments is about 230), the implement time is no more than 1 microsecond. We solve the instances respectively by MC with order $d = 1, d = 2, d = 3$. The average reconstruction rate of 129 haplotype pairs by MC with various orders is illustrated in Figure 1. From Figure 1, clearly when $d = 1$, the reconstruction rate is very high. The larger the order is, the lower the reconstruction rate. This is mainly because SNP fragments are so short that MC can not mine long linkage disequilibrium property only from the given

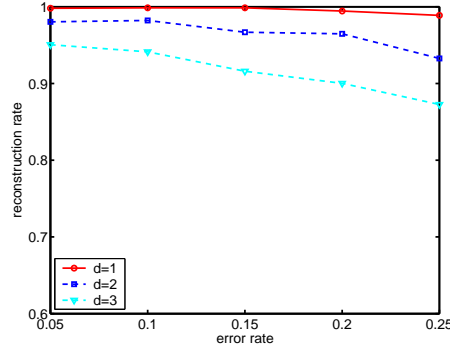


Figure 2: The reconstruction rate of the MC model on simulated data. d is the order of MC model.

information. Large order can only make the frequencies of Markov states very low and the model can not distinguish which are true haplotype fragments. Of course, there may be some individual instances on which MC with a high order has higher reconstruction rate than that with a low order.

Now we use the simulated data to test the reconstruction rate of MC. Firstly, a haplotype with 100 SNP sites are generated randomly, where A, C, G, T appear equiprobably. Secondly, according to a parameter *similarity rate* s (denotes the similarity degree of two haplotypes in a pair), another haplotype is randomly generated. Let $s = 0.5$, and totally 100 pairs of haplotypes are generated. Then 5 instances with about $m = 2nc/L$ SNP fragments are copied from each haplotype pair by the method used above. According to the values of e , totally 500 instances are taken as a test set.

The result on simulated data is similar to that on real data set. The implementing DP algorithm is also very fast. Again solving MC with $d = 1$ obtains most accurate results. Note that MC with $d = 3$ on simulated data has worse performance than on real data. This may be because the correlation (linkage disequilibrium) property between SNPs in simulated data is not so strong as that in real data.

4.2 Comparison with Combinatorial Model

Now we compare MC with a combinatorial model — MEC in terms of reconstruction rate and the implementing time of exact algorithms for these two models. The exact algorithm for MC is the DP algorithm given in Appendix and that for MEC model is a branch and bound (BNB) algorithm [16]. In order to compare the implementing time of two algorithms, we generate 100 instances with error rate 0.1 for 16, 18, 20, 22 SNP sites respectively. The number of SNP fragments in these instances is between 37 and 51. The result on the reconstruction rate of two models is summarized in Table 1 and the comparison of the implementing time of two exact algorithms is illustrated in Figure 3, where the y axis is time (second).

Table 1: Comparison of the reconstruction rate of MC model and MEC model.

Models \ SNPs	16	18	20	22
MEC	0.975	1.000	0.995	0.959
MC	0.988	1.000	0.995	1.000

According to the numerical simulation, for a comparable reconstruction rate, MC is very fast, but the implementing CPU time of MEC model exponentially increases with respect to the number of SNP sites. For instances with more than 25 SNP sites (for this case, the number of SNP fragments is about 60), the implementing CPU time is prohibitively long (several hours).

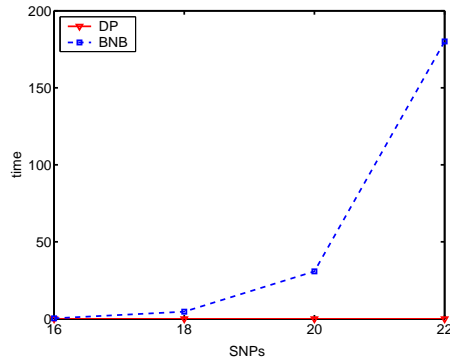


Figure 3: Comparison of the implementing time of DP and BNB.

4.3 Experiment on HapMap Data

In this section, we use the haplotype data downloaded from HapMap web site to test our model. The goal of the international HapMap Project [19] is to compare the genetic sequences of different individuals to identify chromosomal regions where genetic variants are shared. There are genome wide phased genotypes on 23 chromosomes about 4 populations in 2005-03-Phased I. Two of them (HCB and JPT) consist of a sample of unrelated individuals whereas the others (CEU and YRI) are a sample of trios. We choose HCB population including 45 individuals. The haploypes are vectors on $\{0, 1\}$ with 509 SNPs. We use these haplotypes to generate a data set in which the SNP fragments are matrixes on $\{0, 1, -\}$. The result is in Figure 4 which shows the high reconstruction rate of the model. Note that even for these large-scale problems (the number of total fragments in these instances is more than 1000), MC solves them in no more than one second. The implementing time of MEC for instances of such size is prohibitively long, and is shown in Figure 3.

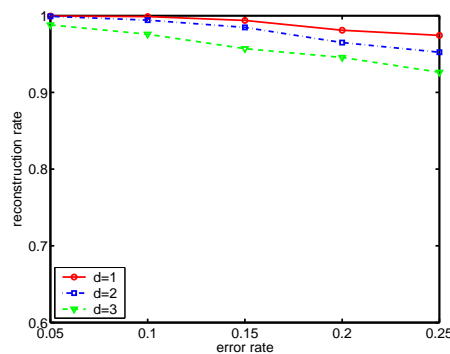


Figure 4: The reconstruction rate of the MC model on HapMap data. d is the order of MC model.

5 Conclusion and Discussion

Haplotype assembly or individual haplotyping is to infer a pair of haplotypes from localized polymorphism data gathered through short genome fragment assembly. It is an important problem in bioinformatics and has been researched by many authors. In this paper, we propose a new statistical model — MC model for haplotype assembly which overcomes several defects of current combinatorial models. Specifically, this model is a consistent one and can accommodate various types of errors in SNP fragments without requirement of the prior information about error type. In addition, the proposed model can be used to solve large-scale problems in a high accuracy because its exact algorithm

is on the order of polynomial-time. Experimental results on several data sets show the effectiveness of the new model.

Although the genotype information is easily obtained from biological experiments, employing the genotypes is still a defect of the MC model. How to use available information to assemble more accurate haplotypes is our goal. As a future topic, we will adopt the Hidden Markov Model (HMM) to model haplotype assembly in which errors in SNP fragments can be naturally viewed as observation errors. Such technique does not need genotype information and may be superior to the MC model.

Acknowledgment

This work is supported by National Natural Science Foundation of China under Grant No. 10471141 and No. 60503004, and Important Research Direction Project of CAS “Some Important Problems in Bioinformatics”.

References

- [1] Bafna, V., Istrail, S., Lancia, G., and Rizzi, R., Polynomial and APX-hard cases of the individual haplotyping problem, *Theor. Comp. Sci.*, 335:109–125, 2005.
- [2] Chakravarti, A., It's raining SMPs, hallelujah? *Nat. Genet.*, 19:216–217, 1998.
- [3] Cilibrasi, R., Iersel, L., Kelk, S., and Tromp, J., On the complexity of the single individual haplotyping problem, *Proc. WABI2005*, 2005.
- [4] Clark, A.G., Inference of haplotypes from PCR-amplified samples of diploid populations, *Mol. Biol. and Evol.*, 7(2):111–122, 1990.
- [5] Daly, M., Rioux, J., Hudson, T., and Lander, E., High-resolution haplotype structure in human genome, *Nat. Genet.*, 29:229–232, 2001.
- [6] Douglas, J.A., Boehnke, M., Gillanders, E., Trent, J.M., and Gruber, S.B., Experimentally-derived haplotypes substantially increase the efficiency of linkage disequilibrium studies, *Nat. Genet.*, 28:361–364, 2001.
- [7] Eronen, L., Geerts, F., and Toivonen, H., A Markov chain approach to reconstruction of long haplotypes, *Proc. 9th Pac. Symp. Biocomput. (PSB'04)*, 9:104–115, 2004.
- [8] Hoehe, M., Kopke, K., Wendel, B., Rohde, K., Flachmeier, C., Kidd, K., Berrettini, W., and Church, G., Sequence variability and candidate gene analysis in complex disease: association of μ opioid receptor gene variation with substance dependence. *Hum. Mol. Genet.*, 9:2895–2908, 2000.
- [9] Gusfield, D., Haplotyping as perfect phylogeny: conceptual framework and efficient solutions, *Proc. Int. Conf. Res. Comput. Mol. Biol. (RECOMB'02)*, 166–175, 2002.
- [10] Lancia, G., Bafna, V., Istrail, S., Lippert, R., and Schwartz, R., SNPs problems, complexity, and algorithms, *Proc. 9th Annu. Eur. Symp. Algorithms (ESA)*, 182–193, 2001.
- [11] Li, Z., Zhou, W., Zhang, X., Chen, L., A parsimonious tree-grow method for haplotype inference, *Bioinformatics*, 21:3475–3481, 2005.
- [12] Lippert, R., Schwartz, R., Lancia, G., and Istrail, S., Algorithmic strategies for the single nucleotide polymorphism haplotype assembly problem, *Brief. Bioinform.*, 3:23–31, 2002.

- [13] Rastas, P., Koivisto, M., Mannila, H., and Ukkonen, E., A hidden markov technique for haplotype reconstruction, *Lecture Notes in Bioinformatics*, 3692:140–151, 2005.
- [14] Stephens J.C., Schneider J.A., Tanguay D.A., Choi J., Acharya T., Stanley S.E., Jiang R., Messer C.J., Chew A., Han J.H., Duan J., Carr J.L., Lee M.S., Koshy B., Kumar A.M., Zhang G., Newell W.R., Windemuth A., Xu C., Kalbfleisch T.S., Shaner S.L., Arnold K., Schulz V., Drysdale C.M., Nandabalan K., Judson R.S., Ruano G., and Vovis G.F., Haplotype variation and linkage disequilibrium in 313 human genes, *Science*, 293:489–493, 2001.
- [15] Venter, J.C., Adams, M.D., *et al.*, The sequence of the human genome, *Science*, 291(5507):1304–1351, 2001.
- [16] Wang, R.S., Wu, L.Y., Li, Z.P., and Zhang, X.S., Haplotype reconstruction from SNP fragments by minimum error correction, *Bioinformatics*, 21(10):2456–2462, 2005.
- [17] Yan, H., Papadopoulos, N., Marra, G., *et al.*, Conversion of diploidy to haploidy, *Nature*, 403:723–724, 2000.
- [18] Zhang, X.S., Wang, R.S., Wu, L.Y., and Chen, L., Models and algorithms for the haplotyping problem, *Curr. Bioinform.*, 1(1):105–114, 2006.
- [19] <http://www.hapmap.org/>

Appendix

A.1 Markov Chain Model

Since G is known, given the order of Markov chain model d , let

$$S_i^G(d) = \{H(i-d+1, i) : H(i-d+1, i) \in G\}, \quad i = d, d+1, \dots, n.$$

$S_i^G(d)$ is said to be the state set of H at the i th site. Let $h \in S_i^G(d)$, $h' \in S_{i+1}^G(d)$. If the suffixal $d-1$ alleles of h is identical to the prefixal $d-1$ alleles of h' , h' is said to be a successor of h , denoted by $h \rightarrow h'$. Obviously, given genotype G and $h_i \in S_i^G(d)$, $i = d, d+1, \dots, n$, if $h_d \rightarrow h_{d+1} \rightarrow \dots \rightarrow h_n$, then $H = h_d \rightarrow h_{d+1} \rightarrow \dots \rightarrow h_n$ is a haplotype compatible with G .

Let $S_t(d) = S_{t+d}^G(d)$, then $\{S_t(d), p^t, t = 0, 1, \dots, n-d\}$ is a MC that describes the haplotype assembly problem, where $p^0(\cdot)$ is the distribution of Markov chain's initial states on $S_0(d)$, and $p^t(\cdot | \cdot)$, $t = 1, 2, \dots, n-d$ is the state transition probability from time $t-1$ to time t :

$$p^t(h_t | h_{t-1}) = \begin{cases} \Pr(h_t | h_{t-1}, X) & \text{if } h_{t-1} \rightarrow h_t, \\ 0 & \text{otherwise,} \end{cases} \quad (6)$$

where $h_{t-1} \in S_{t-1}(d)$, $h_t \in S_t(d)$. The estimation of $P(h_t | h_{t-1}, X)$ will be discussed in the following subsection. For haplotype $H = h_1 \rightarrow h_2 \rightarrow \dots \rightarrow h_{n-d}$, $h_i \in S_i(d)$, $i = 1, 2, \dots, n-d$, its probability can be computed by the Markov chain model:

$$\begin{aligned} \Pr(H | X) &= \Pr(H(1, d) | X) \prod_{t=1}^{n-d} \Pr(H(t+d) | H(t, t+d-1), X) \\ &= p^0(h_0) \prod_{t=1}^{n-d} p^t(h_t | h_{t-1}). \end{aligned}$$

The complement haplotype \bar{H} can be easily obtained from H and G .

A.2 Parameter Estimation

We estimate the state transition probability according to the frequency of haplotype fragments in SNP matrix. Note that the MC model for haplotype assembly is inhomogeneous. In other words, each SNP site has its own state and transition probability which are different from the MC model for sequence alignment. For $h_{t-1} = H(t, t+d-1) \in S_{t-1}(d)$, $h_t = H(t+1, t+d) \in S_t(d)$, if h_t is a successor of h_{t-1} , let $H(t, t+d)$ (i.e. $h_{t-1} \rightarrow h_t$) denote the fragment that h_{t-1} overlaps with h_t in $d-1$ consecutive alleles. Then given the state at time $t-1$ as h_{t-1} , the probability of the state h_t at time t can be estimated as

$$\begin{aligned} \Pr(h_t | h_{t-1}, X) &= \frac{\Pr(h_{t-1}, h_t | X)}{\Pr(h_{t-1} | X)} = \frac{\Pr(H(t, t+d) | X)}{\Pr(H(t, t+d-1) | X)} \\ &\approx \frac{\text{Fr}(H(t, t+d))}{\text{Fr}(H(t, t+d-1))}, \end{aligned} \quad (7)$$

The probability of initial state is estimated as follows

$$p^0(h_0) = \Pr(H(1, d) | X) \approx \text{Fr}(H(1, d)), \quad (8)$$

where $\text{Fr}(\cdot)$ denotes the frequency of a haplotype fragment appearing in SNP matrix.

Specifically, we compute the frequency of a haplotype fragment in such a way: $H(i, j)$ is a haplotype fragment, and X_k is the k th row of SNP matrix X (i.e. the k th SNP fragment), $k = 1, \dots, m$. $X_k(i, j)$ denotes the part of X_k from the i th SNP site to the j th SNP site. If $X_k(i, j)$ has no site that takes value ‘-’ and is identical to $H(i, j)$, then the frequency of $H(i, j)$ is added 1. If $X_k(i, j)$ (not including those $X_k(i, j)$ that each site takes value ‘-’) is identical to $H(i, j)$ except those sites taking value ‘-’ (we assume there are N such sites), then the frequency of $H(i, j)$ is added by $\frac{1}{4N}$ or $\frac{1}{2N}$ depending on the SNP matrix over $\{A, G, C, T, -\}$ or $\{0, 1, -\}$. To avoid that the frequencies of haplotype fragments in a state set are all 0s, let the initial frequencies of the haplotype fragments in state sets be ε , where ε is a very small positive number.

Since we assume that G is known, the transition probability formula (7) should be modified. Let $h_{t-1} = H(t, t+d-1) \in S_{t-1}(d)$, $h_t = H(t+1, t+d) \in S_t(d)$, and h_t is a successor of h_{t-1} . If the $(t+d)$ -th SNP site of G is a homozygous site, set $\Pr(h_t | h_{t-1}, X) = 1$. If it is a heterozygous, say $G(t+d) = \{A, T\}$, which means that given the state at time $(t-1)$ h_{t-1} , the state at the next SNP site can only be A or T. Let $h_t = H(t+1, t+d-1) \cup A$, $f_1 = H(t, t+d-1) \cup A$, $f_2 = H(t, t+d-1) \cup T$ and set $f_3 = H(t, t+d-1) \cup C$, $f_4 = H(t, t+d-1) \cup G$. Then distribute averagely the frequency sum of f_3 and f_4 to f_1 and f_2 , and normalize it:

$$\Pr(h_t | h_{t-1}, X) = \frac{\text{Fr}(f_1) + \frac{1}{2}(\text{Fr}(f_3) + \text{Fr}(f_4))}{\text{Fr}(f_1) + \text{Fr}(f_2) + \text{Fr}(f_3) + \text{Fr}(f_4)}.$$

For SNP matrix on $\{0, 1, -\}$, the idea for parameter estimation is similar.

A3. Dynamic Programming Algorithm

It is easy to see that for $h_t = H(t+1, t+d) \in S_t(d)$, $\bar{h}_t = \bar{H}(t+1, t+d) = G \ominus H(t+1, t+d)$ must be in $S_t(d)$. We can solve the problem (3) by a polynomial-time DP algorithm.

Step 1 Initialize: for all $h_0 \in S_0(d)$, let

$$P(h_0) = p^0(h_0)p^0(\bar{h}_0) = \text{Fr}(h_0)\text{Fr}(\bar{h}_0).$$

Step 2 Iterate: for all $h_t \in S_t(d)$, $t = 1, 2, \dots, n-d$, compute

$$P(h_t) = \max\{P(h_{t-1})p^t(h_t | h_{t-1})p^t(\bar{h}_t | \bar{h}_{t-1}) : h_{t-1} \in S_{t-1}(d)\},$$

$$\phi_t(h_t) = \arg \max \{P(h_{t-1})p^t(h_t | h_{t-1})p^t(\bar{h}_t | \bar{h}_{t-1}) : h_{t-1} \in S_{t-1}(d)\}.$$

Step 3 Stop:

$$P^* = \max \{P(h_{n-d}) : h_{n-d} \in S_{n-d}(d)\}.$$

Step 4 Trace back:

$$h_{n-d}^* = \arg \max \{P(h_{n-d}) : h_{n-d} \in S_{n-d}(d)\}, \quad h_t^* = \phi_{t+1}(h_{t+1}^*),$$

where $t = n - d - 1, n - d - 2, \dots, 0$.

Step 5 Assemble haplotypes

$$H^* = h_0^* \rightarrow h_1^* \rightarrow \dots \rightarrow h_{n-d}^*, \quad \bar{H}^* = \bar{h}_0^* \rightarrow \bar{h}_1^* \rightarrow \dots \rightarrow \bar{h}_{n-d}^*,$$

Each state set contains 2^d states. The probability estimation of each state needs at most m comparisons, so the time cost by each state set is $O(2^d m)$. When d is given and the probabilities of all states are computed, it is easy to see that the running time of the above DP algorithm is $O(2^d n)$ and the time of the complete algorithm is $O(2^d mn)$. Since d is a small constant, so the time complexity of DP is $O(n)$ and the time complexity of the complete algorithm is $O(mn)$. In the following we will prove that (H^*, \bar{H}^*) is the haplotype pair that solves problem (3).

Theorem 1 *given a SNP matrix X and corresponding Markov chain model $MC(d)$, assume genotype G is known. If $P^* > 0$, then the haplotype pair (H^*, \bar{H}^*) determined by above DP algorithm is an optimal haplotype pair compatible with G , i.e. $(H^*, \bar{H}^*) \in C(G)$ and*

$$\Pr(H^* | X) \Pr(\bar{H}^* | X) = \max_{\{H, \bar{H}\} \in C(G)} \Pr(H | X) \Pr(\bar{H} | X).$$

Proof. According to the definition of state set, the states of each state set are all compatible with genotype G . When we assemble a pair of haplotypes, we only assemble the haplotype fragments in the state set. Since h_{t+1}^* is a successor of h_t^* , and \bar{h}_{t+1}^* is a successor of \bar{h}_t^* , $t = 0, 1, \dots, n - d$, the assembled haplotype pair must be compatible with genotype G .

To prove the optimality of (H^*, \bar{H}^*) , we adopt induction on t , $t = 0, 1, \dots, n - d$. If $C(G(1, t+d)) \neq \emptyset$, it follows from the iteration formula of DP algorithm that

$$\begin{aligned} P(h_t) &= \max_{h_{t-1} \in S_{t-1}(d)} P(h_{t-1})p^t(h_t | h_{t-1})p^t(\bar{h}_t | \bar{h}_{t-1}) \\ &= \max_{\substack{H^p = h_0 \rightarrow \dots \rightarrow h_t \\ \bar{H}^p = \bar{h}_0 \rightarrow \dots \rightarrow \bar{h}_t}} P(h_0) \prod_{k=1}^t p^k(h_k | h_{k-1})p^k(\bar{h}_k | \bar{h}_{k-1}) \end{aligned}$$

where $H^p \oplus \bar{H}^p = G(1, t+d)$, otherwise $P(h_t) = 0$. According to the frequency computation of haplotype fragments, the transition probability of $MC(d)$ is nonnegative, and $P^* > 0$ indicates that there exists a pair of haplotype (H, \bar{H}) , $H = h_0 \rightarrow h_1 \rightarrow \dots \rightarrow h_{n-d}$, $\bar{H} = \bar{h}_0 \rightarrow \bar{h}_1 \rightarrow \dots \rightarrow \bar{h}_{n-d}$ compatible with G . Therefore,

$$\begin{aligned} P^* &= \max_{h_{n-d} \in S_{n-d}(d)} P(h_{n-d}) \\ &= \max_{\substack{H = h_0 \rightarrow \dots \rightarrow h_{n-d} \\ \bar{H} = \bar{h}_0 \rightarrow \dots \rightarrow \bar{h}_{n-d}}} P(h_0) \prod_{k=1}^{n-d} p^k(h_k | h_{k-1})p^k(\bar{h}_k | \bar{h}_{k-1}) \\ &= \max_{\{H, \bar{H}\}} \Pr(H | X) \Pr(\bar{H} | X) \end{aligned}$$

where $(H, \bar{H}) \in C(G)$. When (H, \bar{H}) is incompatible with G , $\Pr(H | X) = \Pr(\bar{H} | X) = 0$. Since

$$\sum_{\{H, \bar{H}\} \in C(G)} \Pr(H | X) \Pr(\bar{H} | X)$$

is a constant for given X , the conclusion is proved. \square